CONTRAST-IT e COMPARE-IT
Due nuovi corpora per l'italiano contemporaneo

Anna-Maria De Cesare
Università di Basilea, Università della Svizzera italiana

The goal of this contribution is to describe two new comparable corpora to investigate contemporary Italian: CONTRAST-IT, a multilingual corpus including five languages (Italian, French, Spanish, German and English), and COMPARE-IT, the first monolingual corpus representing written neo-standard Italian in three countries (Italy, Switzerland and Canada). In the first part of this study, we describe the general design of these two new resources, representative of newspaper language, and compare CONTRAST-IT to similar research tools. In the second part, we present in more details COMPARE-IT by focusing on both quantitative and qualitative aspects related to the most frequently used words in the corpus. The results confirm the high comparability of the data included in COMPARE-IT, while also reflecting differences due to language contact phenomena between Italian and the languages spoken and written in the countries where the newspapers are published. As we point out in the conclusion, many (new) grammatical, textual and pragmatic phenomena can be investigated with CONTRAST-IT and COMPARE-IT, including some understudied phenomena described in this study, which deserve further investigation in particular to throw light on the idea that Italian is a weak bicentric language.

Keywords: multilingual and monolingual comparable corpora; contemporary Italian; journalistic texts; lexical analysis; invariable parts of speech

1. Introduzione

L'obiettivo generale di questo contributo è presentare due nuove risorse elettroniche per l'italiano contemporaneo: CONTRAST-IT e COMPARE-IT. Si tratta di due corpora comparabili di testi giornalistici allestiti per condurre indagini su corpora, sia *corpus-based* sia *corpus-driven*, di tipo contrastivo e comparativo (per una caratterizzazione dei termini *contrastivo* e *comparativo*, cfr. De Cesare

et al. 2016: 57-60)[1]. Il CONTRAST-IT è un corpus multilingue, che permette di studiare l'italiano contemporaneo in prospettiva contrastiva con quattro altre lingue europee (francese, spagnolo, tedesco e inglese). Il COMPARE-IT è un corpus monolingue, che permette di comparare la manifestazione dell'italiano neostandard in diversi Paesi e aree geografiche (Italia, Svizzera e Canada). Questo corpus costituisce attualmente un unicum nel panorama di risorse a disposizione per studiare la manifestazione dell'italiano contemporaneo scritto in diversi Paesi. L'obiettivo di questo contributo è anche di mostrare le potenzialità euristiche delle due nuove risorse, illustrando alcuni percorsi di ricerca che sfruttano il COMPARE-IT e fornendo spunti per ulteriori indagini.

Questo contributo è organizzato come segue: il § 2 presenta le caratteristiche generali di CONTRAST-IT e COMPARE-IT, prestando soprattutto attenzione a questioni legate al loro *design*, e descrive le specificità del primo sullo sfondo degli altri corpora ad accesso libero assimilabili alla nuova risorsa; il § 3 descrive poi più dettagliatamente il corpus COMPARE-IT, composto di testi giornalistici redatti in italiano pubblicati in Italia, in Svizzera e in Canada. L'obiettivo principale di questa parte del lavoro è evidenziare l'alta comparabilità dei campioni inclusi nel corpus COMPARE-IT, osservando caratteristiche lessicali, grammaticali (relative a due parti invariabili del discorso) e interpuntive. Metteremo anche a fuoco alcune differenze tra i testi redatti nei tre Paesi, formulando l'ipotesi che queste differenze siano legate a un uso diverso del codice lingua e a interferenze linguistiche prodotte dal contatto tra l'italiano e le lingue parlate e scritte nei Paesi in cui i giornali analizzati sono pubblicati. A guisa di conclusione (§ 4), si menzionano alcune piste di ricerca e domande aperte da indagare più approfonditamente con l'ausilio di CONTRAST-IT e COMPARE-IT, e si accenna ai modi in cui queste due risorse elettroniche possono essere ulteriormente sviluppate e arricchite in futuro, per rispondere a nuovi quesiti.

* Vorrei ringraziare due revisori esterni per i loro preziosi commenti sul contenuto di una versione precedente di questo lavoro.

[1] I corpora CONTRAST-IT e COMPARE-IT sono liberamente accessibili in rete da ottobre 2018. Sono stati creati con dati raccolti nell'ambito di due progetti di ricerca finanziati dal Fondo Nazionale Svizzero e diretti da chi scrive tra il 2011 e il 2018 (per ulteriori informazioni, cfr. http://p3.snf.ch/Project-133716 e http://p3.snf.ch/Project-159273). All'indirizzo https://contrast-it.philhist.unibas.ch/en/home si trova una descrizione dettagliata di queste risorse elettroniche, e altre informazioni d'interesse per chi opera nel campo della linguistica contrastiva e comparativa. La creazione dei due corpora è stata possibile grazie a un prolungamento del secondo finanziamento del FNS e ha coinvolto, oltre alla sottoscritta, Elisa Tekin (preparazione dei campionamenti di testo e aiuto nella creazione del sito web https://contrast.it citato sopra) e Lorenzo Gregori (in veste di curatore tecnico dei corpora e delle piattaforme web che li ospitano).

2. CONTRAST-IT e COMPARE-IT: caratteristiche generali

CONTRAST-IT e COMPARE-IT sono due nuovi corpora comparabili che includono l'italiano e sono di tipo rispettivamente multilingue e monolingue: il CONTRAST-IT comprende sottocorpora comparabili in italiano, francese, spagnolo, inglese e tedesco; il COMPARE-IT include invece testi esclusivamente in italiano, con sottocorpora comparabili di testi redatti in Italia, Svizzera e Canada. I due corpora sono stati allestiti utilizzando i *software open source* del pacchetto NoSketch Engine (Manatee, per la macchina di ricerca, e Bonito, per l'interfaccia grafica; su queste risorse, cfr. Rychlý 2007). Per l'etichettatura delle parti del discorso (POS-*tagging*) e la lemmatizzazione dei testi inclusi nei due corpora si è fatto ricorso al modulo linguistico più recente d'italiano, francese ecc. di TreeTagger[2]. I corpora CONTRAST-IT e COMPARE-IT includono testi pubblicati nella versione elettronica dei principali quotidiani nazionali di sette Paesi (per dettagli, cfr. § 2.1 e le pagine del sito menzionato in nota 1[3]). Le due risorse rappresentano globalmente la lingua (neo-)standard scritta. Esse sono state create per indagare, su base empirica, l'italiano contemporaneo in prospettiva rispettivamente contrastiva e comparativa: il primo corpus è pensato per individuare differenze e somiglianze tra l'italiano e le altre quattro lingue europee disponibili nel corpus; il secondo, per mettere in rilievo differenze e somiglianze tra l'italiano d'Italia e di altri Paesi in cui l'italiano è usato per iscritto nei *mass media* (e in due di essi ha lo statuto di lingua nazionale[4]). CONTRAST-IT e COMPARE-IT sono liberamente accessibili e interrogabili. Le Figure 1 e 2 presentano la pagina d'accesso e l'interfaccia di interrogazione dei due corpora. Le due piattaforme web consentono in particolare di fare ricerche per parole o lemmi, generare liste di

[2] I software relativi alle cinque lingue che ci interessano sono liberamente scaricabili all'indirizzo: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger.

[3] Il corpus CONTRAST-IT è descritto in modo più dettagliato alla pagina: https://contrast-it.philhist.unibas.ch/fileadmin/user_upload/contrast-it/CONTRAST-IT_detailed_corpus_design.pdf; il corpus COMPARE-IT alla pagina: https://contrast-it.philhist.unibas.ch/fileadmin/user_upload/contrast-it/COMPARE-IT_detailed_corpus_design.pdf
Per una descrizione approfondita dei quotidiani online (anche rispetto a quelli cartacei), si rinvia a Bonomi *et al.* 2002, Bonomi 2014 e De Cesare *et al*. 2016, Parte I, cap. 3.1.

[4] Importanti output di ricerca sono già stati prodotti analizzando il corpus ICOCP, composto di gran parte dei dati ora raccolti in CONTRAST-IT e COMPARE-IT (per una descrizione del corpus ICOCP, cfr. De Cesare *et al.* 2014a: 52-62 e De Cesare *et al.* 2016: 139-149). Si vedano per esempio i due lavori collettivi sulle strutture sintatticamente marcate di De Cesare *et al.* 2014b e 2016.

frequenza relative a parole, lemmi o parti del discorso, calcolare le collocazioni, visualizzare la distribuzione di una forma nei testi del corpus, restringere la ricerca a specifiche testate o rubriche tematiche (Cronaca, Economia ecc.), salvare sul proprio computer i risultati di una ricerca.

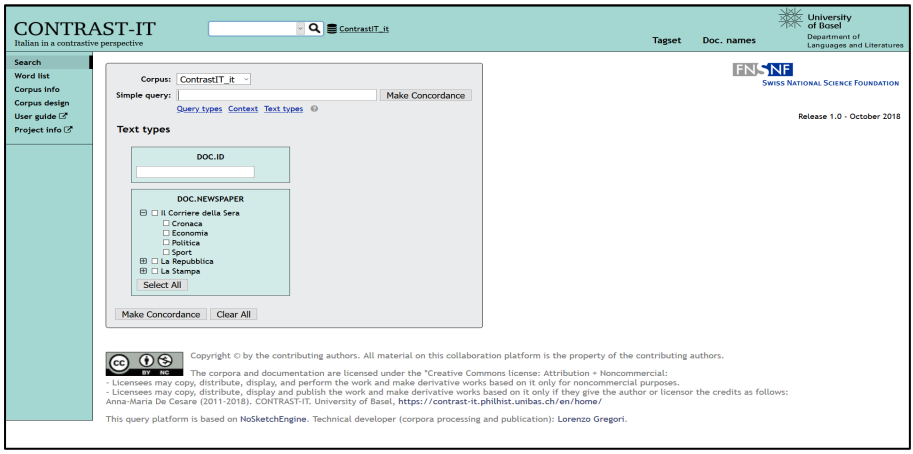

Figura 1. CONTRAST-IT[5]

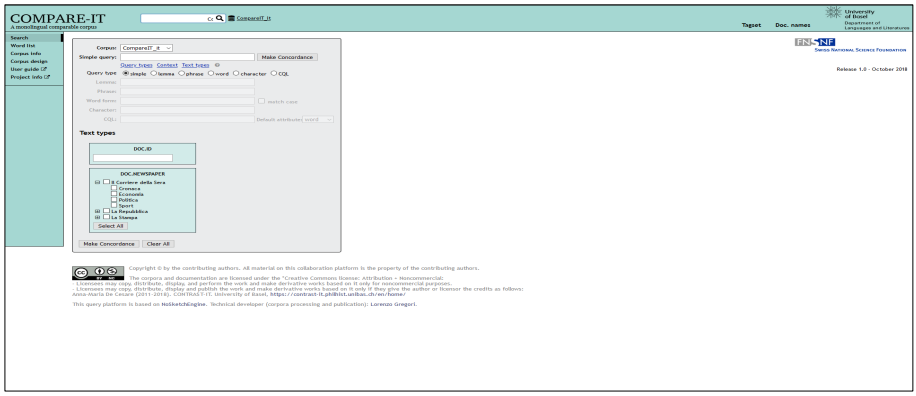

Figura 2. COMPARE-IT[6]

