UN CORPUS DELLA STAMPA ITALIANA LOCALE

Simone TORSANI

ABSTRACT • A corpus of the Italian local press. This paper introduces CoSIL, a corpus of articles from Italian local newspapers containing about 180,000 texts and 66,000,000 words. The corpus was built to provide researchers with a freely downloadable balanced corpus of journalistic texts and a material for linguistic research on online local press, a nowadays-pervasive source of information. Besides the objectives behind the construction of the corpus, the paper describes its design and development, focusing on its representativeness and balance.

KEYWORDS • Corpus Design; Italian Language; Local Press.

1. Introduzione

Il presente contributo illustra la progettazione e la realizzazione del Corpus della Stampa Italiana Locale (CoSIL). CoSIL è un corpus di circa 66.000.000 di parole composto da articoli apparsi in versione digitale sulla stampa locale tra il 2003 e il 2019. Il corpus è diviso in sei categorie, che corrispondono alle sezioni più diffuse nei giornali presi in esame, cioè cronaca, economia e lavoro, politica, sport, cultura ed eventi e spettacoli.

Sebbene esistano già diversi corpora di italiano giornalistico (v. *infra*), CoSIL, per quanto di dimensioni più limitate rispetto ad essi, ha diversi punti di forza e originalità. In primo luogo, è liberamente scaricabile dal sito www.corpus-cosil.it e quindi può essere analizzato anche tramite strumenti diversi rispetto a quelli, pur ricchi, messi a disposizione dalle interfacce web attraverso le quali molti corpora sono consultabili. Inoltre, il corpus è suddiviso in categorie e permette perciò ricerche su ambiti specifici (es. confrontare il linguaggio dello sport e quello della politica). Infine, gli articoli fanno riferimento a fatti propri delle realtà locali, che possono essere trascurati dalla stampa nazionale; per esempio, fatti di microcriminalità (cronaca), questioni relative al tessuto produttivo locale (economia e lavoro) o eventi sportivi amatoriali (sport).

I corpora di linguaggio giornalistico in lingua italiana sono abbastanza numerosi. Il corpus del quotidiano La Repubblica (Baroni et al., 2004), di circa 380 milioni di parole che raccoglie articoli apparsi sul giornale tra il 1985 e il 2000 e accessibile in rete. Gli articoli de La Repubblica, così come quelli de La Stampa e del Corriere della Sera, sono inoltre accessibili in rete tramite gli archivi dei rispettivi quotidiani. In particolare, l'archivio del Corriere permette di effettuare ricerche su argomenti specifici (per es. politica). Non mancano, inoltre, progetti di corpora etichettati. Lo *Italian Content Annotation Bank* (I-CAB, Magnini et al., 2005), composto da articoli comparsi sul quotidiano l'Adige non è pubblico. I-CAB è di dimensioni più ridotte (circa 180.000 parole unità), ma annotato semanticamente.

Poiché CoSIL è costituito da articoli pubblicati in rete, esso confluisce, oltre che nell'alveo dei corpora di linguaggio giornalistico, anche nella famiglia dei *web corpora* (Baroni e Ueyama, 2006), cioè corpora compilati a partire da testi in rete, per l'italiano *Paisà* (Lyding et al., 2014),

itTenTen (Jakubíček et al., 2013) e *itWac* (Baroni et al. 2009). Quello dei *web corpora* è un settore dinamico e complesso, che sarebbe errato ridurre alle sole grandi dimensioni di alcuni corpora. Il settore infatti comprende anche lavori su tipi specifici di testi in rete, per esempio tratti da reti sociali (v. per es. Bosco et al., 2018), questioni più tecniche, come l'eliminazione di testo ridondante (come copyright menu ecc., v. per esempio, Schäfer, 2017) o metodi per categorizzare grande quantità di testi raccolti dalla rete (v. per esempio, Sharof, 2018). Tali questioni sono in parte emerse anche nella realizzazione di CoSIL (v. *infra*).

