LIMITI E POTENZIALITÀ DELL'USO DI DATI EMPIRICI IN LESSICOGRAFIA

Il caso del plurale delle parole composte

M. Silvia MICHELI

ABSTRACT • The goal of this paper is twofold: on the one hand, it deals with the main methodological problems related to the study of Italian compound words using linguistic corpora; on the other hand, it aims at evaluating to what extent data extracted from two corpora of contemporary Italian, which are very different in size and content, can be used to improve Italian dictionaries in regards to the inflection of Italian compounds, about which speakers often have linguistic doubts.

KEYWORDS • Compounds, Corpus Linguistics, Inflection, Lexicography

1. Introduzione

Il contributo che il dato empirico può offrire a una risorsa lessicografica riguarda tipicamente la costruzione del lemmario, la composizione delle glosse e la loro articolazione in accezioni: su questi elementi si è concentrata la riflessione sull'uso di dati estratti da corpora in lessicografia. Poca attenzione è stata finora dedicata al contributo che i dati posso fornire per migliorare le informazioni di natura morfologica che accompagnano i lemmi di un dizionario, su cui non di rado si concentrano i dubbi dei parlanti, rispetto ai quali l'intuizione non è efficace: come osservato da Iacobini, Giuliani (2001: 332), infatti, in morfologia il giudizio del parlante non è dirimente nello stabilire se una parola è possibile o attestata. In particolare, la formazione del plurale di determinate parole – che per diverse ragioni possono costituire delle eccezioni alle tendenze generali[1] – rappresenta un motivo molto frequente per cui un parlante interroga un dizionario. Un interessante, quanto complesso, caso di questo tipo è costituito delle parole composte, una categoria del lessico italiano che negli ultimi decenni è stata oggetto di numerosi contributi teorici[2], orientati prevalentemente alla definizione della categoria e alla classificazione dei suoi membri. Nella riflessione sulla natura dei composti, l'osservazione della flessione

[1] Si pensi ad esempio al plurale delle parole in *-cia* e *-gia* o, come si vedrà in questo contributo, a quello delle parole composte.

[2] Rendere conto della vastissima bibliografia sulle parole composte in italiano non rientra negli obiettivi di questo contributo, che si concentra piuttosto sulle metodologie con cui questo tipo di parole può essere studiato; oltre ai lavori su specifiche tipologie di composto a cui si farà riferimento nel corso dell'articolo, si rimanda a Masini, Scalise (2012) e Iacobini, Thornton (2016: 209-210) per un quadro generale dello stato dell'arte sulla composizione italiana.

costituisce un aspetto particolarmente rilevante, spesso sottovalutato, perché può aiutare a comprendere come i parlanti concepiscono questo tipo di parole[3] – come entità lessicalizzate e immagazzinate nel lessico mentale o come entità trasparenti e scomponibili nelle proprie parti – e quale grado di analizzabilità conservano nella coscienza linguistica.

Requisito fondamentale per l'analisi di questo fenomeno è l'osservazione di dati empirici, di natura sia qualitativa sia quantitativa, la cui raccolta pone tuttavia numerosi problemi metodologici legati a specifiche caratteristiche formali e semantiche proprie delle parole composte. È su questo aspetto che il presente contributo intende soffermarsi, con l'obiettivo di valutare in che modo, ed entro quali limiti, i risultati ottenuti da un'indagine empirica possono essere considerati realmente *dati*, su cui basarsi per integrare o mettere in discussione quanto riportato dalle risorse lessicografiche: allo stato attuale, infatti, i dizionari non sempre si dimostrano strumenti efficaci per i lettori, in alcuni casi perché riportano informazioni discordanti, in altri perché registrano più forme plurali, senza indicare quale sia effettivamente la variante più frequente nell'uso.