[5] http://philhist-contrast-it noske.philhist.unibas.ch/cnt/run.cgi/first_form
[6] http://philhist-contrast-it-noske.philhist.unibas.ch/cmp/run.cgi/first_form

2.1 Caratteristiche generali del corpus design

CONTRAST-IT e COMPARE-IT sono stati allestiti con campioni di testi che presentano le seguenti caratteristiche generali:

- *Rappresentatività*: i campioni di testi raccolti sono tratti dai grandi quotidiani nazionali dei rispettivi Paesi; più precisamente, questi testi sono tratti dalla versione elettronica delle principali testate di 7 nazioni (Italia, Francia, Spagna, Germania, Regno Unito, Svizzera, Canada); sono dunque il prodotto di *mass media* e rappresentano una delle forme più diffuse di testo scritto; inoltre, complessivamente, essi rispecchiano la varietà standard delle lingue incluse nei due corpora; nel caso dell'italiano, essi sono anche rappresentativi della varietà neostandard o dell'uso medio (per cui, cfr. rispettivamente Berruto 1987, 2012 e Sabatini 2011 [1985], così come la discussione in De Cesare *et al.* 2016: 116-120).
- *Autenticità*: i testi raccolti documentano un uso autentico della lingua; in particolare, sono redatti in lingua originale; ciò non esclude che alcune parti di questi testi siano il frutto di una traduzione o adattamento da una o più altre lingue (spesso naturalmente dall'inglese). Si pensi al caso delle citazioni, peraltro frequenti negli articoli dei quotidiani online (su questo punto, si vedano Bonomi *et al.* 2002 per l'italiano e De Cesare *et al.* 2016: 128-134 per una prospettiva che include anche altre lingue). La rielaborazione da un altro testo (in italiano o in un'altra lingua) non è generalmente dichiarata in modo esplicito.
- *Integralità*: tutti i testi sono completi; essi includono anche le principali componenti testuali del paratesto: comprendono la titolazione (composta dal titolo vero e proprio e, facoltativamente, da occhiello, sommario e/o testatina), la firma di chi ha prodotto il testo (giornalista, scrittore, redazione online, agenzia di stampa ecc.), la fonte o le fonti utilizzate per confezionare la notizia (nome della o delle agenzie di stampa) e la data della prima pubblicazione online. Per motivi di copyright, non è possibile dare accesso all'integralità dei due corpora e nemmeno dei singoli testi di cui si compongono.
- *Dimensioni*: il corpus CONTRAST-IT è basato su un campione di testi di ca. 1,5 milioni di parole (per dettagli, cfr. Tabella 1); il corpus COMPARE-IT su un campione di testi di ca. 550.000 parole (si veda il § 3.1 per dettagli). Rispetto ai *mega corpora* di miliardi di parole creati negli ultimi anni (o *in fieri*) con dati ricavati automaticamente dalla rete (come per esempio il corpus TenTen, nella Tabella 1), i corpora CONTRAST-IT e

> COMPARE-IT non sono molto ampi. Si tratta però di corpora molto puliti (che contengono solo testi della stessa tipologia, rappresentativi dell'italiano giornalistico online) e altamente comparabili.

Le due risorse sono composte di testi selezionati manualmente (per dettagli, cfr. i riferimenti nella nota 1); questa procedura, molto costosa in termini di tempo (ma non solo), e con ovvie conseguenze sulla dimensione del corpus, ha il grande vantaggio di garantire il massimo controllo sulla selezione dei dati e di evitare i ben noti problemi legati alla metodologia del *web crawling* (Baroni & Ueyama 2006). Dato il modo in cui CONTRAST-IT e COMPARE-IT sono stati compilati e la loro dimensione, essi assomigliano per certi versi di più ai tradizionali corpora cartacei, anche se i testi provengono dalla rete.

Va infine precisato che sono stati manualmente eliminati dagli articoli raccolti alcuni *boilerplates* tipici dei testi del web e in particolare dei giornali online, come la dichiarazione di copyright e alcuni indirizzi web ricorrenti. Sono state invece mantenute alcune informazioni ritenute importanti a fini della ricerca, e anch'esse ricorrenti, quali la testatina che denota la rubrica della notizia (per es. *politica*), la data di pubblicazione online e la firma dell'articolo (le informazioni che si ripetono nel corpus spiegano alcuni *bias*, come l'alta occorrenza del sostantivo *ottobre* − mese in cui sono stati raccolti molti articoli − e dell'abbreviazione *Red*[azione] *online*).

2.2 Corpora comparabili

I corpora CONTRAST-IT e COMPARE-IT appartengono entrambi alla tipologia dei corpora *comparabili* (per una definizione, cfr. Olohan 2004: 35 e Cresti & Panunzi 2013: 56-57). Le due risorse sono infatti state create a partire da campioni di testi che rispondono agli stessi criteri. Più in particolare, essi condividono le seguenti proprietà testuali, tematiche e semiotiche[7]:

[7] Secondo la distinzione tracciata in Sharoff, Rapp & Zweigenbaum (2013: 3), gli articoli raccolti in CONTRAST-IT e COMPARE-IT rientrano nella categoria dei testi debolmente comparabili (*weakly comparable texts*). In generale, infatti, i testi raccolti nei due corpora descrivono eventi diversi che appartengono allo stesso dominio ("narrow subject domain"). Detto questo, va però anche osservato che molti testi inclusi nei due corpora sono stati raccolti durante lo stesso periodo di tempo e descrivono dunque anche gli stessi eventi mediatici (come la morte di Steve Jobs, di rilevanza internazionale, o il processo di Amanda Knox e Raffaele Sollecito, di interesse soprattutto nazionale).

- *Appartenenza alla stessa tipologia testuale*: i testi raccolti appartengono esclusivamente alla tipologia degli articoli giornalistici pubblicati elettronicamente nell'ambito di quotidiani online; questi testi rientrano nella macro-categoria dei testi informativo-espositivi, la cui funzione pragmatica globale consiste nel trasmettere un sapere (De Cesare 2011); essi rappresentano la lingua scritta trasmessa (Bonomi 2014: 161; De Cesare *et al.* 2016: 114);
- *Organicità tematica*: un gruppo cospicuo di testi è tratto dalle stesse rubriche tematiche, afferenti alla politica, all'economia e allo sport;
- *Data di pubblicazione in rete*: i testi raccolti sono stati pubblicati in rete durante un arco di tempo comune e circoscritto (tra il 2011 e il 2015);
- *Dimensione e bilanciamento*: ampi campioni di testi hanno la stessa dimensione (in termini di numero di parole); questo permette di creare sottocorpora comparabili non solo per dimensione ma anche per rubrica tematica.

CONTRAST-IT e COMPARE-IT si aggiungono ai corpora comparabili già esistenti per studiare l'italiano (per cui si veda l'elenco molto completo e aggiornato di banche dati, corpora e archivi testuali proposto dall'Accademia della Crusca[8] e in Cresti & Panunzi 2013: 200-202), in particolare in prospettiva contrastiva e comparativa. Come vedremo rispettivamente nei §§ 2.2.1 e 2.2.2, essi arricchiscono in vari modi le altre risorse elettroniche a disposizione degli studiosi e consentono di rispondere a nuove domande di ricerca.

2.1.1 *Specificità del* CONTRAST-IT

Per capire le specificità del CONTRAST-IT è utile confrontarlo con gli altri corpora comparabili multilingui disponibili; si veda dunque la Tabella 1 sullo sfondo dell'elenco riportato nella Tabella 2. Si noti che l'elenco della Tabella 2 non è esaustivo: abbiamo tenuto conto unicamente delle risorse elettroniche che includono un sottocorpus in lingua italiana; inoltre, consideriamo solo i corpora liberamente accessibili e interrogabili online[9], composti di testi scritti[10] e redatti

[8] Cfr. http://www.accademiadellacrusca.it/it/link-utili/banche-dati-dellitaliano-scritto-parlato.

[9] L'elenco comprende esclusivamente le riscorse gratuite per chi le usa a fini scientifici e/o didattici; l'accesso ai dati di alcuni corpora prevede una registrazione preliminare.

[10] Un corpus comparabile (di tipo parallelo: per cui, cfr. Gandin 2009) multilingue di testi orali (poi trascritti) è EUROPARL-Direct, che comprende sottocorpora composti di varie coppie di lingue (per l'italiano sono disponibili i *file* 'italiano-francese' e 'francese-italiano'; 'italiano-inglese' e 'inglese-italiano', dove l'italiano è rispettivamente lingua fonte e *target*); il tutto

da nativi[11]. Le due tabelle forniscono indicazioni di massima sulle lingue incluse in ogni corpus; la dimensione complessiva di ogni corpus e almeno anche del sottocorpus italiano; indicano inoltre le tipologie testuali rappresentate in ogni corpus, così come la data di pubblicazione e/o le modalità di raccolta dei testi.

Tabella 1. CONTRAST-IT: corpus comparabile multilingue di testi giornalistici

CONTRAST-IT	https://contrast-it.philhist.unibas.ch/en/corpora/contrast-it-corpus
italiano (Italia), francese (Francia), spagnolo (Spagna), inglese (Regno Unito), tedesco (Germania)	Corpus di testi trasmessi (sul web) Corpus specialistico: testi giornalistici, tratti da quotidiani, pubblicati tra il 2011-2015 Almeno 50% dei testi sono tratti dalle sezioni 'Politica', 'Economia', 'Sport' ± 300.000 parole per sottocorpus Contrast-IT_it: 531 articoli; Contrast-IT_fr: 520 articoli; Contrast-IT_es: 476 articoli; Contrast-IT_en: 404 articoli; Contrast-IT_de: 509 articoli

L'ordine di presentazione dei corpora nella Tabella 2 si basa principalmente su due criteri: il canale di trasmissione dei testi (l'elenco si apre con i corpora di testi pubblicati su carta e si chiude con quelli trasmessi in rete) e la varietà di lingua rappresentata dal corpus in termini di tipologia testuale (consideriamo dapprima i corpora generalisti, poi quelli specialistici).

Tabella 2. Corpora comparabili multilingui di testi scritti (con un sottocorpus d'italiano)

PAROLE	http://www.islrn.org/resources/608-362-291-385-1 (italiano)
italiano, francese (Belgio e Francia), catalano, portoghese, inglese, tedesco, danese, norvegese, svedese, neerlandese, irlandese, greco, finnico	Corpus di testi pubblicati su carta Corpus generalista IT: 3.135.651 parole - testi giornalistici (70%): quotidiani (1992-1996) e periodici (1985-1988); - altre tipologie (30%): libri (1970-1989); miscellanea (1987-1997)

è scaricabile (previa registrazione) dal sito www.idiap.ch/dataset/europarl-direct. Per dettagli sul corpus, cfr. Cartoni, Zufferey & Meyer 2013.

