Costruzione semi-automatica di corpora orientati al genere in lingue morfologicamente ricche: Un paragone fra l'italiano e il serbo

*Maja Miličević, **Silvia Bernardini e **Adriano Ferraresi

(*Università di Belgrado, **Università di Bologna)

Italica Belgradensia 1/2014. 99-114.

Maja Miličević[1]
Silvia Bernardini[2]
Adriano Ferraresi[2*]
Università di Belgrado[1], Università di Bologna[2]

COSTRUZIONE SEMI-AUTOMATICA DI CORPORA ORIENTATI AL GENERE IN LINGUE MORFOLOGICAMENTE RICCHE: UN PARAGONE FRA L'ITALIANO E IL SERBO[**]

Abstract: Il presente contributo tratta di costruzione semi-automatica di corpora dal web sulla base del genere testuale, utilizzando il programma *BootCaT*. In particolare, vengono riportati i risultati di due studi paralleli condotti sull'italiano e sul serbo, due lingue con morfologia flessiva molto ricca. Negli studi sono stati paragonati quattro metodi diversi per la costruzione di corpora orientati al genere piuttosto che all'argomento, basati su parole chiave e sequenze di parole (*n*-grammi) di tre lunghezze diverse (unigrammi, bigrammi e trigrammi). Il genere studiato è stato quello delle ricette culinarie, un genere molto formulaico e molto presente sulla rete Internet. L'analisi dei corpora creati tramite l'uso dei quattro metodi prescelti mostra che per l'italiano i risultati migliori sono conseguiti con parole chiave, mentre per il serbo non c'è un vantaggio significativo di uno dei metodi, ma sia le parole chiave che i bigrammi e i trigrammi producono risultati superiori a quelli ottenuti per l'italiano con parole chiave. Tali risultati confermano l'applicabilità di metodi di costruzione di corpora orientati al genere per lingue diverse dall'inglese e, interpretati in chiave contrastiva, appaiono indicativi dell'importanza delle differenze morfologiche fra le due lingue, specialmente per quanto riguarda la maggiore ricchezza della morfologia nominale in serbo, e la presenza degli articoli in italiano.

Parole chiave: *genere testuale, corpora semi-automatici, BootCaT, parole chiave, n-grammi, morfologia flessiva*

1 INTRODUZIONE

Gli studi linguistici degli ultimi decenni sono stati fortemente segnati dallo sviluppo della linguistica dei corpora. Fra l'altro, grazie all'espansione della rete Internet, la creazione di corpora digitali è diventata facilmente realizzabile anche per utenti senza particolari competenze computazionali come linguisti, traduttori e apprendenti di lingue straniere. Il web costituisce una fonte ricchissima di dati linguistici, e nonostante il problema della mancanza di controllo sull'origine dei dati pubblicati ha il grande vantaggio di rendere facilmente reperibili quantità molto ingenti di testi autentici.

[*] m.milicevic@fil.bg.ac.rs, silvia@sslmit.unibo.it, adriano@sslmit.unibo.it
[**] Questo lavoro è stato in parte condotto nell'ambito del Progetto n. 178004 "Standardni srpski jezik: sintaksička, semantička i pragmatička istraživanja" (Lingua serba standard: esplorazioni sintattiche, semantiche e pragmatiche), finanziato dal Ministero della Pubblica Istruzione, Scienza e Sviluppo Tecnologico della Repubblica di Serbia, di cui fa parte Maja Miličević. La parte del contributo dedicata alla lingua serba è ripresa dall'intervento al workshop *BootCaTters of the world unite!* (Università di Bologna a Forlì, 24 giugno 2013), i cui partecipanti si ringraziano per gli spunti emersi durante la discussione. Gli autori ringraziano inoltre i valutatori anonimi di questo articolo per i loro validi suggerimenti.

Uno dei programmi ampiamente usati per la trasformazione dei dati del web in corpora testuali veri e propri è *BootCaT* (*Bootstrapping Corpora and Terms*; Baroni e Bernardini 2004)[1]. La funzione principale di *BootCaT* è quella di creare in maniera semi-automatica corpora specialistici di dimensioni medio-piccole, in particolare relativi a un determinato argomento (musica classica, cibi per cani, malattie cardiovascolari, ecc.). Il programma interroga il web sulla base di criteri definiti dall'utente (principalmente sotto forma di parole chiave, ovvero termini caratteristici dell'argomento che si vuole studiare), permettendo di restringere i risultati al solo dominio di interesse. Come prodotto finale l'utente ottiene tutto il materiale testuale raccolto dal web in un unico file di testo che può essere interrogato e usato in modi diversi[2]. Alcuni tentativi di superare l'approccio utilizzato in *BootCaT* (si veda in particolare Barbaresi 2014) non sembrano al momento aver prodotto strumenti efficaci e liberamente disponibili per la costruzione di corpora in ambiti circoscritti dall'utente, tanto che questo strumento sembra ad oggi rappresentare la migliore alternativa, dati gli scopi del presente lavoro[3].