Attraverso l'analisi di alcuni *case studies*, si intende quindi proporre una riflessione sul rapporto tra dati empirici e risorse lessicografiche[4]; in particolare, dopo aver discusso quali risorse permettono, o non permettono, di raccogliere dati affidabili e rappresentativi, si metterà a confronto quanto riportato da due dizionari – il Devoto Oli e il Gradit – rispetto alla formazione del plurale di alcune tipologie di composto, e si analizzerà in che misura le informazioni ricavate da due corpora di italiano contemporaneo – itWaC e il corpus del Nuovo Vocabolario di Base – possono integrarlo o metterlo in discussione.

2. Questioni preliminari

Un primo elemento che caratterizza le parole composte e ne rende più complesso il loro studio attraverso i corpora riguarda la bassa frequenza con cui queste ricorrono nell'uso reale dei parlanti: esse rappresentano un insieme di entità lessicali mediamente rare per ragioni di tipo semantico e pragmatico, in quanto veicolano un insieme di significati a cui i parlanti ricorrono solo in determinati contesti o domini testuali. Considerando ad esempio i composti Verbo+Nome – tipologia particolarmente produttiva in italiano – si osserva che essi vengono generalmente utilizzati in funzione agentiva o strumentale: spesso individuano strumenti o figure professionali difficili da incontrare al di fuori di particolari ambiti specialistici o contesti (si pensi ad esempio a *provavalvole* o *narrastorie*). Questo vale in massima parte anche per le altre tipologie di composto: le forme del tipo Aggettivo+Aggettivo (Grossmann, Rainer 2009; D'Achille, Grossmann 2009), ad esempio, si formano quasi esclusivamente per indicare i giocatori di una squadra di calcio (ad es. *bianconero*), l'appartenenza a uno schieramento politico (ad es. *nazionalsocialista*) o la provenienza geografica (ad es. s*erbocroato*). La bassa frequenza con cui le parole composte compaiono nei testi è quindi dovuta in primo luogo a ragioni pragmatiche: nella maggior parte dei casi, esse sono create dai parlanti per nominare concetti o oggetti che non

[3] Su come i composti vengono processati dai parlanti sono disponibili numerosi studi di psicolinguistica: per un quadro generale rimando al volume curato da Libber, Jarema (2007).

[4] I casi particolari che si discuteranno nei paragrafi successivi non sono ovviamente da considerarsi rappresentativi e/o esplicativi in modo esaustivo del fenomeno, quanto piuttosto funzionali a mettere in luce i problemi metodologici legati allo studio dei composti su basi empiriche: l'obiettivo centrale del contributo non è quindi descrivere la formazione del plurale nelle parole composte in italiano, ma discuterne le possibili metodologie di indagine ed evidenziare limiti e potenzialità del contributo che due corpora di italiano contemporaneo possono fornire ai dizionari.

si trovano nell'orizzonte quotidiano degli individui. A tale questione se ne aggiunge un'altra legata più in particolare allo studio del plurale di un sostantivo, non necessariamente composto: nell'uso reale dei parlanti la distribuzione tra singolare e plurale è determinata da fattori difficilmente prevedibili perché strettamente legati al contesto in cui viene prodotto l'enunciato, e molto spesso è sbilanciata a favore del singolare. Questo rende ancora più arduo riuscire a estrarre dai corpora dati quantitativamente significativi relativi alle forme flesse.

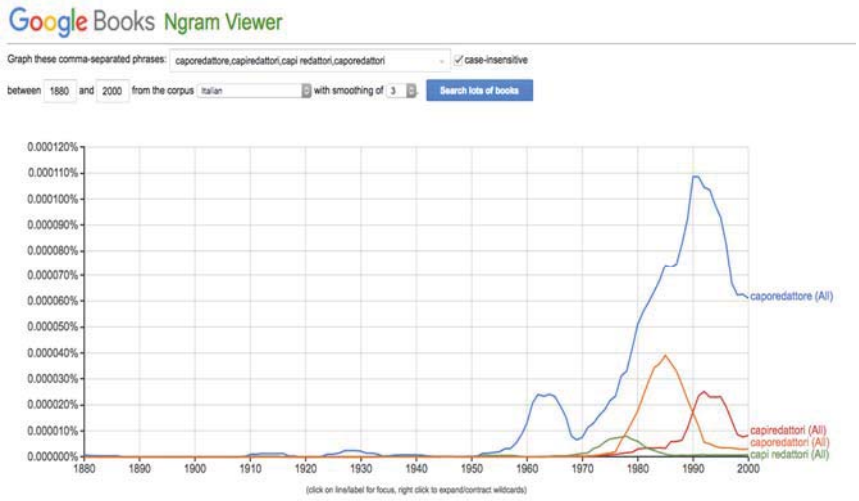
3. I dati empirici: fonti, problemi, metodi