[11] Non teniamo dunque conto qui di corpora comparabili multilingui di testi redatti da non nativi; una risorsa importante in questo campo è VALICO, che comprende testi in italiano L2 confrontabili con testi redatti nella lingua madre degli apprendenti: inglese, tedesco ecc. (cfr. http://www.valico.org/cp.html).

MR. BEAN	http://blog.cbs.dk/mrbean-korpus
italiano e danese	Corpus di testi pubblicati su carta Corpus specialistico: testi narrativi (prodotti sulla base dello stesso input: due film muti della serie televisiva britannica *Mr. Bean*) 27 testi scritti in italiano (ca. 7.800 parole) 20 testi scritti in danese (7.400 parole)
REUTER (RCV2)	https://trec.nist.gov/data/reuters/reuters.html
italiano, francese, spagnolo, spagnolo dell'America Latina, portoghese, tedesco, neerlandese, danese, norvegese, svedese, russo, cinese, giapponese	Corpus di testi pubblicati su carta Corpus specialistico: testi giornalistici (lanci di agenzia) ca. 487.000 notizie Reuters (redatte in lingua originale) Testi pubblicati tra il 20.8.1996 e il 19.8.1997
MLLC - Multilingual and Parallel Corpora	http://www.islrn.org/resources/963-635-729-341-8
italiano, francese, spagnolo, inglese, tedesco, neerlandese	Corpus di testi pubblicati su carta Corpus specialistico: testi giornalistici tratti dalla stampa finanziaria TOT. ca. 93 mio di parole IT: *Sole 24 Ore* (1992); 1,88 mio di parole FR: *Le Monde* (1992-93); 10 mio di parole SP: *Expansion* (1991, 1994); 10 mio di parole IN: *The Financial Times* (1993); 30 mio di parole TE: *Handelsblatt* (1986-88); 33 mio di parole NE: *Het Financieele Dagblad* (1992-1993); 8,5 mio di parole
TenTen Corpora	https://www.sketchengine.eu/documentation/tenten-corpora
+ 30 lingue, tra cui l'italiano	Corpus di testi trasmessi (*web corpus*): scaricati automaticamente dalla rete (tramite SpiderLing) Corpus generalista: testi tratti da siti con estensione '.it', '.fr' ecc. Obiettivo: 10^{10} di parole per ogni lingua IT: itTenTen16 (2016): 4,9 miliardi di parole
NUNC	http://www.bmanuel.org/projects/ng-HOME.html
italiano, francese, spagnolo, inglese, tedesco	Corpus di testi trasmessi Corpus specialistico: CMC, newsgroup 600.000 parole per ogni sottocorpus

WaCky	https://corpora.dipintra.it/public/run.cgi/first_form
italiano, francese, tedesco, inglese	Corpus di testi trasmessi (*web corpus*): scaricati automaticamente dalla rete Corpus generalista: testi tratti da siti con estensione '.it', '.fr' ecc. itWAC: 2 miliardi di parole (tratte dal dominio .it); deWAC: 1,7 miliardi di parole; frWAC: 1,6 miliardi di parole; ukWAC: 2 miliardi di parole
acWaC	https://corpora.dipintra.it/public/run.cgi/first_form
italiano, inglese	Corpus di testi trasmessi (*web corpus*): scaricati automaticamente dalla rete Corpus specialistico: testi accademici acWAC-IT: 1,6 miliardi di parole; acWAC-EN: 74 miliardi di parole
Leipzig Corpora Collection	http://wortschatz.uni-leipzig.de/en/download/
252 lingue	Corpus di testi trasmessi (*web corpus*): scaricati automaticamente dalla rete (tramite Heritrix) Corpus in parte generalista, in parte specialistico (testi tratti da giornali, da pagine casuali del web e da Wikipedia) IT: corpus basato su materiale del 2005-2009; ca. 20 mio di frasi (per 400.000.000 tokens) tratte da giornali pubblicati in Italia, a San Marino e in Svizzera
Swiss SMS corpus	https://sms.linguistik.uzh.ch
italiano (standard), francese (standard), tedesco (standard), altre lingue, varietà e dialetti	Corpus di testi trasmessi Corpus specialistico: sms (CMC) Testi raccolti tra 2009-2010 IT: 1.471 sms; FR: 4.619 sms; TE: 7.287 sms

Tra le principali specificità del corpus CONTRAST-IT, si possono elencare i punti seguenti, tutti indicativi di una grande comparabilità dei dati anche a livello quantitativo (oltre che per il corpus design, per cui si veda sotto). Prima di tutto, ogni sottocorpus monolingue ha una dimensione equiparabile sia in termini di numero di *tokens* sia di parole[12] (cfr. Tabella 3). Per quanto riguarda il primo

[12] Usiamo il termine *parola* (che traduce il termine *word* dell'interfaccia di CONTRAST-IT), ma bisogna precisare che si tratta qui in realtà, più tecnicamente, di *types*, ovvero di forme di parole (Cresti & Panunzi 2013: 111). In questo caso, per intenderci, sono contate come unità distinte parole come *lo, la, i, le, l',* ecc. Non abbiamo dunque chiaramente a che fare con dei lemmi.

dato, relativo al numero di forme grafiche presenti nel corpus (cioè di parole grafiche, ma anche di segni di punteggiatura, simboli ecc.), osserviamo una media di 352.210; i corpora che più si discostano da questo valore (per difetto e per eccesso) sono CONTRAST-IT_fr che ha 12.067 *tokens* in più (questo numero corrisponde a uno scarto dal valore medio di 3,4%) e CONTRAST-IT_es e CONTRAST-IT_en che hanno ca. 8.000 *tokens* in meno (2,3%). Per quanto riguarda il secondo dato, relativo ai *types* (cioè di parole diverse nel corpus), la media è di 298.547; i corpora che maggiormente si discostano da questo valore sono CONTRAST-IT_es, che ha 4924 *types* in meno (1,7%) e CONTRAST-IT_fr che ha 7441 *types* in più (2,5%)[13].

Tabella 3. Dimensione dei sottocorpora di CONTRAST-IT

Sottocorpus	**n. *tokens***	**n. parole**
CONTRAST-IT_es	344.241	293.623
CONTRAST-IT_en	344.127	295.048
CONTRAST-IT_fr	364.278	305.988
CONTRAST-IT_it	360.265	303.510
CONTRAST-IT_de	348.141	294.564

Per quanto riguarda la comparabilità 'quantitativa' dei sottocorpora di CONTRAST-IT, oltre alla media riportata nella Tab. 3 si può anche tenere conto della deviazione standard (ds), una misura di dispersione che permette di avere una stima precisa di quanto i valori si allontanano effettivamente dalla media. Maggiore il valore della ds, maggiore la variabilità nei dati. In CONTRAST-IT, il valore ds relativo ai tokens è 9.432 e dei types, 5.752. Il dato relativo alla ds è dunque modesto (si tratta di poche migliaia di tokens / parole su una media di centinaia di migliaia). Tutto questo − scarti in percentuale riportati nella Tab. 3 e misura di dispersione − ci dice che la variabilità interna dei sottocorpora è contenuta, cioè che i sottocorpora sono simili dal punto di vista quantitativo[14].

Le altre caratteristiche del corpus CONTRAST-IT (sempre rispetto ai corpora presentati nella Tabella 1) riguardano invece aspetti legati alla tipologia e campionatura dei testi; questi aspetti, elencati di seguito, garantiscono la comparabilità dei dati a livello qualitativo.

a. solo testi giornalistici pubblicati in quotidiani elettronici a larga diffusione; molti di questi testi sono pubblicati solo in rete;

[13] Ringrazio Lorenzo Gregori per una parte della descrizione quantitativa dei dati.
[14] Ringrazio Davide Garassino per lo spunto e il calcolo della ds.

b. testi trasmessi sul web, la cui tipologia differisce dalle classiche categorie della CMC (newsgroup, chat, blog, sms; per una discussione, si veda Tavosanis 2011);
c. testi giornalistici prodotti e pubblicati nel terzo millennio (tutti a partire dal 2011);
d. testi tratti da fonti trasparenti e tipologicamente omogenee;
e. testi che rappresentano una sola varietà diatopica di lingua (l'italiano d'Italia, il francese della Francia e via dicendo).

2.1.2 *Specificità del* COMPARE-IT

Il corpus COMPARE-IT costituisce un unicum nel panorama dei corpora attualmente disponibili per lo studio dell'italiano contemporaneo. A nostra conoscenza, è il primo corpus comparabile (liberamente accessibile) che include testi redatti in Italia e fuori d'Italia (al momento include testi redatti in Svizzera e in Canada). Il corpus COMPARE-IT è dunque uno strumento fondamentale per indagare questioni legate alla variazione diatopica dell'italiano contemporaneo scritto di tipo giornalistico e fenomeni di contatto linguistico tra l'italiano e le lingue con le quali l'italiano si trova a interagire a livello locale o nazionale, in particolare il tedesco e il francese (per quanto riguarda la Svizzera) e l'inglese e il francese (per quanto riguarda il Canada). Il corpus è stato creato con un campione di testi altamente comparabili, pubblicati elettronicamente in quotidiani online tra il 2011 e 2013. Le Tabelle 4 e 5 presentano il *design* generale del corpus COMPARE-IT, fornendo informazioni anche sui suoi tre sottocorpora (composti, rispettivamente, di articoli redatti in Italia, in Svizzera e in Canada).

Tabella 4. COMPARE-IT: corpus comparabile monolingue di testi giornalistici in italiano

COMPARE-IT	https://contrast-it.philhist.unibas.ch/en/corpora/compare-it-corpus
italiano d'Italia italiano della Svizzera italiano del Canada	Corpus di testi trasmessi (sul web) Corpus specialistico: testi giornalistici, tratti da quotidiani, pubblicati tra il 2011 e il 2013 ± 550.000 parole complessive

Tabella 5. Dimensione dei sottocorpora di COMPARE-IT

Sottocorpus	**n. parole**	**n. articoli**
COMPARE-IT_IT	300.000	531
COMPARE-IT_CH	150.000	544
COMPARE-IT_CA	100.000	159

3. COMPARE-IT: Comparabilità dei dati

L'obiettivo principale di questa seconda parte del contributo è dimostrare la validità del corpus monolingue comparabile COMPARE-IT per le ricerche linguistiche in chiave comparativa. La validità del corpus si misura in prima istanza in base al grado di comparabilità dei dati disponibili in questa risorsa[15]. Un buon indicatore per valutare il grado di comparabilità dei dati è innanzitutto il lessico, in particolare quello fondamentale. Come sappiamo, infatti, il vocabolario fondamentale di una lingua è quello meno soggetto a variazioni sociolinguistiche: è prodotto stabilmente nei testi, a prescindere da fattori di ordine diatopico, diafasico e perfino diacronico (Chiari & De Mauro 2012: 23).

Per capire meglio la composizione del lessico fondamentale del corpus COMPARE-IT abbiamo analizzato, sia da un punto di vista quantitativo sia qualitativo, varie liste di frequenza generate dal software Sketch Engine[16] a partire dai testi inclusi in questa risorsa.