Estendendo il metodo sopra descritto, alcuni studi recenti hanno esaminato anche la possibilità di utilizzare *BootCaT* per la creazione di corpora orientati al genere testuale piuttosto che al dominio. Visto che, contrariamente al dominio, il genere rappresenta una categoria definita per lo più da fattori esterni al testo (cfr. sezione 2), compiti quali la classificazione automatica di testi sulla base del genere o la creazione automatica di corpora appartenenti a determinati generi sono fra i più difficili nel campo della linguistica computazionale/dei corpora. Come si vedrà di seguito, numerosi tentativi di individuare i correlati del genere all'interno del testo hanno dimostrato che un indizio piuttosto affidabile potrebbe essere costituito da sequenze frequenti di parole, i cosiddetti *n*-grammi. Nei lavori basati su *BootCaT*, però, si è verificato che l'approccio fondato su combinazioni di parole porta a tassi di successo diversi a seconda della lingua. In particolare, Bernardini e Ferraresi (2013) hanno trovato che con trigrammi (sequenze di tre parole) si ottengono buoni risultati per l'inglese ma non per l'italiano; il fattore proposto dagli autori come decisivo per il minor successo del metodo trigrammi in italiano è la ricca morfologia flessiva di questa lingua rispetto all'inglese, soprattutto per quanto riguarda le forme coniugate dei verbi.

Partendo dai risultati delle ricerche precedenti, in questo contributo ci proponiamo di approfondire l'analisi dell'italiano e di effettuare una prima valutazione delle diverse procedure per la creazione semi-automatica di corpora orientati al genere in serbo, un'altra lingua morfologicamente ricca. Di seguito saranno paragonati i risultati di quattro metodi diversi, implementati in *BootCaT*: parole chiave, unigrammi, bigrammi e trigrammi. Il genere testuale oggetto di attenzione sono le ricette culinarie, un genere diffuso sul web e caratterizzato da linguaggio molto formulaico. I risultati mostrano che

---

[1] http://bootcat.sslmit.unibo.it/

[2] Uno degli usi principali di *BootCaT* consiste nell'assistere i traduttori nel processo di documentazione su un dominio di interesse, ma i corpora specialistici sono utili anche per la costruzione di basi di dati terminologiche e, nell'ambito della linguistica computazionale, per compiti di *machine learning* (cfr. Baroni e Bernardini 2004: 1313).

[3] Un approccio interessante, sebbene non ancora del tutto maturo, appare quello proposto da Wong et al. (2011), che consiste nel filtrare i risultati ottenuti dai motori di ricerca in base alla rappresentatività, specificità e autorevolezza dei domini identificati, utilizzando metriche derivate dalla *web science*.

il metodo migliore per l'italiano è quello basato su parole chiave, mentre per il serbo (limitatamente all'alfabeto latino utilizzato in questo studio) tutti i metodi ad eccezione degli unigrammi producono buoni risultati.

2 IL GENERE TESTUALE E LA SUA IDENTIFICAZIONE AUTOMATICA

Il genere testuale è definito in primo luogo dalle funzioni svolte dal testo nella comunità di appartenenza (cfr. Sinclair 2005). Cioè, nelle parole di Swales, "[un] genere costituisce una classe di eventi comunicativi, i cui membri condividono una serie di obiettivi comunicativi" (1990: 58)[4]. Visto che criteri extralinguistici quali appunto l'obiettivo comunicativo di un testo sono difficilmente operazionalizzabili in termini computazionali (come anche nelle ricerche sul web), la selezione automatica di testi appartenenti ad uno stesso genere deve essere realizzata in modi diversi. Come spesso viene notato nella letteratura (cfr. Swales 1990, Crowston and Kwasnik 2004, ecc.), testi appartenenti ad uno stesso genere tendono ad avere in comune anche certe caratteristiche formali (dal punto di vista di stile, struttura e contenuto), il che permette la selezione e classificazione dei generi attraverso criteri linguistici. Fra le caratteristiche interne al testo considerate utili per l'identificazione del genere si possono distinguere quelle strettamente computazionali e non molto intuitive, come le sequenze di caratteri (*character n-grams* in inglese; cfr. Kanaris e Stamatatos 2009), quelle intuitive ma difficilmente ottenibili, come le sequenze di parti del discorso (*POS n-grams*; cfr. Sharoff 2007), e infine quelle intuitive e facilmente ottenibili, come le sequenze di parole[5].

Le combinazioni di parole si sono verificate particolarmente utili come indizi del genere in inglese. Sequenze ininterrotte di quattro parole sono state usate da Biber e Conrad (1999) nel loro studio delle differenze fra conversazione e prosa accademica; sulla base dell'*International Corpus of English* (ICE), Gries e Mukherjee (2010) hanno studiato le variazioni regionali nell'inglese usato in Asia paragonando sequenze di parole di varia lunghezza; Gries et al. (2011) hanno trovato che è possibile distinguere registri e sottoregistri presenti nei corpora *ICE-GB* e *BNC-Baby* grazie agli *n*-grammi in generale, e in particolare ai trigrammi; i bigrammi si sono rivelati la soluzione migliore negli studi condotti da Crossley e Louwerse (2007) e Louwerse e Crossley (2006), presumibilmente perché catturano anche una parte del contesto sintattico e semantico al pari di unità più lunghe, ma a differenza di queste ultime rimangono comunque frequenti, riducendo il rischio di sottorappresentazione nel corpus (*data sparseness* in inglese).