La difficoltà con cui i composti vengono intercettati dai corpora potrebbe indurre a ritenere efficace l'utilizzo del web come corpus da cui estrarre dati empirici, in virtù delle enormi dimensioni e del contenuto eterogeneo. L'uso dei motori di ricerca, come Google o Yahoo, per interrogare il web nell'ambito di analisi linguistiche è stato – e, anche se in misura minore, è ancora – un tema a lungo dibattuto in letteratura (cfr. Kilgarriff, Grafenstette 2003; Crystal 2006; Lüdeling *et al*. 2007; Kilgarriff 2007): la maggior parte degli studiosi intervenuti nel dibattito si è detta contraria all'utilizzo di Google in linguistica, evidenziandone i numerosi limiti e mettendone in discussione la scientificità[5]. Le argomentazioni con cui i detrattori dell'uso di Google hanno sostenuto la loro posizione discendono tutte da una caratteristica intrinseca e ineliminabile del web: il suo contenuto cambia continuamente ed è perciò di fatto inconoscibile per l'utente; ne consegue che le interrogazioni nei motori di ricerca non sono replicabili e le informazioni che se ne possono estrarre non sono confrontabili e quindi, semplicemente, non possono essere definiti *dati*[6], sulla base dei quali elaborare/verificare teorie linguistiche o migliorare il contenuto di grammatiche e dizionari. Le grandi potenzialità della rete possono piuttosto essere sfruttate per creare corpora di grandi dimensioni, come testimonia il progetto WaCky, nell'ambito del quale è nato il webcorpus di italiano contemporaneo itWaC. Nonostante i numerosi problemi metodologici, l'uso di Google in linguistica gode comunque di una certa vitalità: nell'ambito degli studi di morfologia, la sua validità come strumento di ricerca è stata sostenuta da Hathout *et al*. (2008) e Montermini (2015)[7]. In particolare, nei due contributi gli autori sostengono che, con le dovute cautele, il web possa essere utilizzato proficuamente per estrarre dati quantitativi[8] al fine di valutare la produttività di meccanismi di formazione delle parole o la portata di fenomeni morfologici anche rari e quindi difficilmente osservabili attraverso i corpora tradizionali. Costituirebbe inoltre uno degli aspetti più interessanti del web la possibilità di analizzare la creatività lessicale in contesti informali e spontanei quali forum, blog, etc. (ivi: 72): tuttavia, questo non sembra un valido argomento per sostenerne la maggiore efficacia rispetto ai corpora tradizionali, dal momento che esistono corpora costituiti anche da pagine di forum o blog, come il già citato itWaC, o pensati appositamente per lo studio della Comunicazione Mediata dal Computer (CMC), come uno dei sottocorpora che costituiscono il Nuovo Vocabolario di Base[9], o Web2Corpus_IT (Beißwenger *et al*. 2016; Chiari 2016).

[5] Per una più dettagliata disamina dei limiti dei motori di ricerca si rimanda a Kilgarriff (2007).
[6] Per una delimitazione del concetto di dato linguistico si vedano Iannàccaro (2000) e Lehmann (2004).
[7] Nei due contributi l'uso del web come fonte di dati è applicato allo studio di alcuni suffissi del francese e dell'italiano.
[8] Dati di cui, pur riconoscendone la diversità rispetto a quelli estratti dai corpora tradizionali, si sostiene la affidabilità nell'ambito di un'indagine scientifica: «In no case should this 'new' kind of data be taken as a weakness for the theories deduced from it» (Hathout *et al*. 2008: 82).
[9] Per la descrizione del corpus si veda il paragrafo 5.