2. Composizione del corpus

2.1 Criteri per la selezione dei quotidiani

La presente sezione illustra i criteri utilizzati per la selezione dei quotidiani da includere nel corpus: la licenza sotto la quale sono distribuiti i contenuti del giornale, la sua registrazione come testata giornalistica e la natura locale del quotidiano.

In rete si trovano numerosi quotidiani ma, per ragioni di diritti di riproduzione, la scelta si è limitata, per il momento, a quelle testate che distribuiscono i loro contenuti sotto specifiche combinazioni di licenze *Creative Commons*. Le licenze *Creative Commons* sono composte da diverse condizioni: per esempio, una condizione stabilisce che "è necessario attribuire all'autore la paternità del prodotto" (BY). Tra le condizioni, ve n'è una ("non si possono realizzare opere derivate a partire dall'opera", ND), piuttosto ambigua sia perché diversamente interpretata dalle differenti legislazioni nazionali sia, soprattutto, perché non è chiaro se, ed entro quali limiti, un corpus costituisca o meno un'opera derivata. Per tali motivi, l'uso di testi rilasciati sotto tale specifica condizione è in genere sconsigliato nella compilazione di corpora (Kamocki e Ketzan, 2014). La scelta di non includere questi contenuti è stata perciò adottata anche nella compilazione di CoSIL. Il modello di riferimento nella scelta delle licenze è il corpus Paisà e sono state pertanto incluse solo opere rilasciate sotto una combinazione delle seguenti condizioni: attribuire la paternità dell'opera (BY), uso non commerciale (NC) e redistribuzione con la stessa licenza dell'opera originale (SA). La questione dei diritti è un tema molto importante e discusso nel settore e, visti anche gli sviluppi recenti, per esempio, l'uso di testi da reti sociali, non si limita più oggi alle sole licenze di distribuzione, ma tiene in conto anche elementi come la protezione dei dati personali (su GDPR e testi dalla rete v. tra gli altri, Basile, Lai e Sanguinetti, 2018 e Bosco et al., 2018).

Ulteriore condizione necessaria all'inclusione nel corpus è costituita dalla registrazione del sito come testata giornalistica presso un tribunale. Sebbene la legislazione permetta oggi di non registrare presso un tribunale i giornali in rete come testate giornalistiche (v. Sentenza 23230 della Corte di Cassazione del 13 giugno 2012), una maggiore rigidità in questo senso elimina ogni possibile ambiguità identificando inequivocabilmente come giornale ogni sito utilizzato per il corpus.

Ultima condizione necessaria all'inclusione è che la testata in questione sia un quotidiano locale. Vi sono infatti in rete diversi giornali che rilasciano i loro contenuti secondo le licenze *Creative Commons*, ma si tratta in alcuni casi di riviste specializzate e non quotidiani locali: un esempio per tutti è quello di Unimondo, giornale della Fondazione Fontana Onlus. Pubblicazioni di questo tipo non sono pertanto state incluse.

I quotidiani scelti sono riportati in tabella 1, insieme alle licenze sotto le quali sono rilasciati i loro contenuti.

Tabella 1 - I giornali del corpus

Quotidiano	**Città**	**Licenza**
CasertaSera	Caserta (Campania)	by-nc-sa/2.5/it/
Cervianotizie	Cervia (Emilia Romagna)	by/2.5/it/
Cesenanotizie	Cesena (Emilia Romagna)	by/2.5/it/
Ciavula	Caulonia/Gioiosa Ionica (Calabria)	by-sa/3.0/
Ferrara24Ore	Ferrara (Emilia Romagna)	by/4.0/
Forli24Ore	Forlì (Emilia Romagna)	by/4.0/
Forlinotizie	Forlì (Emilia Romagna)	by/2.5/it/
Gazzetta di Lucca	Lucca (Toscana)	by-nc-sa/4.0/
Gazzetta di Massa e Carrara	Massa e Carrara (Toscana)	by-nc-sa/4.0/
Gazzetta di Pistoia	Pistoia (Toscana)	by-nc-sa/4.0/
Gazzetta di Viareggio	Viareggio (Toscana)	by-nc-sa/4.0/
Il quotidiano	Ascoli (Marche)	by-nc-sa/2.5/it/
Lo Schermo	Lucca (Toscana)	by-nc-sa/3.0/it/
Lugonotizie	Lugo (Emilia Romagna)	by/2.5/it/
News-town	L'Aquila (Abruzzo)	by-nc/3.0/it/
Quartaparete	Napoli (Campania)	by/3.0/it/
Ravenna24Ore	Ravenna (Emilia Romagna)	by/4.0/
Rimininotizie	Rimini (Emilia Romagna)	by/2.5/it/
Rovigo24Ore	Rovigo (Veneto)	by/4.0/
Seitorri	Campobasso (Molise)	by-nc-sa/4.0/