3.1 COMPARE-IT: analisi del lessico presente nelle tre sezioni del corpus

L'analisi del lessico presente nelle tre sottosezioni del corpus COMPARE-IT si basa sul confronto di tre liste di frequenza, create impostando la ricerca in Sketch Engine nel modo seguente: lempos (che sta per 'lemma + parte del discorso'), 2000 primi risultati. Queste liste ricoprono dunque in larga misura il vocabolario fondamentale dell'italiano, vale a dire i vocaboli di massimo uso che da soli tendono a coprire mediamente ca. l'85% delle occorrenze di tutti i testi scritti e parlati (per cui si veda Chiari & De Mauro 2012, 2014[17]).

---

[15] Provare la comparabilità del corpus CONTRAST-IT attraverso un'analisi lessicale non è invece possibile perché il *tagger* usato per italiano, francese, spagnolo, tedesco e inglese è diverso. In questo caso, ci dobbiamo accontentare dei parametri quantitativi esposti nel § 2.2.1. (per una discussione sulla comparabilità dei dati raccolti in corpora comparabili, in particolare facendo ricorso a test statistici, cfr. Köhler 2013). Una motivazione importante che giustifica invece un'analisi approfondita del solo corpus COMPARE-IT è l'interesse che questo corpus presenta nel panorama attuale degli strumenti d'indagine a disposizione per lo studio dell'italiano contemporaneo scritto in Italia e fuori dai confini italiani.

[16] Questo studio (pilota) è stato condotto prima della creazione della piattaforma web che ospita il corpus COMPARE-IT. Per questo motivo, facciamo riferimento all'interfaccia di Sketch Engine (piuttosto che a quello di NoSketch Engine; su questi aspetti tecnici, cfr. § 2). Per dettagli sul funzionamento di Sketch Engine, cfr. Kilgarriff *et al.* 2004.

[17] La lista del nuovo vocabolario fondamentale dell'italiano contemporaneo è presentata in neretto in https://www.dropbox.com/s/mkcyo53m15ktbnp/nuovovocabolariodibase.pdf?dl=0. Il vocabolario fondamentale non include nomi propri, sigle e acronimi, numeri in cifre e sim-

Le tre liste sono ricavate dai campioni di COMPARE-IT indicati nella Tabella 6. Come è facile osservare, questi dati non sono comparabili a tutti gli effetti: il campione di testi redatti in Canada è più piccolo degli altri due. Dato però che la nostra analisi verte solo sui 2000 lessemi più frequentemente usati nelle tre sottosezioni del corpus e che, come indicato per esempio da Chiari & De Mauro 2012, 2014, questi lessemi sono relativamente stabili da un campione di testi a un altro, la diversa ampiezza dei campioni è da ritenersi poco significativa.

Tabella 6. Dimensione dei campioni usati nell'analisi del lessico (corpus COMPARE-IT)

Sottocorpus	**n. parole**
COMPARE-IT_IT	150.000
COMPARE-IT_CH	150.000
COMPARE-IT_CA	100.000

Le liste di frequenza prodotte con Sketch Engine si prestano a svariate analisi quantitative e qualitative. Per la presente indagine, abbiamo scelto di occuparci dei due aspetti seguenti: (i) il tasso di copertura dei primi 2000 lemmi (§ 3.1.1.) e, dato che tra i primi 2000 lemmi di ogni lista di frequenza compaiono molti lessemi funzionali, con contenuto semantico di tipo istruzionale, (ii) il tasso di copertura delle preposizioni e degli avverbi (§ 3.1.2.). Proponiamo poi un'analisi qualitativa dei dati che entrano nelle tre liste di frequenza relative alle due parti invariabili del discorso, iniziando dalle preposizioni (§ 3.1.3.) per passare poi agli avverbi (§ 3.1.4)[18] .

3.1.1 *Tasso di copertura dei 2000 lemmi più frequenti*

Il primo aspetto di cui teniamo conto è il tasso di copertura dei 2000 lemmi più frequenti (nelle tre liste generate a partire dal corpus COMPARE-IT), vale a dire il numero di occorrenze del corpus coperte dai 2000 lemmi più frequenti. Questa copertura, come mostrano i dati della Tabella 7, è molto simile nei tre campioni (il dato percentuale è calcolato in base al numero di parole totale in ogni sottocorpus: ca. 150.000 in IT-IT/IT-CH e 105.000 in IT-CA).

boli vari. Nelle liste da noi create con Sketch Engine sono invece inclusi i nomi propri, le sigle e gli acronimi (queste due ultime categorie sono trattate perlopiù come nomi).

[18] La stessa indagine è pensabile per altre parti del discorso, quale la congiunzione. La scelta di trattare in modo dettagliato solo la preposizione e l'avverbio è legata al paradigma relativamente circoscritto di lemmi che entrano in entrambe le classi così come agli interessi teorici di chi scrive.

Tabella 7. Tasso di copertura dei primi 2000 lemmi (corpus COMPARE-IT)

Sottocorpus	n. occ. Coperte	tasso di copertura
COMPARE-IT_IT	124.273	82,85%
COMPARE-IT_CH	128.283	85,52%
COMPARE-IT_CA	91.118	86,78%

Lo scarto, pur apparendo come minimo (di 3-4%), di copertura tra IT-IT (ca. 83%) e IT-CH e IT-CA (ca. 86-87%) nel corpus COMPARE-IT (cfr. Tabella 7) va tematizzato. Questi dati indicano infatti che il lessico più frequentemente usato nei quotidiani elettronici redatti e pubblicati in Italia (in questo caso i primi 2000 lemmi) copre complessivamente una parte minore del campione rispetto a quella che copre la stessa fetta del lessico nei quotidiani elettronici redatti e pubblicati in Svizzera e in Canada. Lo scarto tra il campione italiano e svizzero (i due corpora più direttamente comparabili) concerne ca. 4.000 occ.

I dati nella Tabella 8, in cui i 2000 lemmi più frequenti sono suddivisi per fascia, permettono di osservare che sono rispettivamente le fasce "interne" a coprire la maggioranza dei dati nel sottocorpus IT-CH (lo scarto più importante riguarda la fascia 501-1000) e la prima nel caso di IT-CA (sempre rispetto a IT-IT; le fasce che coprono più dati del campione sono in grigio).

Tabella 8. Tasso di copertura (t. d. c.) di varie fasce di lemmi (corpus COMPARE-IT)

	IT_IT		IT_CH		IT_CA	
Ranghi	**n. occ.**	**t. d. c.**	**n. occ.**	**t. d. c.**	**n. occ.**	**t. d. c.**
1-100	72.005	57,21%	71.807	55,66%	52.852	58,02%
101-200	11.005	8,74%	11.382	8,82%	7.774	8,54%
201-300	6.852	5,44%	7.542	5,85%	4.919	5,40%
301-400	5.081	4,04%	5.652	4,38%	3.779	4,15%
401-500	4.138	3,29%	4.630	3,59%	3.039	3,34%
501-1000	13.520	10,74%	14.655	11,36%	9.507	10,44%
1001-2000	13.251	10,53%	13.343	10,34%	9.228	10,13%
TOT.	125.852	100%	129.011	100%	91.098	100%

Per capire meglio la differenza nel tasso di copertura riscontrata nei tre campioni di testi è necessario approfondire l'analisi, tenendo anche conto di aspetti qualitativi. Ci limitiamo qui ad alcune osservazioni di massima legate all'uso dei nomi propri che denotano i tre Paesi in cui i quotidiani del corpus hanno la loro sede (Italia, Svizzera e Canada). La distribuzione dei nomi propri è in Tabella 9.

Tabella 9. Tasso di copertura (n. di occ.) di aluni nomi propri (corpus COMPARE-IT)

	IT_IT	**IT_CH**	**IT_CA**
Lemmi	**n. occ.**	**n. occ.**	**n. occ.**
Italia	333	73	193
Svizzera	9	517	0
Canada	5	17	376
Tot. occ.	347	607	569
t.d.c.	0,23% di 150.000	0,4% di 150.000	0,54% di 105.000

Rispetto al sottocorpus IT-IT, in IT-CH e IT-CA i nomi propri presi in considerazione coprono una parte più ampia del corpus (0,4% e 0,54% vs. 0,23%). A ben guardare, in IT-CA sono principalmente tre i nomi propri che coprono la maggior parte dei dati (*Italia, Canada* e *Toronto*). In particolare, il nome *Italia* ricorre molto più frequentemente in IT-CA (193 occ./105.000 parole) che non in IT-CH (73 occ./150.000 parole). Questo dato si spiega con il fatto che il sottocorpus IT-CH si compone prevalentemente di articoli tratti da rubriche locali, che si occupano della realtà ticinese (e della Confederazione) piuttosto che di quella italiana.

3.1.2 *Tasso di copertura di preposizioni e avverbi nelle liste dei 2000 lemmi più frequenti*

Per capire meglio la composizione del lessico che entra a far parte delle nostre liste di frequenza ci soffermiamo ora su due parti del discorso: la preposizione e l'avverbio.[19] Queste due parti invariabili del discorso si distinguono da vari punti di vista, che in questa sede non è il caso di presentare in dettaglio (per approfondimenti, cfr. De Cesare 2019a). Per lo scopo della presente indagine ci basta indicare che la preposizione coincide con una classe di parole al tempo stesso chiusa (non soggetta ad arricchirsi di nuove entrate lessicali) e circoscritta (comprende poche forme: *di, da, a* ecc.; cfr. *infra* nella Tabella 10), mentre l'avverbio è una classe di parole non solo aperta (è tuttora possibile coniare nuove entrate lessicali, con un processo derivativo che coinvolge il morfema *-mente*), ma anche molto ampia (si pensi al cospicuo gruppo degli avverbi di maniera terminanti in *-mente*; su questa classe semantico-funzionale, cfr. Sgroi 2011). Queste due parti del discorso svolgono poi funzioni sintattiche diverse: le

[19] È doveroso osservare che la nostra indagine verte sull'output prodotto dal *tagging* di Sketch Engine, concepito in base a un'idea relativamente tradizionale delle parti del discorso. Per una discussione teorica sulle parti del discorso, in particolare sulla definizione delle parti invariabili, cfr. rispettivamente Salvi 2013 e De Cesare 2019a.

preposizioni sono una componente basica (strutturale) della costruzione frasale e sintagmatica; gli avverbi si configurano invece perlopiù come costituenti facoltativi, non richiesti dal nucleo frasale. Date le caratteristiche intensionali ed estensionali assai diverse delle due classi di parole di cui ci occupiamo, ci aspettiamo di trovare grande omogeneità nelle tre liste di frequenza soprattutto nel caso delle preposizioni. Inoltre, ci aspettiamo di osservare un diverso tasso di copertura delle preposizioni e degli avverbi (inclusi nelle liste dei 2000 lemmi più frequenti).