Si fondano su informazioni ricavate da Google le considerazioni di Montermini (2008) sulla formazione del plurale dei composti Nome+Nome e Verbo+Nome: l'analisi si basa su un campione di composti di dimensioni molto ristrette del quale viene osservata la flessione interrogando il motore di ricerca, facendo precedere ciascuna forma dal proprio articolo determinativo (quindi *i*/*gli* per i composti maschili, *le* per i femminili)[10], al fine di osservare quali siano i pattern più diffusi e se fattori come il genere e la vocale finale dei costituenti abbiano un ruolo nella formazione del plurale. Oltre alla ristrettezza del campione esaminato, un limite fortissimo al lavoro nel suo insieme è costituito da due aspetti metodologici: *in primis*, va considerato che anteponendo l'articolo determinativo flesso a tali forme di fatto vengono esclusi dai risultati i casi in cui i composti sono preceduti da un articolo indeterminativo o da nessun articolo; inoltre, i valori numerici forniti (ivi: 174) si riferiscono al numero delle pagine web indicizzate da Google nel momento in cui è stata effettuata la ricerca[11], non al numero delle occorrenze.

I vantaggi offerti dalla rete – dimensioni, eterogeneità di contenuto, facilità di interrogazione – si rivelano quindi solo apparenti se commisurati ai numerosi problemi che conseguono dal suo utilizzo nell'ambito di indagini linguistiche di natura empirica, per le quali, è bene ribadirlo, gli strumenti scientificamente più validi e affidabili sono i corpora. Come osservato da Kilgarriff (2007), il contributo che il web può dare agli studi linguistici è quindi limitato alla costruzione di corpora costituiti da testi provenienti dalla rete, come quelli del progetto WaCky (Baroni, Kilgarriff 2006), o il corpus di Google Ngram Viewer (Michael *et al*. 2012)[12]. In particolare, nell'ambito di uno studio sulla formazione del plurale delle parole composte, tale risorsa può essere utilizzata per verificare, in diacronia, la presenza di forme plurali concorrenti nell'uso dei parlanti, in modo da integrare i dati estratti dai corpora, che fotografano il fenomeno solo sul



[10] Montermini (2008:166; nota 8).

[11] Questo ovviamente implica che le ricerche non sono né confrontabili né replicabili.

[12] Pur nei suoi limiti, Google Ngram Viewer (d'ora in poi GNV) costituisce uno strumento utilizzabile nell'ambito di un'indagine linguistica perché permette di attingere a un vastissimo corpus annotato e chiuso, costituito da una parte consistente dei testi digitalizzati da Google Libri, e da cui possono essere estratti dati replicabili e confrontabili. È importante sottolineare che il contenuto di GNV non coincide precisamente con quello di Google Libri (sui cui limiti si veda Gomez Gane 2008): mentre infatti il secondo viene continuamente ampliato tramite l'aggiunta di nuovi testi digitalizzati, il primo è stato finora aggiornato soltanto una volta, nel 2012, e ha quindi contenuto costante.

piano sincronico. In Figura 1 si può osservare, a titolo di esemplificazione, il grafico relativo al composto *caporedattore*.

Figura 1. Attestazioni delle forme del lemma caporedattore nel corpus di Google Ngram Viewer

Diversamente da quanto riportato dal Gradit e dal Devoto Oli, il composto risulta attestato prima del 1962; in particolare, la più antica occorrenza di *caporedattore* risalirebbe al 1883, anno in cui la forma è attestata nel seguente brano tratto dalla rivista *La civiltà cattolica*:

> E tanto più realmente invalida, quanto più apparentemente validissima agli occhi degli ignari, è in primo luogo quella difesa che pel primo scoperse L. Wogue gran rabbino di Parigi e Caporedattore dell'*Univers israelite* […][13].