Per quanto riguarda l'italiano, Baroni et al. (2004) hanno riportato buoni risultati per la classificazione per generi dei testi componenti il corpus *la Repubblica* sulla base di unigrammi non-lemmatizzati. Per la lingua serba, Vitas et al. (2006) hanno osservato pochi bigrammi e trigrammi frequenti in comune (ovvero poco *overlap*) tra due corpora di testi narrativi e due corpora di testi giornalistici, il che potrebbe essere

---

4 "A genre comprises a class of communicative events, the members of which share some set of communicative purposes."

5 Per sequenze di parole non si intendono solo unità lessicali o sintattiche, ma gruppi di parole che appaiono spesso insieme in un determinato insieme di testi.

riconducibile alla rilevanza degli *n*-grammi per la classificazione del genere, ma anche all'incidenza di fenomeni di *data sparseness*.

Gli studi che hanno esaminato la possibilità di creare corpora orientati al genere con *BootCaT* hanno ottenuto risultati conformi a quelli citati sopra. Per l'inglese l'utilizzo di trigrammi, o alternativamente un misto di trigrammi e parole chiave, costituisce un metodo più efficace rispetto alle sole parole chiave, ma un risultato simile non è stato replicato per l'italiano. In particolare, Bernardini e Ferraresi (2013) hanno paragonato l'approccio "tradizionale" di *BootCaT*, basato su parole chiave, con il metodo fondato su trigrammi frequenti nel genere studiato, in questo caso i foglietti illustrativi dei medicinali, in inglese e italiano. La valutazione dei quattro corpora creati ha rivelato che il metodo trigrammi dà risultati superiori al metodo parole chiave soltanto per l'inglese, suggerendo per l'italiano una possibile influenza della ricca morfologia flessiva, in primo luogo i verbi coniugati. Adottando una metodologia simile, Dalan (2012) ha studiato il genere delle guide web universitarie in inglese e ha dimostrato che combinando trigrammi e parole chiave si ottengono risultati migliori rispetto ai metodi singoli; i trigrammi rimangono comunque un'alternativa valida e appena meno efficiente.

Mentre si può concludere con una certa sicurezza che il metodo basato su *n*-grammi può essere usato con successo in *BootCaT* per il reperimento di testi appartenenti ad un dato genere, rimangono da chiarire le motivazioni a cui ricondurre le limitazioni di questo metodo osservate in una lingua morfologicamente più ricca dell'inglese, come l'italiano.

3 CORPORA SEMI-AUTOMATICI DI RICETTE CULINARIE IN ITALIANO E IN SERBO

Per esaminare se, e attraverso quali metodi, si possono creare corpora orientati al genere con *BootCaT* anche per lingue con sistemi morfologici più ricchi dell'inglese, in questo studio ci concentriamo sull'italiano e sul serbo. Entrambe le lingue possiedono una morfologia flessiva molto ricca, una prevalentemente nell'ambito verbale (l'italiano), l'altra anche nel sistema nominale (il serbo)[6].

3.1 Metodo

L'approccio "standard" di *BootCaT* prevede un primo passo in cui l'utente sceglie un determinato numero di termini (*seeds*) (singole parole chiave o sintagmi caratteristici) che rappresentano il dominio di interesse. Questi termini vengono poi automaticamente combinati in gruppi (*tuples*) e inviati al motore di ricerca (al momento *Bing*); un gruppo di termini costituisce una ricerca (*query*). Come primo risultato si ottiene una lista di URL di pagine contenenti i gruppi di termini richiesti. L'utente può anche utilizzare alcune opzioni che influiscono sul risultato finale (lunghezza delle *tuple*, limitazioni di dominio web, ad esempio solo pagine dal dominio *.it* o *.rs*) ed eventualmente escludere URL giudicate non rilevanti. Nel passaggio finale, il contenuto delle pagine selezionate viene scaricato, ripulito del codice HTML, trasformato in testo semplice e salvato come un unico file.

---

6 Oltre alle categorie flessive nominali presenti anche in italiano (e l'accordo dell'aggettivo con il nome), il serbo possiede un sistema di declinazioni con sette casi e quattro classi di sostantivi.

Il metodo fondato su *n*-grammi, confrontato in questo studio a quello tradizionale, prevede l'inserimento, al posto delle parole chiave, di sequenze frequenti di parole (*n*-grammi appunto) nel genere prescelto. Visti i risultati degli studi precedenti, i quali hanno identificato come rilevanti non solo le sequenze di più parole, ma anche gli unigrammi non lemmatizzati (le parole più frequenti del genere in esame, in forma flessa), si sono creati quattro corpora diversi per ciascuna lingua, sulla base rispettivamente di (1) parole chiave, (2) unigrammi, (3) bigrammi, e (4) trigrammi. Per quanto riguarda invece il genere testuale, la scelta delle ricette culinarie è dovuta al fatto che esse costituiscono un genere familiare e molto presente sul web, essendo allo stesso tempo sufficientemente specializzato e convenzionale (si veda p. es. Arendholz et al. 2013; cfr. anche Bernardini e Ferraresi 2013). Inoltre, questo genere sta attirando sempre più attenzione nel campo della linguistica computazionale; le ricette culinarie in serbo, ad esempio, sono recentemente state studiate dal punto di vista dell'ampliamento del WordNet e dei dizionari morfologici elettronici per questa lingua (Vujičić Stanković et al. 2014).