Dati i vincoli delle licenze, la distribuzione geografica è piuttosto limitata, ma copre comunque una parte considerevole del territorio nazionale.

Figura 5 - Distribuzione geografica del corpus (in grigio le regioni rappresentate)

Le notizie, come prevedibile, fanno in genere riferimento alla realtà locale (v. *infra*), anche se non mancano notizie relative ad eventi di interesse nazionale (*Primarie del Pd, ad avere la meglio è Zingaretti*, da Ravenna 24 ore) o riguardanti realtà locali diverse dal territorio di

riferimento della testata (*I "disegni smisurati" del '900 italiano. Fino al 18 marzo in mostra al Casino dei Principi di Villa Torlonia a Roma*, da SeiTorri, Molise).

2.2 Categorie

Una caratteristica del corpus è la divisione delle notizie in categorie (ognuna delle quali raccolta in un documento a sé) e il bilanciamento tra le categorie stesse (sul bilanciamento v. *infra*). Le categorie sono state identificate a partire da quelle utilizzate nella sezione "Stampa" del Perugia Corpus (PEC, Spina, 2014), un corpus di riferimento per la lingua italiana[1].

La prima categoria **(A1 - Cronaca)** contiene notizie di cronaca o attualità. Si tratta di fatti in molti casi trascurati dalla cronaca nazionale come, tra gli altri, microcriminalità (*Droga. Alfonsine: rinvenuta una coltivazione di marijuana sull'argine del fiume Reno*, da Lugonotizie), sicurezza stradale (*L'Aquila, controlli della Stradale: la notte tra sabato e domenica ritirate 9 patenti*, da NewsTown), o, infine, notizie legate agli aspetti della vita quotidiana delle realtà locali (*Alife, arriva il micronido: al via le iscrizioni. Ginocchio: 'un grande aiuto per le famiglie'*, da Caserta Sera).

La seconda categoria **(A2 – Economia e lavoro)** contiene notizie legate all'economia e al lavoro. Le notizie in questa sezione si concentrano non tanto su argomenti di macroeconomia, come finanza o mercati, quanto sul tessuto produttivo locale (*Confindustria L'Aquila, Fracassi lascia il timone a Podda: 'Fusione con Teramo è un passaggio obbligato'*, da NewsTown), problemi legati all'occupazione (*Ilcea, i lavoratori incontrano il presidente della Provincia*, da Rovigo 24 ore) o all'agricoltura (*Nubifragio a Ravenna: ingenti danni in agricoltura*, da Lugo Notizie).

Anche le notizie nella categoria "politica" **(A3 - Politica)** hanno in genere un focus più ristretto rispetto alla stampa nazionale e si concentrano sulle attività amministrative, come commissioni comunali o regionali (*Venerdì si riunisce in Comune la commissione sulla trasparenza degli appalti*, da Lo Schermo), elezioni e liste civiche (*L'Aquila, primarie Pd: presentata mozione "Abruzzo per Martina"*, da News Town) o a interrogazioni/dichiarazioni su problemi locali (*Ancisi (LpR [Lista per Ravenna, n.d.A.]): 'Inutili gli 81mila euro spesi per le nuove fioriere, servono barriere mobili'*, da Ravenna 24 ore).