Un interessante contributo che GNV può dare alla lessicografia riguarda infatti la possibilità di retrodatare la prima attestazione di una parola con una certa facilità, rendendo accessibile un vastissimo repertorio di testi digitalizzati in pochi secondi. Rispetto alla formazione del plurale, il grafico permette di attestare la presenza di due forme plurali in concorrenza (*caporedattori* e *capiredattori*), che, con alterne fortune, hanno convissuto nel corso della seconda metà del Novecento[14].

Nonostante contribuiscano a definire un quadro più chiaro del fenomeno, le informazioni ricavate da GNV possono soltanto integrare, ma non sostituire, il contributo di natura quantitativa e qualitativa dei corpora tradizionali, a cui è necessario rivolgersi per integrare o rivedere il contenuto di risorse lessicografiche o di grammatiche.

Nella scelta del corpus più adatto da cui attingere i dati, due aspetti sembrano particolarmente importanti da valutare: le dimensioni e il contenuto. Quanto debba essere grande un corpus per poter essere considerato rappresentativo e fonte di dati attendibili costituisce una questione tutt'altro che risolta nella linguistica dei corpora. Marc Brysbaert e Boris New (2009) hanno osservato come le dimensioni ottimali di un corpus dipendano strettamente dalla frequenza delle parole oggetto dell'indagine: più esse sono rare, maggiori devono essere le dimensioni del corpus[15]. D'altra parte, gli stessi studiosi sottolineano che l'utilizzo di risorse troppo estese, il cui contenuto potrebbe non essere rappresentativo della lingua o di un suo particolare dominio, è indubbiamente rischioso e non sembra garantire risultati migliori[16].

Nello studio dei composti non va inoltre sottovalutata la possibilità di analizzare tutti i contesti d'uso attraverso le concordanze[17], la cui osservazione è imprescindibile per poter andare oltre il dato quantitativo e valutare il comportamento delle forme nell'uso dei parlanti. Nello studio del plurale delle parole composte questo elemento è particolarmente importante anche

[13] La rivista è stata fondata da un gruppo di gesuiti a Napoli nel 1850 ed è ancora attiva. La citazione è tratta dalla sezione "Cronaca contemporanea" (*La civiltà cattolica*, 1883, vol. 12, p. 606).

[14] Come si vedrà nel paragrafo 7.1, tale dato sarà confermato da due corpora di italiano contemporaneo.

[15] Per i dati su cui si basano le riflessioni dei due studiosi rimando al loro contributo: Brysbaert, New (2009: 977-90).

[16] In particolare, gli autori del contributo ritengono che «For most practical purposes, a corpus of 16-30 million words suffices for reliable word frequency norms. In particular, there is no evidence that a corpus of 3 billion words is much better than a corpus of 30 million words»; d'altra parte, un corpus con meno di 16 milioni di occorrenze non fornisce dati significativi per parole che presentano una frequenza inferiore a 100 occorrenze. Occorre sottolineare che quanto sostenuto dai due studiosi si riferisce in primis all'inglese (Ivi: 980).

[17] In quest'ottica, un corpus come il CORIS (Favretti *et al*. 2002), che per ragioni legate ai diritti d'autore permette di visualizzare solo una parte delle concordanze, non costituisce una risorsa efficace nello studio della flessione dei composti.

perché permette di individuare quando un composto viene usato come invariabile, caso frequente nelle forme Verbo+Nome.

4. Il campione esaminato: consistenza e metodologia di raccolta

I dati relativi ai *case studies* che si analizzeranno sono stati raccolti nell'ambito di un'indagine volta a descrivere la formazione del plurale di più tipologie di composti (Micheli 2016); di seguito si riporta una sintetica descrizione della metodologia adottata per creare il campione di parole composte di cui si è analizzata la flessione.