Soffermiamoci dapprima sul tasso di copertura di preposizioni e avverbi nelle tre sottosezioni del corpus COMPARE-IT. Prima di commentare i dati (presentati nella Tabella 10), va osservato che i risultati relativi alle preposizioni riguardano unicamente quelle semplici, mentre i risultati relativi agli avverbi riguardano sia gli avverbi semplici sia derivati[20]. Inoltre, diversamente da quanto proposto per calcolare il tasso di copertura dei primi 2000 lemmi, il tasso di copertura delle preposizioni semplici e degli avverbi è calcolato questa volta in base al numero di occorrenze coperte dai 2000 lemmi più frequenti (i dati da considerare questa volta sono dunque quelli riportati nella colonna centrale della Tabella 7).

Tabella 10. Tasso di copertura. Preposizioni semplici e avverbi (corpus COMPARE-IT)

Tag (suffisso)	**Parametri**	**IT_IT**	**IT_CH**	**IT_CA**
Preposizioni	Lemmi	30	30	29
(-i)	N. occ.	16.556	16.327	11.286
	Tasso di copertura	13,3%	12,7%	12,4%
Avverbi	Lemmi	95	104	90
(-r)	N. occ.	6.142	5.527	3.553
	Tasso di copertura	4,9%	4,3%	3,9%

Il tasso di copertura delle preposizioni semplici e, *mutatis mutandis*, degli avverbi che compaiono tra i 2000 lemmi più frequenti delle tre parti del corpus COMPARE-IT è molto omogeneo: si aggira attorno a 12-13% nel caso delle preposizioni e a 4-5% in quello dell'avverbio. Sommando i due dati percentuali, si giunge a un tasso complessivo di ca. 16-18%. Questo risultato è naturalmente

[20] Nelle liste di frequenza prodotte con Sketch Engine, le preposizioni articolate sono taggate in modo diverso, mediante il suffisso -x (non -i); questo suffisso è però applicato a molti altri lemmi (negazione, *si, che*), oltre che ad altre parti del discorso (articoli, definiti e indefiniti), ragione per cui abbiamo deciso di non tenere conto delle preposizioni articolate in questo studio.

atteso: nei tre campioni, hanno maggiore copertura i nomi, i verbi e gli aggettivi. Va però chiarito che il tasso di copertura complessivo delle preposizioni è in realtà più alto: include anche le occorrenze coperte dalle preposizioni articolate.

3.1.3 *Analisi qualitativa delle preposizioni*

Nella Tabella 11 si propone la lista completa delle preposizioni semplici incluse tra i 2000 lemmi più frequenti delle tre sottosezioni del corpus COMPARE-IT; i lemmi compaiono in ordine di frequenza decrescente.

Tabella 11. Tasso di copertura dei primi 2000 lemmi (corpus COMPARE-IT)

IT_IT (30 *types*)	**IT_CH (30 *types*)**	**IT_CA (29 *types*)**
di, a, in, per, con, da, tra, ad, su, dopo, da|di, contro, secondo, senza, fino, rispetto, oltre, fra, verso, sotto, durante, entro, attraverso, nonostante, presso, de, davanti, fin, **intorno**, **pro**	di, in, a, per, con, da, ad, da|di, contro, tra, su, dopo, secondo, fra, senza, fino, rispetto, durante, sotto, verso, oltre, entro, presso, nonostante, de, **sino**, attraverso, davanti, **lungo**, tramite	di, in, per, a, con, da, ad, tra, da|di, su, dopo, durante, secondo, senza, fino, verso, attraverso, contro, oltre, presso, entro, nonostante, tramite, rispetto, sotto, fin, de, fra, davanti

Le tre liste di frequenza contengono sostanzialmente lo stesso paradigma di preposizioni. Per quanto riguarda il numero di *types*, sono 30 in IT-IT e IT-CH e 29 in IT-CA. Tutte le preposizioni incluse nel campione IT-CA si ritrovano anche negli altri due campioni, seguendo in gran parte anche lo stesso rango d'uso, soprattutto in testa di lista. Da notare, il fatto che le due preposizioni "grammaticali" *di* e *a*, che fungono da meri segna-caso e non sono dotate di contenuto semantico proprio (Salvi 2013; De Cesare 2019a), occupano posizioni molto alte nelle tre liste. La preposizione *di* si colloca addirittura in testa assoluta delle tre liste[21].

Tra le tre liste di frequenza vi sono però anche alcune differenze degne di nota e approfondimento. Prima di tutto, si rilevano quattro *hapax legomena* (sono i lemmi in grassetto nella Tabella 11): *intorno* e *pro* compaiono unicamente

[21] La posizione alta nelle tre liste di frequenza delle preposizioni *di* e *a* costituisce un dato empirico importante per una sottoclassificazione della parte del discorso *preposizione*, effettuata sullo sfondo della differenza tra parole "vuote" e "piene": le preposizioni grammaticali, associate a un contenuto istruzionale di tipo appunto grammaticale, si distribuiscono nei dati in modo diverso rispetto alle preposizioni non-grammaticali, dotate di contenuto semantico denotativo (danno indicazioni di tipo spaziale, temporale, ecc.).

in IT-IT; mentre *sino* e *lungo*[22] sono esclusivi della lista IT-CH. Si osserva inoltre che *tramite* non compare in IT-IT (ma è presente negli altri due sottocorpora) e *fin* non è presente in IT-CH (ma compare nei due altri sottocopora). Queste peculiarità distribuzionali sono da imputare in gran parte alla scelta di soffermarsi in modo tassativo sui primi 2000 lemmi più frequenti delle tre sottosezioni del corpus COMPARE-IT. I dati della Tabella 12 permettono infatti di osservare che le preposizioni "speciali", che non compaiono in tutte e tre le liste di frequenza, sono in realtà anch'esse presenti nei campioni ma occupano un rango d'uso più basso.

Tabella 12. Lista delle preposizioni semplici "speciali" (ordinate secondo il rango d'uso decrescente in IT-IT)

	IT_IT		**IT_CH**		**IT_CA**	
	rango	N. occ.	rango	N. occ.	Rango	N. occ.
fin	1701	11	2362	7	1213	11
intorno	1867	10	5525	2	3023	4
pro	1888	10	3077	5	4962	2
lungo	2972	5	1800	10	5106	2
sino	3214	5	1152	17	7764	1
tramite	3247	5[23]	1885	10	862	16

In generale, quando non fanno parte dei 2000 lemmi più frequenti, le preposizioni "speciali" (incluse nella Tabella 12) compaiono tra i 3000 lemmi più frequenti successivi del campione, e fanno dunque con buona probabilità parte della fascia del vocabolario di base detta di *alto uso*: è il caso di *fin* in IT-CH (rango 2362), *pro* in IT-CH e IT-CA (che occupa rispettivamente i ranghi 3077 e 4962), *tramite* in IT-IT (rango 3247); possiamo includere anche *lungo* in IT-IT (rango 2972), che rappresenta un caso limite in IT-CA (in questo campione compare al rango 5106; ricordiamo però che le nostre liste includono nomi propri, sigle e acronimi, rigorosamente esclusi dalle liste del vocabolario di base del NVDB).

Diversi sono invece i casi di *intorno* e soprattutto di *sino*. Si tratta infatti di preposizioni che non compaiono sempre tra i primi 5000 lemmi più frequenti dei campioni analizzati, ma in una delle tre sottosezioni in ranghi ancora più bassi, e

[22] A scanso di equivoci: si tratta qui naturalmente solo del lemma in funzione preposizionale. L'uso aggettivale di *lungo* corrisponde a un lemma diverso.

[23] A parità di frequenza d'uso, come nel caso di *lungo, sino, tramite* in (IT-IT), si dovrebbe tenere conto dell'indice di dispersione.

hanno una distribuzione molto differente nei tre campioni: *intorno* occupa rispettivamente il rango 1867, 3023 e 5525 in IT-IT, IT-CA e IT-CH; *sino*, il rango 1152, 3214 e 7764 in IT-CH, IT-IT e IT-CA. Se si scarta l'ipotesi che la diversa distribuzione di *intorno* e *sino* nei dati è da imputare a fattori contingenti al campionamento (come suggerisce la distribuzione omogena delle altre preposizioni), la differenza interlinguistica venuta alla luce potrebbe indicare preferenze d'uso diverse nei tre campioni d'italiano. Si apre dunque una pista di ricerca interessante in ottica di linguistica comparativa, che possiamo approfondire alla luce di altri dati.

Particolarmente interessante è il caso della preposizione *sino*, che compare tra i primi 2000 lemmi più frequenti in IT-CH (vocabolario fondamentale), tra i 2000 e 5000 lemmi più frequenti in IT-IT (vocabolario di alto uso) e nella lista successiva in IT-CA (vocabolario comune). Per capire meglio le differenze distribuzionali riscontrate nei tre campioni, è utile tenere conto della frequenza della sua variante sinonimica, ovvero *fino* (nel NVDB sia *fino* sia *sino* fanno parte del vocabolario fondamentale): vi sono 99 occ. di *fino* in IT-IT, 96 occ. in IT-CH e 53 occ. in IT-CA. Nei tre campioni, la distribuzione della preposizione *sino* rispetto a *fino* è ben diversa: mentre la variante *fino* è chiaramente preferita in tutti e tre i campioni (da notare la frequenza assoluta molto simile in IT-IT e IT-CH), si osserva che *sino* è una variante solidamente attestata nel campione IT-CH (17 occ. di *sino* rispetto alle 96 occ. di *fino*), ma decisamente marginale in IT-IT (5 occ. di *sino* vs. 99 occ. di *fino*) e IT-CA (1 occ. di *sino* vs. 53 occ. di *fino*).

Un'analisi delle occorrenze di *sino/fino* nel corpus COMPARE-IT permette poi di osservare che entrambe le forme sono perlopiù usate come preposizioni complesse (in funzione intransitiva[24]): sono generalmente seguite dalla preposizione *a*, in alcuni casi anche da *in* (la seconda forma si trova praticamente solo nella locuzione *fino/sino in fondo*). L'analisi delle occorrenze di *fino* nei tre sottocorpora permette però di rilevare un esempio in cui questa forma è usata come preposizione semplice (in funzione transitiva), in cui cioè è direttamente seguita da un sintagma nominale; nel caso in questione, riprodotto al punto (1), potrebbe trattarsi di un refuso ma anche, dato che l'occorrenza è presente nel sottocorpus IT-CA, di un'interferenza con l'inglese (*until that day* > *fino quel giorno*). Il testo in cui compare questo uso peculiare di *fino* + SN è un'intervista a Marina Nevat, scrittrice iraniana trasferitasi a Toronto. Di sicuro, la giornalista del *Corriere Canadese* (Rosanna Bonura) che ha intervistato la scrittrice traduce le sue parole dall'inglese verso l'italiano.

[24] Sul concetto di *preposizione transitiva* e *intransitiva*, vedere Graffi (1994: 47).

(1) Dopo averlo mostrato alle guardie, Marina fu liberata e portata via da Ali in una macchina. **Fino** quel giorno Marina non ebbe idea di quello che accadde agli altri prigionieri, fu detto che probabilmente si trattò di una finta esecuzione. (COMPARE-IT, IT-CA, corriere.com)

L'analisi dei dati relativi alle preposizioni *sino/fino* fa dunque emergere un uso idiosincratico del codice, nel caso dell'italiano elvetico (altri casi messi in luce riguardano la sintassi e il lessico funzionale; per cui, cfr. rispettivamente De Cesare *et al.* 2014b e De Cesare 2017), e il ben noto fenomeno di interferenza linguistica, dovuto al contatto con l'inglese, nel caso dell'italiano "canadese".