Visto che sia le parole chiave sia gli *n*-grammi dovevano essere frequenti nel genere di interesse, come punto di partenza per la costruzione di corpora semi-automatici sono stati utilizzati corpora di ricette creati (semi-)manualmente. Più specificamente, attraverso semplici ricerche in *Google* sono stati identificati fra dieci e quindici siti dedicati alle ricette culinarie (quasi tutti divisi in categorie) per ognuna delle due lingue. Pagine tratte dai siti prescelti sono poi state scaricate con *BootCaT* utilizzando al posto delle *tuple* gli indirizzi dei siti preceduti dall'operatore *site:* (cfr. Bernardini et al. 2010). La successiva ripulitura manuale dei file, tramite la quale sono stati rimossi tutti i frammenti di testo non appartenenti al genere delle ricette (incluse le righe contenenti le URL, aggiunte automaticamente da *BootCaT*), ha permesso di ottenere un corpus italiano di 221.455 parole e un corpus serbo di 200.608 parole, entrambi di ottima qualità[7,8].

Per i corpora semi-automatici basati su parole chiave, le parole *seed* sono state ottenute tramite la funzione "*Keyword List*" del programma di analisi testuale *AntConc*[9], che paragona la frequenza delle parole nel corpus di interesse rispetto ad un corpus di riferimento[10]. Per l'italiano il corpus di riferimento utilizzato è stato un sottocorpus tratto da *Europarl*[11] (1.567.331 parole), mentre per il serbo si è creato un corpus di testi narrativi e giornalistici *ad hoc* (1.584.920 parole)[12]. Gli *n*-grammi sono stati estratti dagli stessi corpora manuali, sempre con *AntConc*, utilizzando la funzione

---

7 Tutti i copora in questo studio sono stati costruiti utilizzando la versione a riga di comando di *BootCaT* (invece di quella a interfaccia grafica). La versione utilizzata è quella descritta in Ljubešić (2013).

8 Una procedura simile per la creazione di un web corpus di ricette culinarie in serbo è stata adoperata da Vujičić Stanković et al. (2014), i quali però hanno usato programmi diversi, creati appositamente.

9 www.antlab.sci.waseda.ac.jp/antconc_index.html.

10 L'estrazione delle parole chiave è stata fatta per valore di *Log-Likelihood*.

11 http://www.statmt.org/europarl/

12 La parte più consistente del corpus è stata raccolta da Duško Vitas, Miloš Utvić e Cvetana Krstev, con l'aiuto degli studenti del Dipartimento di Scienze Informatiche della Facoltà di Filologia (Università di Belgrado), mentre un'altra porzione dei testi è stata fornita da Tanja Samardžić. Ringraziamo tutti i colleghi per il loro prezioso aiuto.

"*Clusters > N-Grams*" e definendo la lunghezza dei *cluster* come uno, due e tre[13]; l'estrazione di *n*-grammi non richiede l'uso di corpora di riferimento[14].

Per la creazione dei corpora semi-automatici sono state usate le prime 50 parole/sequenze delle rispettive liste. Al fine di produrre un numero simile di parole inserite nella *query* per i diversi metodi, la lunghezza delle *tuple* è stata impostata a 5 per i corpora basati su parole chiave e unigrammi, a 4 per quelli basati su bigrammi e a 3 per quelli basati su trigrammi. A parte la trasformazione in caratteri minuscoli, nelle fasi della selezione delle parole *seed*, della creazione delle *tuple* e della selezione delle URL non sono stati effettuati interventi manuali. Esempi di gruppi di parole utilizzati per la creazione degli otto corpora semi-automatici sono riportati in tabella 1.

	Italiano	**Serbo**[15]
Parole chiave	piatto patate ciotola burro fate bene aggiungete crema padella caldo	1 pecite kašika fil stavite minuta mleka ulje šećera kuvajte
Unigrammi	quindi i con a di forno in minuti si cottura	stavite preko od vode ne priprema staviti umutiti pa za
Bigrammi	"in un" "con il" "pizzico di" "e fate" "circa minuti" "il tutto" "la pasta" "con una"	"i na" "ostaviti da" "i pecite" "se ohladi" "minuta na" "kada se" "100 g" "30 minuta"
Trigrammi	"a temperatura ambiente" "e fate cuocere" "con un cucchiaio" "di sale e" "cucchiai di olio" "di tanto in"	"posolite i pobiberite" "u podmazan pleh" "čokolade za kuvanje" "papirom za pečenje" "u zagrejanoj rerni" "i beli luk"

Tabella 1: Esempi di *tuple* utilizzate per la costruzione dei corpora semi-automatici

Per ogni lista di *seed* sono state create 20 *tuple* e il numero di pagine da scaricare per *tupla* è stato fissato a 20, senza alcuna limitazione di dominio Internet.

3.2 Valutazione dei corpora

Le due tabelle sottostanti riassumono i dati principali sui corpora creati; la tabella 2 riguarda l'italiano, la tabella 3 il serbo. Le informazioni riportate sono il numero di URL da cui sono stati scaricati testi, il numero di parole nel corpus (escludendo le righe contenenti le URL), e la percentuale di pagine appartenenti al genere studiato, stimata attraverso un'analisi a campione.