In una prima fase sono stati estratti dal Devoto Oli 2014 tutti i lemmi classificati come 'composti' e appartenenti alle tipologie Nome+Nome, Aggettivo+Nome, Nome+Aggettivo, Aggettivo+Aggettivo, Verbo+Nome.

Composti selezionati dal Devoto Oli 2014					
Nome+Nome	Aggettivo+Nome	Nome+Aggettivo	Aggettivo+Aggettivo	Verbo+Nome	Totale
516	177	105	65	997	1860

Tabella 1. Lemmi selezionati dal Devoto Oli 2014: distribuzione quantitativa rispetto alla tipologia di composto

Com'è noto, per il plurale delle prime quattro tipologie l'italiano ammette tre tipi di flessione: una interna, in cui solo il primo costituente viene flesso; una esterna, in cui la marca di plurale è posta sul margine destro del lessema; una doppia, in cui entrambi i costituenti vengono flessi. Per i composti Verbo+Nome, in cui il primo costituente rimane sempre invariabile, si può invece distinguere tra composti che rimangono invariabili e composti che presentano due forme distinte, una per il singolare e una per il plurale.

Nella seconda fase di raccolta sono state raccolte le forme plurali; al fine di intercettare nei corpora anche le forme non previste dalle grammatiche o dai dizionari[18], si è scelto di includere nel campione tutti i possibili tipi di plurale (con flessione interna, esterna e doppia). Ogni composto è stato quindi flesso al plurale in tre modi differenti: per *capostazione*, ad esempio, sono state create le forme *capistazione*, *capostazioni*, *capistazioni*. Nel caso di composti del tipo Verbo+Nome è stato flesso solo il secondo costituente, essendo il primo un elemento verbale invariabile. Il vantaggio di lavorare con un formario esploso è che permette di intercettare tutti i possibili tipi di plurale formulabili sfruttando le possibilità morfologiche dell'italiano, a

[18] Le indicazioni contenute nei dizionari e nelle grammatiche, insieme alla tradizione scolastica, costituiscono per Serianni le «classiche fonti della norma linguistica» (Serianni 2014,: 239). Il concetto di norma è stato, ed è tuttora, ampiamente discusso in letteratura da vari studiosi: in particolare, due posizioni diverse sono state assunte da Serianni (2004; 2014) e Sgroi (2010; 2016). Legato al concetto di norma è quello di errore, a cui è dedicato un volume recentemente curato da Grandi (2015), in cui sono raccolti numerosi contributi che, da punti di vista differenti, definiscono e interpretano il rapporto tra grammatica ed errore, regole ed eccezioni, nelle lingue naturali. Nel caso del plurale dei composti, come messo in luce in Micheli (2016: 229-33) e come si vedrà nei paragrafi successivi, le grammatiche e i dizionari non sono concordi nel definire una norma chiara rispetto alla formazione del plurale di queste parole; in assenza di una norma, parlare di 'errori' in riferimento agli usi incerti dei parlanti non è del tutto appropriato, essendo l'errore un concetto che si definisce sempre in relazione a una norma.

prescindere dalla loro attestazione nei dizionari o nelle grammatiche, ed evitare i numerosi errori che il POS tag compie con questo tipo di forme[19].

Un ulteriore elemento di cui si è tenuto conto è stato la forma grafica con cui le forme compaiono nei testi: soprattutto nel caso delle neoformazioni sono infatti molto frequenti nell'uso oscillazioni grafiche tra la forma univerbata, quella con il trattino e quella in cui i due costituenti sono separati dallo spazio. Al fine di intercettare tutti i composti, a prescindere dalla grafia presentata, ciascuna forma del campione, singolare e plurale, è stata trascritta nelle tre varianti: il lemma "capostazione", ad esempio, compare nel campione nelle forme *capostazione*, *capo-stazione*, *capo stazione*, *capostazioni*, *capo stazioni*, *capo-stazioni*, *capistazioni*, *capi stazioni*, *capi-stazioni*, *capistazione*, *capi stazione*, *capi-stazione*.