La categoria sport **(A4 - Sport)** non mostra, apparentemente, grosse differenze rispetto alla cronaca nazionale perché riporta in genere di eventi sportivi del tutto analoghi a quelli a carattere nazionale. L'unica differenza di rilievo è forse un maggiore equilibrio tra i diversi sport. Mentre nella stampa nazionale il calcio ha, in genere, una posizione predominante, nella stampa locale anche altri sport sembrano ben rappresentati proprio perché essa raccoglie e valorizza gli eventi sportivi e gli atleti del territorio. Si trovano notizie, per esempio, di basket, (*L'OraSì vince ancora con super Smith: a Jesi è 78-88*, da Ravenna 24 ore), pallavolo (*Sport: congratulazioni alle ragazze del VP Volley per il successo nel torneo giovanile di Rovereto*, da Lo Schermo), o sci (*Sci di fondo, Antonio Sassano sul podio del Criterium Interappenninico di Barrea*, da SeiTorri).

La sezione Cultura **(A5 - Cultura)** contiene notizie relative a eventi culturali, come mostre (*L'Aquila: inaugurata mostra 'Il Gran Sasso nell'animo. Paesaggi di Fulvio Muzi'*, da NewsTown), presentazioni di libri (*Oggi la presentazione di 'Dirsi di sì nelle città UNESCO'*, da Ravenna 24 ore) o convegni (*Musei del Polesine, se ne parla alla Vangadizza di Badia*, da Rovigo 24 ore).

[1] V. *infra* per una descrizione delle modalità di selezione delle notizie relativamente alle categorie

L'ultima categorie **(A6 - Spettacoli)** raccoglie notizie relative alle arti performative e, quindi, a eventi come concerti (*Il quartetto di Sara Jane in concerto al Sax Pub di Lugo con il nuovo disco 'In mancanza d'aria'*, da LugoNotizie), recensioni di spettacoli teatrali (*"Il gabbiano", elogio e denuncia del nulla che siamo*, da QuartaParete) o eventi legati al cinema (*Cronaca di una passione: chiude a Massa il tour del film di Fabrizio Cattani*, da La Gazzetta di Massa e Carrara).

Come già accennato, rispetto alla stampa nazionale gli eventi riportati nella stampa in oggetto hanno un focus diverso, concentrato, appunto, sulle realtà locali. Questo tratto distintivo di CoSIL emerge, come visto, in misura diversa nelle categorie. In alcune di esse, infatti, le notizie sono in genere analoghe rispetto alla stampa nazionale; per es. nella sezione "Spettacoli" si trovano recensioni del tutto analoghe a quelle che si trovano nella stampa nazionale. In altre categorie, invece, le notizie si concentrano su eventi propri della realtà locale, per esempio, nella sezione "Cronaca" si trovano notizie relative a fatti di microcriminalità.

2.3 Raccolta e categorizzazione dei dati

2.3.1 Raccolta degli articoli

Gli articoli sono stati estratti in parte tramite il software BootCat, un programma progettato per la creazione di *web corpora* (Baroni e Bernardini, 2004). Nella maggior parte dei casi gli indirizzi dei singoli articoli sono stati raccolti dal sito del quotidiano tramite script, quando ciò non è stato possibile, come nel caso de ilquotidiano.it, la ricerca dei testi è stata realizzata tramite motore di ricerca, impostando un filtro all'origine per non estrarre pagine non rilasciate sotto le licenze richieste. I testi raccolti sono stati in seguito trattati attraverso semplici script sviluppati *ad hoc* per ovviare a problemi minori emersi durante l'estrazione, come, per esempio, l'eliminazione di duplicati ed elementi ridondanti estratti dal programma come parte del testo. Come anticipato nella parte dedicata ai web corpora, questi sono problemi comuni in questo ambito, tanto che una parte della ricerca si occupa della progettazione di strumenti per la loro risoluzione. Nel caso presente, tuttavia, tali problemi sono stati piuttosto contenuti grazie anche al fatto che i testi sono stati raccolti da un numero limitato di fonti ed è stato, per esempio, facile identificare ed eliminare i pur numerosi elementi ridondanti, come indicazioni su come pubblicare commenti. I diversi articoli sono stati infine raccolti in un unico documento di testo per categoria (per es. A1.txt contiene tutti gli articoli della sezione "Cronaca"). All'interno del documento ogni articolo è associato a un numero che è ripreso in un documento a parte (in questo caso, A1_credits.txt) che contiene alcuni metadati dell'articolo: quotidiano, indirizzo e licenza. Per esempio, in A1.txt il testo "1) *Cesa. I carabinieri della stazione di Cesa hanno eseguito un ordine di carcerazione per un 38enne del luogo (…)* è associato in A1_credits.txt a "1) giornale: casertasera.it / link: https://casertasera.it/2016/04/20/416/ / licenza: by-nc-sa/2.5/it/"