3.1.4 *Analisi qualitativa degli avverbi*

Nella Tabella 13 si propone la lista completa degli avverbi inclusi tra i 2000 lemmi più frequenti nelle tre sottosezioni del corpus COMPARE-IT; i lemmi sono anche qui ordinati in frequenza decrescente.

Tabella 13. Lista degli avverbi (corpus COMPARE-IT)

IT_IT (95 *types*)	**IT_CH (104 *types*)**	**IT_CA (90 *types*)**
anche, più, poi, solo, oggi, già, ancora, però, sempre, molto, così, invece, ora, ieri, mai, prima, quindi, meno, bene, quasi, soprattutto, proprio, fuori, infatti, almeno, circa, certo, fa, poco, dunque, qui, appena, tanto, forse, troppo, adesso, grazie, allora, oltre, ben, insieme, pure, soltanto, sì, subito, inoltre, infine, ormai, meglio, no, cioè, intanto, avanti, lì, spesso, addirittura, insomma, via, magari, davvero, direttamente, presto, nemmeno, ecco, male, pur, domani, neanche, tuttavia,	anche, più, oggi, ancora, già, solo, poi, invece, molto, ora, però, sempre, inoltre, così, infatti, meno, ieri, quindi, tuttavia, circa, oltre, soprattutto, troppo, prima, pure, meglio, bene, dunque, poco, attualmente, quasi, sì, grazie, almeno, allora, mai, proprio, fa, qui, domani, infine, tanto, no, ormai, subito, finora, avanti, pur, davvero, forse, nuovamente, piuttosto, ben, appena, soltanto, certo, adesso, maggiormente, spesso, intanto, addirittura, particolarmente, complessivamente, fuori,	anche, più, molto, sempre, solo, così, poi, ancora, proprio, già, oggi, invece, ora, mai, quindi, circa, davvero, qui, ieri, grazie, avanti, fa, inoltre, tanto, soprattutto, però, insieme, prima, oltre, meno, bene, infatti, subito, spesso, almeno, ben, allora, quasi, indietro, sì, assieme, fuori, poco, forse, adesso, via, sicuramente, appena, no, finora, meglio, certo, infine, ovviamente, anzi, presto, pur, dopo, addirittura, dietro, attualmente, domani, ormai, lì, purtroppo, personalmente,

altrimenti, piuttosto, davanti, veramente, certamente, probabilmente, sotto, eppure, attualmente, là, particolarmente, dietro, ovviamente, nuovamente, peraltro, indietro, duramente, rispettivamente, sicuro, anzi, sicuramente, finalmente, decisamente, stavolta, finora, persino	insieme, nettamente, ecco, insomma, finalmente, appunto, semplicemente, male, assieme, sicuramente, fortemente, ovviamente, completamente, leggermente, direttamente, nemmeno, malgrado, recentemente, probabilmente, anzi, rapidamente, ulteriormente, presto, peraltro, pertanto, contro, sotto, chiaramente, altrimenti, lì, dopo, inizialmente, sicuro, stasera, certamente, assolutamente, praticamente, via, solamente, unicamente	semplicemente, abbastanza, ufficialmente, davanti, assolutamente, completamente, estremamente, magari, soltanto, sicuro, solamente, intanto, particolarmente, fortemente, male, certamente, naturalmente, appunto, troppo, perciò, altrimenti, direttamente, maggiormente, cioè

Complessivamente, ci sembra che si possa sostenere che i dati delle tre parti del corpus COMPARE-IT sono in larga misura omogenei (anche se è palese che l'omogeneità dei dati relativi agli avverbi è minore di quella osservata per le preposizioni). Vi è buona omogeneità prima di tutto per quanto riguarda il numero di avverbi più frequenti, inteso in termini di *types*: sono 95 in IT-IT, 104 in IT-CH e 90 in IT-CA. I dati sono simili anche per quanto riguarda la correlazione tra forma e rango d'uso degli avverbi: gli avverbi più frequenti sono morfologicamente semplici; a questo si aggiunge che in testa di lista si trova in modo stabile l'avverbio *anche*, il prototipo della classe dei focalizzatori (per una descrizione di questa classe, cfr. De Cesare 2019a: 90-95). Gli avverbi derivati con il suffisso *-mente* compaiono invece sempre e soprattutto nella seconda parte delle tre liste. Infine, un'osservazione sulla forma degli avverbi: in tutte e tre le liste, gli avverbi derivati sono meno numerosi di quelli semplici: rappresentano il 15% degli avverbi in IT-IT (14/95 occ.), il 25% in IT-CH (26/104 occ.) e il 19% in IT-CA (17/90 occ.).

L'omogeneità dei dati viene meno soprattutto quando si osserva più attentamente la lista degli avverbi in *-mente* presenti nelle tre sottosezioni del corpus COMPARE-IT, proposta nella Tabella 14 (da notare che includiamo nella lista anche *altrimenti*, terminante in *-menti*, variante arcaica del suffisso *-mente*: su questo punto, cfr. De Cesare, Albom & Cimmino 2017: 89, N2).

Tabella 14. Lista degli avverbi in *-mente* (in ordine di frequenza decrescente)

IT_IT (14)	direttamente, altrimenti, veramente, certamente, probabilmente, attualmente, particolarmente, ovviamente, nuovamente, duramente, rispettivamente, sicuramente, finalmente, decisamente
IT_CH (26)	attualmente, nuovamente, maggiormente, particolarmente, complessivamente, nettamente, finalmente, semplicemente, sicuramente, fortemente, ovviamente, completamente, leggermente, direttamente, recentemente, probabilmente, rapidamente, ulteriormente, chiaramente, altrimenti, inizialmente, certamente, assolutamente, praticamente, solamente, unicamente
IT_CA (17)	sicuramente, ovviamente, attualmente, personalmente, semplicemente, ufficialmente, assolutamente, completamente, estremamente, solamente, particolarmente, fortemente, certamente, naturalmente, altrimenti, direttamente, maggiormente

I dati della Tabella 14 permettono di mettere a fuoco le seguenti differenze tra le tre liste: (i) un numero più elevato di avverbi in *-mente* in IT-CH (26 *types* vs. 14 in IT-IT e 17 in IT-CA), possibilmente da spiegare come la conseguenza dell'influsso del francese sull'italiano elvetico (cfr. De Cesare 2016: 460[25]); (ii) un numero relativamente basso di avverbi che si trovano nelle tre liste (gli avverbi in comune sono solo 7; cfr. le forme in grassetto nella Tabella 15); (iii) un numero elevato di *hapax legomena*, cioè di avverbi presenti in una sola lista (4 casi in IT-IT; 10 in IT-CH e 4 in IT-CA, tutti sottolineati nella Tabella 15). Le differenze tra le nostre tre liste non possono essere ricondotte in blocco al campionamento: i sottocorpora IT-IT e IT-CH sono perfettamente comparabili sia per quanto riguarda la quantità dei dati (150.000 parole in ogni sottocorpus), sia le sezioni tematiche rappresentate (Politica, Economia, Sport).

[25] In questo settore della lingua, il tedesco dovrebbe esercitare un influsso minore del francese perché il suffisso ted. *-(er)weise* si usa solo con determinate categorie semantico-funzionali di avverbi: quelli di frase (si veda il caso di *bedauerlicherweise* 'purtroppo') e quelli connettivi (cfr. *beispielsweise* 'per esempio').

Tabella 15. Lista degli avverbi in *-mente* (in ordine alfabetico)

IT_IT (14)	**altrimenti, attualmente**, **certamente,** decisamente, **direttamente,** duramente, finalmente, nuovamente, **ovviamente, particolarmente,** probabilmente, rispettivamente, **sicuramente,** veramente
IT_CH (26)	**altrimenti,** assolutamente, **attualmente**, **certamente,** chiaramente, complessivamente, completamente, **direttamente,** finalmente, fortemente, inizialmente, leggermente, maggiormente, nettamente, nuovamente, **ovviamente, particolarmente,** praticamente, probabilmente, rapidamente, recentemente, semplicemente, **sicuramente,** solamente, ulteriormente, unicamente
IT_CA (17)	**altrimenti,** assolutamente, **attualmente**, **certamente,** completamente, **direttamente,** estremamente, fortemente, maggiormente, naturalmente, **ovviamente, particolarmente,** personalmente, semplicemente, **sicuramente,** solamente, ufficialmente

Per capire le differenze tra le tre liste riportate nella Tabella 15, è utile prendere in considerazione l'elenco degli avverbi in *-mente* che fa ormai parte del vocabolario fondamentale (che compare cioè tra i primi 2000 lemmi più frequenti del NVDB; per una discussione sulla crescita di questi avverbi nel vocabolario del terzo millennio, cfr. De Cesare 2019b). Questo elenco è riportato in Tabella 16, in ordine alfabetico (non abbiamo dati sulla loro frequenza assoluta in quanto l'accesso al corpus NVDB non è libero).

Tabella 16. Lista degli avverbi in *-mente* nel vocabolario fondamentale

NVDB (14)	altrimenti, assolutamente, certamente, chiaramente, completamente, direttamente, effettivamente, esattamente, evidentemente, facilmente, finalmente, immediatamente, lentamente, naturalmente, ovviamente, particolarmente, perfettamente, personalmente, praticamente, probabilmente, semplicemente, sicuramente, solamente, talmente, veramente

Tra le tre liste del corpus COMPARE-IT e quella del NVDB vi sono differenze importanti. Prima di tutto, una notevole quantità di avverbi in *-mente* del NVDB (8 su 25) non compare in nessuna delle tre liste (si tratta di *effettivamente, esattamente, evidentemente, facilmente, immediatamente, lentamente, perfettamente, talmente*); solo 6 avverbi del NVDB si ritrovano in tutte e tre le liste (*altrimenti, certamente, direttamente, ovviamente, particolarmente, sicuramente*). È interes-

sante notare poi che tra i rimanenti avverbi in *-mente* del NVDB vi è più omogeneità con il campione IT-CH (8 avverbi in comune: *assolutamente, chiaramente, completamente, finalmente, praticamente, probabilmente, semplicemente, solamente*) e IT-CA (6 avverbi in comune: *assolutamente, completamente, naturalmente, personalmente, semplicemente, solamente*) che non con IT-IT (solo 3 avverbi in comune: *finalmente, probabilmente, veramente*).