Il campione raccolto attraverso le due fasi risulta costituito da 27.609 forme, di cui nella Tabella 2 si riporta la distribuzione quantitativa rispetto alla tipologia di composto.

Tipologia di composto	AA	AN	NA	NN	VN	Totale
Forme dei lemmi	828	2.411	4.550	12.886	6.934	27.609

Tabella 2. Forme dei lemmi raccolte per l'indagine: distribuzione quantitativa rispetto alla tipologia di composto

5. Le risorse lessicografiche: il Gradit e il Devoto Oli 2014

Come risorse lessicografiche di riferimento si è scelto di adottare due dizionari molto diversi per architettura e postulati teorici – il Grande Dizionario Italiano dell'Uso[20] e il Devoto Oli 2014[21] – in modo da osservare il trattamento riservato a tale fenomeno tanto da parte di un dizionario che dichiaratamente intende porsi il più vicino possibile all'uso reale dei parlanti quanto da parte di uno di impostazione più tradizionale. I due dizionari strutturano le informazioni riguardanti il plurale dei lemmi in modo diverso: la versione elettronica del Gradit fornisce sempre la forma plurale di ciascun lemma, si essa regolare o atipica, indicando in nota eventuali forme concorrenti o meno comuni; il DO riporta il plurale soltanto nei casi in cui si abbia una flessione interna o doppia, mentre laddove non si forniscono indicazioni si intende implicitamente che il plurale presenti una regolare flessione esterna, come se si trattasse di un lessema semplice.

6. Il Corpus del Nuovo Vocabolario di Base e itWaC: limiti e potenzialità

Nella scelta della fonte di dati empirici per la presente indagine sono stati presi in considerazione tre fattori: le dimensioni, la possibilità di consultare le concordanze, il contenuto. Tra i corpora di riferimento attualmente disponibili per l'italiano, la risorsa che è parsa più vicina

[19] I motivi per cui si è scelto di lavorare con un formario esploso sono essenzialmente due: 1) nei corpora annotati attualmente disponibili per lo studio dell'italiano non è previsto un tag specifico per le parole composte che permetta di estrarle automaticamente; 2) le frequenti oscillazioni riguardanti la grafia con cui i composti ricorrono nei testi determina numerosi errori nel POS tag, limitandone l'efficacia nell'estrazione di dati relativi a questo tipo di parole.

[20] In particolare, del Grande Dizionario Italiano dell'Uso si è consultata la versione elettronica, del Devoto Oli quella cartacea.

[21] D'ora in poi Gradit e DO.

a soddisfare tali requisiti è il Corpus del Nuovo Vocabolario di Base, curato da Isabella Chiari e Tullio De Mauro (cfr. Chiari, De Mauro 2012); tale corpus, costituito da 18 milioni di occorrenze, permette di consultare le concordanze ed è bilanciato in sei sottocorpora a seconda della tipologia testuale (Stampa, Saggistica, Spettacolo, Comunicazione Mediata dal Computer, Letteratura, Parlato). Un limite di questa risorsa è legato alle sue dimensioni: essa ha infatti una estensione commisurata all'obiettivo per cui è stata creata, ossia lo studio del vocabolario di base, ma insufficiente per lo studio di entità lessicali più rare come le parole composte. L'utilizzo di questo solo corpus limiterebbe quindi l'estrazione di dati quantitativi in grado di descrivere il fenomeno: per questo motivo si è scelto di affiancargli un altro corpus di dimensioni maggiori come secondo riferimento. In virtù della sua estensione si è scelto di utilizzare itWaC (cfr. Baroni *et al.* 2009): il corpus è costituito da circa due miliardi di occorrenze e rappresenta lo strumento più esteso per lo studio dell'italiano contemporaneo; le significative dimensioni ne fanno quindi una risorsa in grado di intercettare anche entità lessicali a bassa frequenza. L'utilizzo di itWaC nell'ambito di indagini morfologiche non è raro[22], ma presenta almeno due limiti che è necessario considerare: l'assenza di un bilanciamento per generi testuali e l'assenza di una documentazione sul contenuto. Entrambi gli aspetti dipendono dalla procedura di acquisizione automatica dei testi, che di fatto rende opaco il contenuto effettivo del corpus e non permette di strutturarlo secondo criteri di bilanciamento.