2.3.2 Categorizzazione

La suddivisione in categorie ha costituito l'aspetto più problematico nel lavoro di costituzione del corpus. La situazione più semplice, ma anche più frequente, è quella in cui la sezione del giornale corrisponde alla categoria del corpus, come per esempio Sport. In tale caso le notizie sono state automaticamente inserite nella categoria corrispondente. Tuttavia, si sono ripetutamente presentate due situazioni di ambiguità che hanno reso impegnativa la

categorizzazione. Nel primo caso la notizia si trova in una categoria che corrisponde a due o più categorie di CoSIL. Il caso più evidente è costituito da Cultura e Spettacoli, due categorie che in molti giornali (per es. news-town.it) sono accorpate in un'unica sezione. In questi casi i testi sono stati suddivisi tramite un semplice script che effettua una prima classificazione sulla base di parole chiave nel testo (per esempio, "cinema", "teatro", "concerto" individuano articoli relativi alla sezione "Spettacoli"). I risultati sono stati successivamente controllati per correggere eventuali errori nella categorizzazione. Nel secondo caso la notizia si trova in una sezione, ma è associata tramite etichetta anche ad altre (anche più di due) sezioni. Nei giornali stampati, infatti, una notizia non può che trovar posto in una sola sezione; la natura ipertestuale dei giornali in rete, invece, permette facilmente di superare ogni gabbia tipologica precostituita facendo sì che, in alcuni casi, una notizia possa –legittimamente- comparire in più di una sezione. Per esempio, la notizia *De Pascale su visita di Matteo Salvini al Centro Olio Eni: riprendere il dialogo con i lavoratori* (da Cervia Notizie) è etichettata nel sito come "Politica", "Cronaca" ed "Economia" e fa in effetti riferimento a tutte e tre le sezioni. Questo fa sorgere il problema di dove categorizzare una notizia come quella appena citata. In tali casi si è scelto di mantenere solo quegli articoli etichettati con una sola delle sei categorie del corpus.

3. Dimensioni, bilanciamento e rappresentatività

CoSIL è composto da un totale di 66.871.172 parole unità e da 180.070 documenti unici. L'attenzione alla distribuzione dei dati rispetto alle categorie non è casuale, ma costituisce un tentativo di recepire l'ampio dibattito nella disciplina sui principi di bilanciamento e rappresentatività. I testi sono stati infatti raccolti e selezionati secondo criteri predefiniti che tendono, tramite un'organizzazione tematica, a renderlo il più rappresentativo possibile sotto il profilo della variabilità linguistica.

Sebbene rappresentatività e bilanciamento costituiscano uno degli aspetti centrali nella progettazione di un corpus, la questione rimane comunque sempre aperta e, ogni volta, di difficile soluzione (v. tra gli altri, McEnery e Hardie, 2011 a cui si rimanda per una panoramica delle diverse proposte in merito). Il modello adottato nella realizzazione di CoSIL è costituito dal Perugia Corpus (PEC); Spina (2014) offre, nella sezione quotidiani del corpus, una suddivisione delle diverse categorie di notizie che corrispondono alle normali sezioni di un quotidiano: editoriale, politica, economia, cronaca, esteri, cultura, sport, lettere e spettacolo. Tale suddivisione, tuttavia, è relativa alla stampa nazionale ed è stato necessario adattarla alle caratteristiche dei quotidiani in oggetto. Tale adattamento è consistito nell'eliminazione di alcune categorie, come "esteri", e nella ridefinizione delle percentuali delle restanti.