Gli scarti tra le tre liste del corpus COMPARE-IT e il gruppo dei fondamentali del NVDB possono spiegarsi in parte con il fatto che NVDB è un corpus generalista, e mira a registrare un uso "neutro" della lingua, mentre COMPARE-IT è un corpus specialistico, che rappresenta la lingua della stampa quotidiana. Detto questo, quasi tutti gli avverbi in *-mente* che si trovano solo nel NVDB compaiono anche nei nostri campioni, ma in ranghi d'uso più bassi, generalmente tra i 2000 e 5000 lemmi più frequenti: è il caso di *perfettamente* nelle tre liste, ma anche di *immediatamente, esattamente* e *evidentemente* in IT-IT; *effettivamente, facilmente, perfettamente, esattamente, evidentemente* in IT-CH; *facilmente, immediatamente, perfettamente* in IT-CA. Vi sono anche due eccezioni importanti: *talmente* è assente nel campione IT-CH e *lentamente* nel campione IT-CA. Queste assenze possono spiegarsi solo alla luce di un campionamento più ampio di dati, di cui però al momento non disponiamo.

3.2 COMPARE-IT: analisi dei tokens nei sottocorpora IT-IT e IT-CH

Il grado di comparabilità di due sottosezioni del corpus COMPARE-IT, quelle composte di testi redatti in Italia e in Svizzera, può essere valutato anche in base a un altro indicatore: il numero di occorrenze in ogni sottocorpus. Con 'occorrenze' s'intende i *tokens*, ovvero le più piccole unità grafiche: parole (grafiche), segni di punteggiatura, numeri in cifre, sigle, acronimi e simboli vari[26] (gli spazi bianchi non contano). Il numero di *tokens* presente nei due sottocorpora in questione è riportato nella Tabella 17 (i dati sono di nuovo quelli calcolati con Sketch Engine)[27].

Tabella 17. Dimensione dei sottocorpora di COMPARE-IT

Sottocorpus	**n. parole**	**n. *tokens***
COMPARE-IT_IT	150.027	179.912
COMPARE-IT_CH	150.261	181.428

[26] Cfr. Cresti & Panunzi (2013: 87).

[27] Dato che è quantitativamente più ridotto, escludiamo qui dall'analisi i dati relativi al sottocorpus canadese.

I dati quantitativi riportati nella Tabella 17 permettono di osservare che i due sottocorpora di COMPARE-IT, pur essendo composti dallo stesso numero di parole (ca. 150.000), includono un numero lievemente diverso di *tokens*. Lo scarto numerico tra i due sottocorpora non è però molto significativo: vi sono 1.516 *tokens* in più in IT-CH; abbiamo dunque un altro indicatore per affermare che i due sottocorpora di COMPARE-IT presentano un elevato tasso di comparabilità.

La comparabilità dei dati è peraltro confermata da un'analisi più attenta dei *tokens* che differiscono dalle parole grafiche (punteggiatura, simboli vari e numeri in cifre). Alcuni primi dati sono riportati in Tabella 18. Il *tagset* per l'italiano (di Sketch Engine) permette di valutare in modo preciso la frequenza assoluta dei segni interpuntivi in posizione interna (PUN) e finale di frase (SENT); permette inoltre di valutare la frequenza assoluta dei numeri in cifre (NUM) e di una serie di altri elementi non lessicali (raggruppati nella categoria NOCAT, che include per esempio le virgolette uncinate e i simboli [-] e [%])[28].

Tabella 18. Frequenza assoluta dei *tokens* diversi dalle parole grafiche

TAG	IT-IT	IT-CH
PUN	14.131	15.083
SENT	6.574	7.060
NOCAT	5.250	3.152
NUM	3.822	5.926
TOT.	29.777	31.221

La metà dei *tokens* (non consideriamo le parole grafiche) coincide con segni di punteggiatura interni alla frase (PUN). È interessante notare che la frequenza assoluta dei segni interni alla frase è molto simile nei due sottocorpora (lo scarto è solo di ca. 1000 occ. a favore di IT-CH), soprattutto se valutata in base al numero complessivo di occorrenze (ca. 30.000 in IT-IT e 31.000 in IT-CH). Nei due sottocorpora si registra però una differenza nel rango d'uso del *tag* PUN (che non si coglie nei dati della Tabella 18): in IT-IT, questo *tag* occupa il quarto posto nella lista di frequenza dei *tags* (dopo i nomi, le preposizioni e i nomi propri); in IT-CH, PUN occupa invece il terzo rango (dopo i nomi e le preposizioni). Questa differenza non intacca naturalmente il nostro giudizio complessivo sulla comparabilità dei dati; piuttosto, va letta come una spia di un diverso uso della punteggiatura nei due sottocorpora. Questa ipotesi può essere avvalorata da

[28] Il *tagset* italiano di Sketch Engine non prevede un'etichetta speciale per le sigle e le abbreviazioni. Queste forme sono etichettate con altre POS, ed entrano perlopiù nella classe del nome (tag: NOUN).

un confronto più fine dei segni che fanno parte della categoria PUN (per cui si veda la Tabella 19, che riporta la lista di frequenza dei principali segni inclusi nel conteggio PUN).

Tabella 19. Rango d'uso di alcuni segni interpuntivi (PUN)

PUN	IT-IT	Rango	PUN	IT-CH
,	8.448	1	,	8.090
:	1.514	2	"	1.938
"	1.158	3	:	1.684
(	886	4	)	1.156
)	893	5	(	1.150
'	680	6	'	341
;	142	7	;	289
	TOT. 13.721			TOT. 14.648

Nei due sottocorpora di COMPARE-IT vi è un ordinamento praticamente parallelo dei segni PUN: il segno in assoluto più frequente è la virgola (rango 1), quello meno frequente il punto e virgola (rango 7); occupano lo stesso rango d'uso anche le parentesi tonde (ranghi 4-5) e l'apostrofo (rango 6). Nei due sottocorpora, si registra invece una differenza nel rango d'uso dei due punti (rango d'uso 2 in IT-IT e 3 in IT-CH) e delle virgolette alte (rango d'uso 3 in IT-IT e 2 in IT-CH). A ben guardare, la differenza tra i due sottocorpora riguarda però sostanzialmente l'uso delle virgolette, poiché i due punti hanno una frequenza d'uso molto simile (1.514 occ. in IT-IT e 1.684 occ. in IT-CH). Sensibilmente diversa nei due sottocorpora è invece la frequenza assoluta delle virgolette alte (1.158 in IT-IT e 1.938 in IT-CH, con uno scarto di 780 occ.). Questa differenza è dovuta alle scelte grafiche delle due testate: in IT-IT è molto più frequente l'uso delle virgolette uncinate (ca. 1425 occ., contro ca. 340 in IT-CH).

Se si osserva più da vicino la frequenza assoluta dei segni PUN, è facile rilevare altre differenze importanti. Si registra per esempio uno scarto importante nella frequenza d'uso della virgola, più frequente in IT-IT che in IT-CH (+ 358 occ. nel primo sottocorpus). Oltre alle virgolette, di cui abbiamo già parlato, due altri segni sono più frequenti nel sottocorpus IT-CH: le parentesi (con uno scarto di 270 occ.) e il punto e virgola (ben due volte più frequente in IT-CH). Dal dato sulla frequenza d'uso delle parentesi, si evince che i testi del sottocorpus IT-CH racchiudono più spesso di quelli del sottocorpus IT-IT informazioni secondarie tra parentesi tonde. Una ricerca più approfondita sull'uso delle parentesi tonde permette anche di individuare una differenza qualitativa notevole tra i due sotto-

corpora: nel sottocorpus IT-CH le parentesi racchiudono molto spesso delle sigle, come nel caso riportato in (2), o viceversa (ma più di rado) lo scioglimento di sigle, come in (3); queste due configurazioni interpuntive sono praticamente assenti nel sottocorpus IT-IT, dove le parentesi racchiudono perlopiù commenti metalinguistici e dati percentuali.

(2) secondo l'Ufficio federale dell'agricoltura (UFAG) dovrebbero permettere un aumento della […] (COMPARE-IT, IT-CH; cdt.ch, sezione "Politica")

(3) A premere per un addio al Libor è soprattutto la Cftc (Commodity Futures Trading Commission, l'autorità per i future e i derivati), secondo la quale […] (COMPARE-IT, IT-CH; cdt.ch, sezione "Economia")

4. Osservazioni conclusive

In questo contributo abbiamo presentato i corpora comparabili CONTRAST-IT e COMPARE-IT, due nuovi strumenti di ricerca per condurre indagini *corpus-based* e/o *corpus-driven*. Queste due nuove risorse sono state pensate in primo luogo per promuovere la ricerca sull'italiano contemporaneo in prospettiva contrastiva e/o comparativa.

Nella prima parte di questo contributo abbiamo descritto il design generale di queste risorse elettroniche e chiarito in che modo si distinguono dai corpora comparabili (liberamente accessibili) già esistenti. Abbiamo anche accennato al fatto che il corpus multilingue CONTRAST-IT si presta a molteplici ricerche sulle caratteristiche linguistiche (lessicali, ma anche sintattiche e morfologiche), interpuntive e testuali dell'italiano (giornalistico) odierno in prospettiva contrastiva con altre lingue europee (BIB?). Integrando i dati del corpus CONTRAST-IT con quelli di altri corpora disponibili (elencati in Tabella 1), è ormai anche possibile tracciare la diffusione di un neologismo nella stampa elettronica di diversi Paesi europei (anche dopo il 2010), indagare se vi sono differenze "nazionali" tra la lingua dei quotidiani pubblicati su carta e in rete, o semplicemente controllare la validità dei risultati ottenuti in un'indagine basata sul corpus CONTRAST-IT. Il corpus COMPARE-IT si presta a sua volta ad approfondire questioni di contatto linguistico, in particolare legate a fenomeni d'interferenza tra l'italiano e l'inglese, il francese e il tedesco; questo corpus − che costituisce attualmente un unicum nel ricco inventario di riscorse elettroniche disponibili per studiare l'italiano − permette anche di capire se vi sono norme linguistiche o

stilistiche diverse nei paesi in cui l'italiano è usato per iscritto, in particolare in ambito giornalistico.

Nella seconda parte del contributo ci siamo soffermati sul corpus COMPARE-IT. L'obiettivo principale era mostrare la grande comparabilità dei dati inclusi nelle tre parti del corpus e dunque, complessivamente, la qualità di questa risorsa per la ricerca linguistica. La comparabilità dei dati è stata valutata in base all'omogeneità di vari indicatori (perlopiù quantitativi), relativi al lessico e alla punteggiatura. Alcuni indicatori sono naturalmente più affidabili di altri: pensiamo al tasso di copertura dei 2000 lemmi più frequenti o ancora al tipo di preposizione che compare tra i 2000 lemmi più frequenti. Altri sono più soggetti a variare da un testo all'altro: nel caso del numero e del tipo di *tokens* diversi dalle parole grafiche abbiamo comunque potuto osservare una grande omogeneità dei dati nei testi giornalistici redatti in Italia e in Svizzera.