---

[13] Chiaramente, gli unigrammi non sono dei *cluster* veri e propri, ma parole singole ordinate per frequenza, che potevano essere ottenute anche tramite la funzione "*Word List*".

[14] In uno studio futuro sarebbe interessante esaminare anche gli "*n*-grammi chiave", cioè sequenze di parole ottenute dal confronto con un corpus di riferimento. Essendo questa opzione assente in *AntConc*, qui ci limitiamo a *n*-grammi frequenti all'interno di un dato corpus.

[15] Parole chiave: 1, cuocete [al forno], cucchiaio.NOM, crema.NOM/ACC, mettete; minuti.GEN, latte.GEN, olio.NOM/ACC, zucchero.GEN, cuocete. Unigrammi: mettete, sopra, di, acqua.GEN, no; preparazione.NOM, mettere, battere, poi, per. Bigrammi: "e a", "lasciare che", "e cuocete [al forno]", "si raffredda"; "minuti.GEN a", "quando si", "100 g", "30 minuti.GEN". Trigrammi: "salate e pepate", "in teglia.ACC unta.ACC", "cioccolato.GEN da cucina.ACC"; "carta.INS da forno.ACC", "in forno.LOC preriscaldato.LOC", "e aglio.NOM/ACC".

	Parole chiave	Unigrammi	Bigrammi	Trigrammi
N URL	367	365	351	306
N parole	153.316	446.701	452.199	606.563
URL rilevanti	84%	44%	76%	78%

Tabella 2: Dati sui corpora semi-automatici di lingua italiana

	Parole chiave	Unigrammi	Bigrammi	Trigrammi
N URL	314	339	324	321
N parole	167.605	204.266	191.521	265.382
URL rilevanti	90%	72%	90%	88%

Tabella 3: Dati sui corpora semi-automatici di lingua serba

Come si evince dalle tabelle riassuntive, sia il numero di URL che quello di parole sono, in media, più elevati per l'italiano che per il serbo. La differenza è particolarmente evidente per il numero di parole: per tutti i metodi tranne quello basato su parole chiave i corpora di italiano hanno dimensioni maggiori almeno del doppio rispetto a quelli di serbo. Una prima ispezione manuale indica che le differenze sono dovute in particolare ad alcune pagine singole molto lunghe in italiano (come un blog di circa 45.000 parole, con un centinaio di ricette), e non a differenze sistematiche nell'organizzazione delle pagine in italiano e in serbo o ai metodi adoperati. Tale caratteristica potrebbe essere approfondita in studi futuri.

Valori ancora più interessanti per la nostra analisi emergono dall'osservazione delle percentuali di pagine rilevanti. Allo scopo di valutare le diverse procedure per la creazione di corpora orientati al genere in italiano e in serbo qui usiamo il cosiddetto criterio della *precision*, il cui valore è calcolato dividendo il numero di pagine appartenenti al genere target per il numero totale di pagine valutate. Per ogni corpus sono stati selezionati in modo casuale e valutati 50 testi (corrispondenti a 13,6-16,3% del totale). La percentuale di URL rilevanti ottenuta sulla base di questi campioni è stata utilizzata per stimare la composizione degli interi corpora. Va notato che sono state valutate come rilevanti anche le pagine in cui solo alcuni contenuti appartengono al genere target (per es. nei casi in cui il testo scaricato è costituito da una ricetta seguita da commenti dell'autore e/o dei lettori).

Come si vede molto chiaramente dalle tabelle 2 e 3 i risultati ottenuti per il serbo sono migliori di quelli per l'italiano, in totale come in ogni singola categoria. Le parole chiave rimangono il metodo più efficace per l'italiano anche quando il corpus mira a rappresentare un genere testuale invece che un argomento. Nell'ordine, le parole chiave selezionano l'84% di testi rilevanti, seguite da trigrammi e bigrammi (78% e 76% rispettivamente) e infine da unigrammi, con i quali si ottiene meno della metà delle pagine appartenenti al genere desiderato. Gli unigrammi producono il risultato meno soddisfacente anche per il serbo, mentre i risultati di tutti gli altri metodi si attestano attorno al 90%. Nel confronto tra le due lingue, l'accuratezza (in termini di *precision*) del metodo unigrammi in serbo si avvicina a quella dei trigrammi e bigrammi in italiano (72%).

In un'altra (parziale) ispezione manuale dei corpora si è scoperto che le pagine non corrispondenti al genere target sono tipicamente simili per argomento (per lo più articoli su cibo e salute), o fanno parte di una sezione non rilevante di pagine appartenenti al genere target (di solito commenti dei lettori)[16]. Per la parte restante, si osservano URL con sezioni in lingua sbagliata (soprattutto inglese e spagnolo), pagine simili per genere ma non per argomento (per es. consigli sull'abbigliamento), e pagine non riconducibili né al genere né all'argomento (per es. articoli su personaggi famosi o descrizioni di componenti hardware).