Tabella 2 - Frequenza delle principali categorie in un campione di giornali (l'asterisco indica una discriminazione semiautomatica)

	forli24ore	**cervianotizie**	**lo schermo**	**news-town**	**media**
Cronaca	43,90%	32,00%	10,93%	42,23%	32,27%
Economia	15,30%	13,60%	15,48% *	4,46%	12,21%
Politica	18,72%	12,80%	10,87% *	20,64%	15,76%
Sport	8,45%	12,80%	14,31%	9,13%	11,17%
Cultura	9,84% *	8,80%	29,24% *	17,70% *	16,39%
Spettacoli	3,80% *	20,00%	19,17% *	5,85% *	12,21%
TOTALE	100,00%	100,00%	100,00%	100,00%	100,00%

Nei quotidiani di CoSIL la categoria Editoriale non è presente in tutte le testate e si è perciò deciso di ometterla. Come prevedibile, inoltre, non esiste nei giornali una categoria Esteri

e le (poche) notizie relative ad avvenimenti internazionali sono in genere integrate nelle categorie politica o economia. Anche le lettere presentano una distribuzione poco chiara; in alcuni casi (es. Gazzetta di Viareggio) esse sono contenute in rubriche apposite, mentre in altri (es. Ravenna 24 ore) sono integrate nelle diverse categorie. Perciò, data la loro scarsa consistenza numerica, si è scelto anche in questo caso di non includerle nel corpus. Le altre categorie del PEC sono state mantenute, ma con una sostanziale modifica delle percentuali di rappresentazione ottenute sulla base della distribuzione delle stesse in un campione di giornali. Come si può osservare nella Tabella 2, le categorie presentano valori assai diversi da giornale a giornale: si è pertanto scelto di aggiustare i valori del PEC tenendo presente la media delle distribuzioni calcolata su un campione dei giornali del corpus. Rispetto al PEC si notano differenze marcate relativamente a cultura (15,44% rispetto all'11,60% di PEC), spettacolo (12,03 contro 5,20%) e sport (12,13% contro 7,40%). Questi dati sono in linea con la natura stessa dei giornali nei quali gli eventi locali, come manifestazioni sportive o spettacoli, hanno un grosso peso. La categoria Economia, infine, è stata rinominata in Economia e Lavoro per i motivi esposti nel par. 2.2.

Le percentuali delle diverse categorie calcolate sulla loro frequenza media sono riportate in Tabella 3

Tabella 3 - Bilanciamento delle categorie

codice	**categoria**	**Perc. sul totale**	**Perc. in PEC**
A1	cronaca	29,95%	27,0%
A2	economia e lavoro	12,29%	9,80%
A3	politica	18,16%	17,70%
A4	sport	12,13%	7,40%
A5	cultura	15,44%	11,60%
A6	spettacoli	12,03%	5,20%
TOTALE		100%	100%

Tabella 4 - Composizione di CoSIL

	numero documenti	**parole unità**	**parole unità per testo**	**parole tipo**	**Rapporto tra parole tipo e unità (type/token ratio)**
A1	64.270	20.031.554	311,68	203.223	0,01
A2	18.635	8.219.666	441,01	111.347	0,01
A3	29.607	12.142.730	410,13	136.454	0,01
A4	20.977	8.111.449	386,68	126.515	0,01
A5	27.281	10.323.207	378,40	184.016	0,02
A6	19.300	8.042.566	416,71	164.342	0,02
	180.070	66.871.172			

4. Sviluppi futuri e conclusioni

Il contributo presenta la progettazione e lo sviluppo del corpus CoSIL, una risorsa che si concentra su una parte diffusa dei testi in lingua italiana, cioè la stampa locale.