Nel corso dell'indagine comparativa dei campioni di COMPARE-IT abbiamo anche registrato alcune differenze degne di nota e di approfondimento. Queste differenze riguardano in particolare l'uso degli avverbi in *-mente*, di alcune preposizioni (soprattutto *sino* rispetto alla variante sinonimica, e meno marcata, *fino*) e di alcuni segni interpuntivi (virgolette e parentesi). Scartata l'ipotesi che queste differenze siano il prodotto di un campionamento "difettoso", in particolare non omogeno, abbiamo cercato di formulare alcune prime risposte esplicative. Per capire se le differenze riscontrate nei campioni analizzati sono il sintomo di usi diversi del codice lingua, eventualmente anche legate a diverse norme d'uso (in linea per esempio con le proposte di Ferrari *et al.* 2009 e De Cesare 2016), se sono la conseguenza di fenomeni di contatto linguistico o se sono spiegabili in altri modi ancora, è necessario indagare più a fondo ognuna di queste differenze. Bisogna dunque proporre altre ricerche empiriche (sia quantitative sia qualitative) ed eventualmente considerare altri dati, tratti da corpora simili nel design a COMPARE-IT.

Altrettanto importante, ed è su questo punto che ci preme insistere in chiusura, è continuare a proporre nuovi strumenti di ricerca per lo studio dell'italiano contemporaneo e lavorare all'ampliamento delle risorse già disponibili, ma spesso di dimensioni piuttosto modeste (soprattutto se confrontate con i mega corpora dell'ultima generazione). Tempo e risorse permettendo, il nostro intento è di ampliare i due corpora in varie direzioni. Il corpus CONTRAST-IT potrebbe essere arricchito di nuove sottosezioni, comprendenti testi redatti per esempio in altre lingue romanze (romeno, portoghese). Per il corpus COMPARE-IT si prevede invece di includere testi tratti da quotidiani redatti in italiano in altre nazioni (San Marino, Città del Vaticano, Croazia, Stati Uniti d'America, Argentina, Uruguay ecc.). Un corpus comprendente testi redatti in italiano in diverse aree

geografiche permetterebbe per esempio di sondare la manifestazione dei tratti neo-standard (e dell'uso medio) in parti del mondo anche molto distanti. Inoltre, un corpus con sottosezioni di testi redatti nei Paesi in cui l'italiano ha lo statuto di lingua (semi-)ufficiale (Italia, Svizzera, Croazia e Slovenia) permetterebbe di indagare un'altra questione fondamentale (e relativamente nuova) in campo sociolinguistico: l'idea che l'italiano sia una lingua debolmente pluricentrica, vale a dire una lingua che ha elaborato propri standard nazionali leggermente divergenti nei Paesi in cui l'italiano è appunto lingua (semi-)ufficiale (su queste questioni, cfr. Berruto 2011, Pandolfi 2017, Moretti & Pandolfi 2019; per un'indagine sulla sintassi marcata, cfr. De Cesare *et al.* 2014b).

Riferimenti bibliografici

Baroni, M. & Ueyama, M. 2006. Building general- and special-purpose corpora by Web crawling. Proceedings of the 13th NIJL International Workshop on Language Corpora: Their compilation and application. Tokyo: NIJL, 31-40.

Berruto, G. 1987. *Sociolinguistica dell'italiano contemporaneo*. Roma: Carocci.

Berruto, G. 2011. Italiano lingua pluricentrica? In A. Overbeck, W. Schweickard & H. Völker (eds), *Lexicon, Varietät, Philologie. Romanistiche Studien, Günter Holtus zum 65. Geburtstag*. Berlin-New York: de Gruyter Mouton, 14-26.

Berruto, G. ²2012. *Sociolinguistica dell'italiano contemporaneo*. Roma: Carocci.

Bonomi, I. 2014. L'italiano giornalistico dalla carta al web: costanti e novità. In E. Garavelli & E. Suomela-Härmä (eds), *Dal manoscritto al web: canali e modalità di trasmissione dell'italiano. Tecniche, materiali e usi nella storia della lingua*, *Atti del XII Congresso SILFI* (Helsinki, 18-20 giugno 2012), vol. 2. Firenze: Cesati, 161-178.

Bonomi, I. et al. 2002. La lingua dei quotidiani on line. In I. Bonomi (ed.), *L'italiano giornalistico. Dall'inizio del '900 ai quotidiani on line*. Firenze: Cesati, 267-349.

Cartoni, B., Zufferey, S. & Meyer, T. 2013. Using the Europarl corpus for cross-linguistic research. In M.-A. Lefer & S. Vogeleer (eds), *Interference and normalization in genre-controlled multilingual corpora*. Amsterdam/Philadelphia: Benjamins, 23-42.

Chiari, I. & De Mauro, T. 2012. The new basic vocabulary of Italian: problems and methods. *Statistica applicata* 22: 21-35.

Chiari, I. & De Mauro, T. 2014. The New Basic Vocabulary of Italian as a linguistic resource. In R. Basili, A. Lenci & B. Magnini (eds), *First Italian Conference on Computational Linguistics* 2014 (9-10 December 2014). Pisa: Pisa University Press, 113-116.

Cresti, E. & Panunzi, A. 2013. *Introduzione ai corpora d'italiano*. Bologna: il Mulino.

De Cesare, A.-M. 2011. Espositivi, testi. In R. Simone (dir., con la collab. di G. Berruto & P. D'Achille), *Enciclopedia dell'italiano*, vol. 2. Roma: Treccani, 1474-1478.

De Cesare, A.-M. 2016. L'italiano giornalistico della Svizzera (italiana): caratteristiche morfosintattiche. In B. Moretti, E. Maria Pandolfi, S. Christopher & M. Casoni (eds), *L'italiano in Svizzera* [SILTA XLV/3 2017], 453-464.

De Cesare, A.-M. 2017. Per un altro tassello dell'italiano come lingua (debolmente) bicentrica: l'uso di *pure* e *neppure* nell'italiano giornalistico d'Italia e della Svizzera

italiana. In B. Moretti, E. M. Pandolfi, S. Christopher & M. Casoni (eds), *Linguisti in contatto 2. Ricerche di linguistica italiana in Svizzera. Atti del convegno dell'Osservatorio linguistico della Svizzera italiana* (Bellinzona, 19-21 novembre 2015). Bellinzona: Osservatorio linguistico della Svizzera italiana, 146-159.

De Cesare, A.-M. 2019a. *Le parti invariabili del discorso.* Roma: Carocci.

De Cesare, A.-M. 2019b. Sulla crescita degli avverbi in *-mente* nel vocabolario fondamentale. Dall'italiano del secondo al terzo millennio. In B. Moretti, A. Kunz, S. Natale & E. Krakenberger (eds), *LII Congresso SLI «Le tendenze dell'italiano contemporaneo rivisitate»*. Milano, Officinaventuno: 203-220.

De Cesare, A.-M., Garassino, D., Agar Marco, R., Baranzini, L. 2014a. Form and frequency of Italian Cleft Constructions in a Corpus of Electronic News. A Contrastive Perspective with French, Spanish, German and English. In A.-M. De Cesare (ed.), *Frequency, Forms and Functions of Cleft Constructions in Romance and Germanic. Contrastive, Corpus-based Studies*. Berlin: de Gruyter Mouton [*Trends in Linguistics* 281], 49-99.

De Cesare, A.-M., Garassino, G, Agar Marco, R., Albom A. & Cimmino, D. 2014b. L'italiano come lingua pluricentrica? Riflessioni sull'uso delle frasi sintatticamente marcate nella scrittura giornalistica online. *Studi di grammatica italiana* XXXIII: 295-363.

De Cesare, A.-M., Garassino, G, Agar Marco, R., Albom A. & Cimmino, D. 2016. *Sintassi marcata dell'italiano dell'uso medio in prospettiva contrastiva con il francese, lo spagnolo, il tedesco e l'inglese. Uno studio basato sulla scrittura dei quotidiani online*. Frankfurt am Main: Peter Lang.

De Cesare, A.-M., Albom, A. & Cimmino, D. 2017. Avverbi in -MENTE nelle lingue romanze e didattica dell'intercomprensione. *Studi Italiani di Linguistica Teorica e Applicata* XLV(1): 61-94.

Ferrari, A. et al. 2009. La lingua dei quotidiani ticinesi. In B. Moretti, E. M. Pandolfi & M. Casoni M. (eds), *Linguisti in contatto. Ricerche di linguistica italiana in Svizzera. Atti del convegno dell'Osservatorio linguistico della Svizzera italiana* (Bellinzona, 16-17 novembre 2007). Bellinzona: OLSI, 281-298.

Gandin, S. 2009. Linguistica dei corpora e traduzione: definizioni, criteri di compilazione e implicazioni di ricerca dei corpora paralleli. *AnnalSS* V: 133-152.

Graffi, G. 1994. *Sintassi*. Bologna: il Mulino.

Kilgarriff, A. et al. 2004. The Sketch Engine. In G. Williams & S. Vessier (eds), *Proceedings of Eleventh EURALEX International Congress*. Lorient, Faculté des Lettres et des Sciences Humaines: Université de Bretagne Sud, 105-116.

Köhler, R. 2013. Statistical Comparability: Methodological Caveats. In S. Sharoff, R. Rapp, P. Zweigenbaum & P. Fung (eds), *Building and Using Comparable Corpora*. Springer: Berlin/Heidelberg, 77-91.

Moretti, B. & Pandolfi, E. M. 2019. Standard svizzero vs. standard italiano. In T. Krefeld & R. Bauer (eds), *Lo spazio comunicativo dell'Italia e delle varietà italiane*, Versione 44. In: Korpus im Text. url: http://www.kit.gwi.uni-muenchen.de/?p=12725&v=1.

Olohan, M. 2004. *Introducing Corpora in Translation Studies*. Routledge: London & New York.

Pandolfi, E. M. 2017. Italian in Switzerland: the dynamics of polycentrism. In M. Cerruti, C. Crocco, S. Marzo (eds), *Towards A New Standard: Theoretical and Empirical Studies on the Restandardization of Italian*. Berlin: Mouton de Gruyter, 321-362.

Rychlý, P. 2007. Manatee/Bonito - A Modular corpus Manager. In P. Sojka & A. Horák (eds), *1st Workshop on Recent Advances in Slavonic Natural Language Processing.* Brno: Masaryk University, 65-70.

Sabatini, F. 2011 [1985]. 'L'italiano dell'uso medio': una realtà tra le varietà linguistiche italiane. In V. Coletti et al. (eds), *L'italiano nel mondo moderno. Saggi scelti dal 1968 al 2009*. Tomo II. Napoli: Liguori, 3-36. [originariamente: In Holtus, G. & Radtke, E. (eds), *Gesprochenes Italienisch in Geschichte und Gegenwart*. Tübingen: Narr, 154-184].

Salvi, G. 2013. *Le parti del discorso*. Roma: Carocci.

Sharoff, S., Rapp, R. & Zweigenbaum, P. 2013. Overviewing important aspects of the last twenty years of research in Comparable Corpora. In S. Sharoff, R. Rapp, P. Zweigenbaum & P. Fung (eds), *Building and Using Comparable Corpora*. Springer: Berlin/Heidelberg, 1-17.

Sgroi, S. C. 2011. Maniera, avverbi di. In R. Simone (dir, con la collab. di G. Berruto & P. D'Achille), *Enciclopedia dell'italiano*. Treccani: Roma, 849-850.

Tavosanis, M. 2011. *L'italiano del web*. Roma: Carocci.