L'ultimo criterio valutato è stato l'*overlap* fra le URL contenute nei diversi corpora. Per l'italiano non ci sono URL condivise fra due (o più) corpora. Paragonando il corpus manuale con i corpora semi-automatici si scopre che anche gli *overlap* parziali (diverse pagine dello stesso sito) sono pochissimi, circa 3-6 URL. Fra i corpora semi-automatici, invece, i casi di *overlap* parziale sono numerosi, soprattutto fra siti spesso non rilevanti come *Yahoo Answers* o *Wikipedia*. La situazione è abbastanza simile per il serbo, visto che anche in questo caso non si osservano URL condivise. I casi di *overlap* parziale, però, anche se meno numerosi che nei corpora di italiano, sono più rilevanti e riguardano pagine diverse degli stessi siti appartenenti al genere target. In generale, nonostante i *seed* derivino da una medesima fonte, i corpora risultanti sono quindi molto diversi gli uni dagli altri.

3.3 Discussione

I risultati riportati nella sezione precedente confermano che la selezione automatica di pagine web sulla base del genere testuale è più complessa rispetto a quella per dominio di appartenenza. In particolare, si è verificato che per il genere testuale delle ricette culinarie si ottengono risultati diversi per lingue diverse, in questo caso l'italiano e il serbo, pur essendo entrambe le lingue morfologicamente ricche.

Per quanto riguarda l'italiano i risultati ottenuti non sono affatto deludenti, se si eccettua il metodo unigrammi. Le percentuali di pagine rilevanti fra il 76 e l'84% sono molto simili a quelle riportate da Bernardini e Ferraresi (2013), ovvero 80% per le parole chiave e 70% per i trigrammi (con una procedura di valutazione basata su giudizi di professionisti esterni allo studio). Inoltre, questi risultati sono superiori a quelli ottenuti per l'inglese da Dalan (2012) (parole chiave: 34%, trigrammi: 76%, metodo "misto": 81%), e parzialmente anche a quelli di Bernardini e Ferraresi (2013), i quali riportano per questa lingua un'accuratezza del 40% per parole chiave e del 90% per trigrammi. Tenendo conto delle differenze fra i generi studiati, e in particolare del grado di convenzionalità più basso delle guide web studiate da Dalan (2012) rispetto a foglietti illustrativi (Bernardini e Ferraresi 2013) e ricette culinarie (lo studio presente), si può comunque concludere che, sia per l'italiano che per l'inglese, è possibile creare corpora semi-automatici orientati al genere con precisione abbastanza alta. Resta invece il fatto che, contrariamente a quanto accade per l'inglese, il risultato migliore in italiano si ottiene usando parole chiave.

Sebbene manchino per il serbo punti di riferimento nei lavori precedenti, sarebbe lecito ipotizzare che i metodi basati su bigrammi e trigrammi siano meno efficaci di

---

[16] La presenza di sezioni non rilevanti può essere legata al fatto che *BootCaT*, nella fase di ripulitura dei testi, elimina dalle pagine scaricate parti di testo quali menu di navigazione ecc..

quelli basati su parole chiave, così come accade in italiano, avendo entrambe le lingue una morfologia flessiva ricca. In realtà questo non accade: i due metodi riescono infatti a selezionare numeri elevati di pagine orientate al genere target in serbo. Inoltre, in questo studio sono stati ottenuti risultati migliori per il serbo rispetto all'italiano anche con parole chiave (90 contro 84%) e unigrammi (72 contro 44%). In generale quindi, tutti i metodi sembrano funzionare meglio in serbo che in italiano.

Per quanto riguarda il metodo parole chiave, uno dei fattori che sicuramente ha avuto un ruolo importante nello studio presente è il genere testuale studiato. Come detto sopra, le ricette culinarie rappresentano un genere caratterizzato da un legame stretto con un dominio tematico (alimentari e bevande), avendo di conseguenza un numero elevato di parole chiave tipiche. Questo elemento può aver contribuito alla qualità dei corpora basati su parole chiave in entrambe le lingue studiate; rimane quindi da verificare se un risultato simile riapparirebbe nel caso di generi meno legati a campi semantici ristretti.

Il ruolo della morfologia flessiva, d'altro canto, è più complesso. Una correlazione diretta fra la ricchezza morfologica della lingua e l'accuratezza dei metodi basati su bigrammi e trigrammi chiaramente non esiste; se esistesse, ci sarebbe un vantaggio per l'italiano piuttosto che per il serbo. Il vantaggio per il serbo sembra invece risultare almeno in parte dalla natura marcatamente sintetica del suo sistema di morfologia nominale, diverso da quello dell'italiano. Nell'ambito nominale l'italiano possiede un sistema analitico che richiede l'uso di preposizioni per esprimere le relazioni grammaticali espresse in serbo tramite desinenze flessive. L'alta frequenza di preposizioni che ne consegue (le cinque preposizioni più frequenti, *di*, *con*, *a*, *in* e *per*, costituiscono l'11,55% del corpus manuale italiano, mentre i loro equivalenti *od*, *sa*, *na*, *u* e *za* rappresentano l'8,68% del corpus manuale serbo) non dovrebbe però di per sé influire così tanto sui metodi utilizzati in questo studio, visto che sono sempre state usate le prime 50 parole/sequenze più frequenti nel corpus manuale senza riguardo alle frequenze esatte, e visto che le preposizioni sono fra le parole più frequenti anche in serbo. La frequenza delle preposizioni in italiano diventa invece particolarmente importante quando si considera insieme all'articolo.