Allo stato attuale il corpus è limitato ai soli testi grezzi che lo compongono e mancano alcuni tra i principali strumenti in genere collegati a un corpus, come l'annotazione linguistica e un motore di ricerca sul sito. Questi due elementi costituiscono i prossimi ambiti di sviluppo del progetto. In particolare, è allo studio l'elaborazione tramite UDPipe, una pipeline per la tokenizzazione, lemmatizzazione, annotazione e parsing sintattico dei testi. Un vantaggio di UDPipe è che esso può essere addestrato su UD-Italian, il treebank di riferimento per l'italiano.

I primi esperimenti, condotti sull'interfaccia in rete di UDPipe, sono molto incoraggianti. Lo stesso vale per l'installazione della piattaforma CQPweb per l'interrogazione in rete del corpus. Una volta installato CQPweb si prevede di arricchire ulteriormente i metadati del corpus per rendere ancora più efficace l'interrogazione.

CoSIL si trova al crocevia di diverse tipologie di corpus da ognuna delle quali cerca di prendere i vantaggi. Come web corpus realizzato da testo *Creative Commons* COSIL può essere scaricato liberamente; come in un corpus di riferimento le categorie sono bilanciate in modo da rendere il corpus un campione affidabile e permettere ricerche mirate.

RIFERIMENTI BIBLIOGRAFICI

Aliprandi, Simone (2013), *Creative Commons: manuale operativo*, Milano, Ledizioni.

Baroni, Marco, Bernardini, Silvia (2004). *BootCaT: Bootstrapping Corpora and Terms from the Web*, in *Proceedingsof the 4th Language Resources and Evaluation Conference (LREC'04)*..

Baroni Marco et al. (2004), *Introducing the La Repubblica Corpus: A Large, Annotated, TEI (XML)-compliant Corpus of Newspaper Italian*. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*.

Baroni Marco et al. (2009), *The WaCky wide web: a collection of very large linguistically processed web-crawled corpora*, in *Language resources and evaluation*, *43*(3), 209-226.

Baroni, Marco, Ueyama, Motoko (2006), *Building general-and special-purpose corpora by web crawling*, in *Proceedings of the 13th NIJL international symposium, language corpora: Their compilation and application*.

Basile, Valerio, Lai, Mirko, Sanguinetti, Manuela (2018), *Long-term Social Media Data Collection at the University of Turin*. In *Fifth Italian Conference on Computational Linguistics (CLiC-it 2018)*.

Bosco, Cristina et al. (2018), *Overview of the EVALITA 2018 Hate Speech Detection Task*. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*.

Jakubíček, M Miloš et al. (2013), *The TenTen corpus family*, in *7th International Corpus Linguistics Conference CL*.

Kamocki, Paweł, Ketzan, Erik. (2014), *Creative Commons and Language Resources: General Issues and what's new in CC 4.0*. In: *CLARIN Legal Issues Committee (CLIC)-White Paper Series*.

Lyding, Verena et al. (2014), *The PAISA'Corpus of Italian Web Texts*, in 9*th Web as Corpus Workshop (WaC-9)@ EACL 2014* (pp. 36-43).

Magnini, Bernardo et al. (2006), *I-CAB: The Italian Content Annotation Bank*, in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*

McEnery, Tony, Hardie, Andrew (2011), *Corpus linguistics: Method, theory and practice*, Cambridge, Cambridge University Press.

Sharoff, Serge (2018), *Functional Text Dimensions for the annotation of web corpora*, in *Corpora*, *13*(1), pp. 65-95.

Spina, Stefania (2014), *Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione*, in *First Italian Conference on Computational Linguistics CLiC-it 2014* (Vol. 1, pp. 354-359). Pisa University Press.

SIMONE TORSANI • Lecturer and researcher in Educational linguistics at the Università di Genova. His research interests involve language teaching and ICT, digital humanities, corpus linguistics, and reading skills acquisition.

E-MAIL • s.torsani@gmail.com