La presenza e soprattutto le diverse forme dell'articolo in italiano (*il*, *lo*, *l'*, *la*, *i*, *gli*, *le*; *un*, *uno*, *una*, *un'*) potrebbero infatti essere il fattore decisivo in questo studio, considerato che il serbo non possiede articoli. In effetti, guardando le liste dei *seed* usate per creare i corpora semi-automatici basati su *n*-grammi, si osserva che una parte sostanziale di parole in italiano è costituita da articoli e preposizioni articolate, che appaiono 15 volte fra gli unigrammi, 12 fra i bigrammi e 20 fra i trigrammi, portando, insieme alle sole preposizioni, ad un numero maggiore di parole e sequenze di parole funzionali rispetto al serbo (32 contro 22 per gli unigrammi, 16 contro 10 per i bigrammi, 7 contro 2 per i trigrammi). Queste parole sono fra quelle con meno contenuto lessicale nella lingua e per la maggior parte sono comuni a diversi generi testuali, contribuendo ad un aumento nel corpus del numero di pagine non appartenenti al genere target. Sono anche quelle che maggiormente influiscono sul metodo unigrammi: delle 20 *tuple* create, 4 contenevano solo parole funzionali (per es. "quindi", "i", "con", "a", "di"), e altre 4 una sola parola lessicale (per es. "che", "una", "anche", "olio", "gli"); non sorprende dunque che proprio questo metodo ottenga un valore di *precision* sostanzialmente inferiore rispetto agli altri (44%). I bigrammi e i trigrammi hanno prodotto risultati superiori a quelli ottenuti con gli unigrammi in ogni

probabilità grazie ad un numero maggiore di elementi lessicali, ma anche elementi funzionali più tipici del genere studiato (in particolare quelli con valore temporale e consecutivo/finale)[17], rimanendo comunque indietro rispetto ai bigrammi e trigrammi in serbo, dove elementi caratteristici del genere delle ricette culinarie sono ancora più presenti[18].

La qualità dei corpora di italiano creati usando bigrammi e trigrammi sembra quindi dipendere più dagli articoli e dall'assenza di declinazioni nominali che dalla presenza in questa lingua di complesse coniugazioni verbali. Il fatto che il metodo trigrammi dà buoni risultati per un'altra lingua analitica come l'inglese probabilmente ha a che fare non solo con il sistema di desinenze verbali molto semplificato, ma anche con il fatto che l'inglese possiede una forma unica per l'articolo determinativo (al posto delle sette forme dell'italiano) e due per l'articolo indeterminativo (contro quattro in italiano) e non usa preposizioni articolate. Anche queste supposizioni, però, dovranno essere verificate su generi meno convenzionali e tramite paragoni multipli fra l'inglese, l'italiano e il serbo (o altre lingue caratterizzate da tipologie morfologiche simili).

4 CONCLUSIONE

I risultati dei due studi paralleli condotti sull'italiano e serbo evidenziano una notevole rilevanza delle caratteristiche morfologiche per la creazione semi-automatica di corpora orientati al genere in *BootCaT*. Per il serbo, la lingua con più morfologia flessiva delle due, sia parole chiave che bigrammi e trigrammi hanno dato prova di grande affidabilità come indizi discriminanti nella selezione automatica di testi appartenenti a uno specifico genere testuale. Per l'italiano invece, le parole chiave si sono ancora una volta rivelate un metodo migliore rispetto agli *n*-grammi nell'isolare il genere testuale. I fattori decisivi dietro questa differenza sembrano essere le caratteristiche della morfologia nominale e il numero e tipologia di parole funzionali frequentemente usate nella lingua.

In conclusione, questo contributo ha mostrato che la creazione semi-automatica di corpora sulla base del genere rappresenta un compito realizzabile con *BootCaT* per lingue diverse, ma che il metodo più adatto e più accurato varia a seconda dalla lingua. Agli studi futuri rimane il compito principale di approfondire le ricerche finora condotte sull'italiano e sul serbo estendendo l'analisi a generi testuali meno convenzionali e/o utilizzando come termini di ricerca *n*-grammi chiave, derivati tramite paragoni con corpora di riferimento, al posto di *n*-grammi "semplici". In questo modo dovrebbe anche essere possibile escludere in italiano gli *n*-grammi formati principalmente di preposizioni e articoli, frequenti nella lingua generale e non caratteristici di un genere specifico. In ambito serbo sarebbe infine interessante ampliare lo studio tramite l'inclusione di testi in alfabeto cirillico.

---

[17] I trigrammi funzionali, ad esempio, sono maggiormente costituiti da locuzioni congiuntive quali *in modo da* e *fino a che*.

[18] Paragonando le liste utilizzate in questo studio con quelle riportate da Vitas et al. (2006), rappresentative della lingua generale, si trovano un unico bigramma e un unico trigramma in comune, specificamente *da se* 'che [cong.] si', il primo bigramma di tutti i corpora, e *da bi se* 'affinché, in modo che', presente solo nella lista del corpus meno generale dei quattro corpora studiati.

5 RIFERIMENTI BIBLIOGRAFICI

Arendholz, J., Bublitz, W., Kirner, M. e Zimmermann, I. (2013). Food for thought - or, what's (in) a recipe? In C. Gerhardt, M. Frobenius e S. Ley (a cura di), *Culinary Linguistics: The Chef's Special* (pp. 119-138). Amsterdam: John Benjamins.

Barbaresi, A. (2014). Finding viable seed URLs for web corpora: A scouting approach and comparative study of available sources. In *Proceedings of the 9th Web as Corpus Workshop* (pp. 1-8).

Baroni, M. e Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms form the web. In *Proceedings of LREC 2004* (pp. 1313-1316).

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G. e Mazzoleni, M. (2004). Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. In *Proceedings of LREC* 2004 (pp. 1771-1774).

Bernardini, S. e Ferraresi, A. (2013). Old needs, new solutions: Comparable corpora for language professionals. In S. Sharoff, R. Rapp, P. Zweigenbaum e P. Fung (a cura di), *Building and Using Comparable Corpora* (pp. 303-319). Dordrecht: Springer.

Bernardini, S., Ferraresi, A. e Gaspari, F. (2010). Institutional academic English in the European context: A web-as-corpus approach to comparing native and non-native language. In A. L. López e C. J. Rosalía (a cura di), *Professional English in the European Context: The EHEA Challenge* (pp. 27-53). Bern: Peter Lang.

Biber, D. e Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselgard e S. Oksefjell (a cura di), *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 181-190). Amsterdam: Rodopi.

Crossley, S. A. e Louwerse, M. M. (2007). Multi-dimensional register classification using bi-grams. *International Journal of Corpus Linguistics, 12*, 453-478.

Crowston, K. e Kwasnik, B. H. (2004). A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences - Track 4* (pp. 40100a).

Dalan, E. (2012). *Costruzione automatica di corpora orientati al genere e fraseologia: Il caso delle guide web in inglese degli Atenei europei*. Tesi di Laurea Magistrale, Università di Bologna.

Gries, S. Th. e Mukherjee, J. (2010). Lexical gravity across varieties of English: An ICE-based study of n-grams in Asian Englishes. *International Journal of Corpus Linguistics, 15*, 520-548.

Gries, S. Th., Newman, J. e Shaoul, C. (2011). N-grams and the clustering of registers. *Empirical Language Research Journal, 5*.

Kanaris, I. e Stamatatos, E. (2009). Learning to recognize webpage genres. *Information Processing and Management, 45*, 499-512.

Louwerse, M. M. e Crossley, S. A. (2006). Dialog act classification using n-gram algorithms. In G. Sutcliffe e R. Goebel (a cura di), *Proceedings of the 19th International Florida Artifficial Intelligence Research Society* (pp. 758-763). Menlo Park, CA: AAAI Press.

Ljubešić, N. (2013). Helping *BootCaT* to catch the Babel fish: Getting encoding, content and language right. Intervento al workshop "BootCaTters of the world unite!", Forlì, 24.6.2013.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. *Proceedings of Web as Corpus Workshop*. Louvain-la-Neuve.

Sinclair, J. (2005). Corpus and text: Basic principles. In M. Wynne (a cura di), *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxford: Oxbow Books.

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

Vitas, D., Pavlović-Lažetić, G. e Krstev, C. (2006). About word length counting in Serbian. In P. Gryzbek (a cura di), *Contributions to the Science of Text and Language: Word Length Studies and Related Issues* (pp. 301-317). Dordrecht: Springer.

Vujičić Stanković, S., Krstev, C. e Vitas, D. (2014). Enriching Serbian WordNet and electronic dictionaries with terms from the culinary domain. In H. Orav, C. Fellbaume e P. Vossan (a cura di), *Proceedings of the Seventh Global WordNet Conference* (pp. 127-132). University of Tartu, Estonia.

Wong, W., Liu, W. e Bennamoun, M. (2011). Constructing specialised corpora through analysing domain representativeness of websites. *Language Resources and Evaluation, 45*, 209-241.

SEMI-AUTOMATIC CREATION OF GENRE-BASED CORPORA FOR MORPHOLOGICALLY RICH LANGUAGES: COMPARING ITALIAN AND SERBIAN

Summary

This article deals with methods for the semi-automatic construction of genre-oriented corpora from the web, drawing on the *BootCaT* toolkit. In particular, it reports the results of two parallel studies on Italian and Serbian, chosen as examples of languages with very rich inflectional morphology. The two studies compare four different methods to create genre-, rather than topic-oriented corpora, based on keywords and sequences of words (*n*-grams) of different lengths (unigrams, bigrams and trigrams). The genre under scrutiny is that of cooking recipes, a genre that is very formulaic and also very frequent on the web. The analysis of the corpora created using the four different methods shows that the best results for Italian are achieved by the keyword method, while for Serbian no single method substantially outperforms the others; furthermore, the results obtained for Serbian are consistently better than those for Italian. As well as confirming the potential of genre-oriented methods of corpus construction for languages other than English, these results can be interpreted in a contrastive perspective, as highlighting the importance of morphological differences between the two languages, particularly as concerns the richer nominal morphology of Serbian compared to Italian, as well as the absence vs. presence of articles.

Key words: *genre, semi-automatic corpora, BootCaT, keywords, n-grams, inflectional morphology*