	Corpus del NVDB	itWaC
Estensione	18 milioni di occorrenze	2 miliardi di occorrenze
Tipologia	Corpus di riferimento	Web corpus
Bilanciamento	Bilanciato secondo sei tipologie testuali (Letteratura, Stampa, Saggistica, Comunicazione Mediata dal Computer, Spettacolo, Parlato)	Nessun bilanciamento

Tabella 3. Caratteristiche strutturali dei due corpora

Si tratta di strumenti molto diversi tra loro quanto a struttura, contenuto e finalità, che possono quindi fornire informazioni diverse nel corso dell'indagine: il corpus per il NVdB permette di ottenere i dati sulla frequenza delle forme attestate nei vari sottocorpora e di osservare le concordanze, fondamentali per chiarire il comportamento delle forme, soprattutto di quelle invariabili; itWaC è in grado – per la sua estensione – di fornire dati più consistenti sulla frequenza delle forme che andranno confrontati con quelli del corpus del NVdB. Non va tuttavia sottovalutato il fatto che la decisione di utilizzare due corpora molto diversi pone ovviamente il problema di dover interpretare dati solo in parte confrontabili, perché estratti da risorse diverse. Nei paragrafi successivi si osserverà in che misura i due corpora possono fornire dati significativi sulla flessione di alcune particolari tipologie di parole composte.

[22] Si veda ad esempio Baroni *et al*. 2007.

7. Risultati dell'indagine: un quadro generale

Un primo dato che va osservato per valutare le potenzialità e i limiti dei due corpora riguarda il numero dei lemmi del campione che vi risultano attestati (Tabella 4).

	Corpus del NVdB	itWaC
Lemmi del campione	1886	
Lemmi attestati nel corpus	789	1468
Forme dei lemmi attestate nel corpus	1232	2869
Occorrenze delle forme dei lemmi attestate	19.066	1.648.885

Tabella 4. Composti attestati nei due corpora

Come prevedibile date le dimensioni significative del corpus, in itWaC è attestata una parte considerevole del campione di forme (circa il 70%); diversamente, nel corpus del NVdB ne è attestato circa il 40%. Tali dati offrono tuttavia soltanto una panoramica generale rispetto alle potenzialità dei due corpora: per valutare entro quali limiti essi possono fornire dei dati significativi sulla flessione delle parole composte occorre infatti osservare quanti e quali lemmi sono attestati con un numero di occorrenze sufficiente per poter comprendere e descrivere fondatamente il comportamento delle forme. La frequenza dei lemmi costituisce infatti un aspetto fondamentale per valutare l'attendibilità del contributo di un corpus; nondimeno, osservare quali sono i composti che ricorrono più frequentemente nei due corpora permette di ricavare informazioni anche sul loro contenuto.

Corpus del Nuovo Vocabolario di Base		
Rango	Lemma	Frequenza
1	Mezzogiorno	408
2	Mezzanotte	388
3	Capolavoro	385
4	Pomodoro	356
5	Portafoglio	352
6	Marciapiede	259
7	Palcoscenico	246
8	Piattaforma	239
9	Asciugamano	233
10	Pianoforte	228
11	Portavoce	200
12	Centrodestra	195
13	Reggiseno	185
14	Passaporto	181

ITWAC		
Rango	Lemma	Frequenza
1	Centrosinistra	81.381
2	Piattaforma	54.421
3	Mezzogiorno	52.566
4	Salvaguardia	51.348
5	Capolavoro	43.279
6	Centrodestra	39.904
7	Capogruppo	37.044
8	Portavoce	31.915
9	Capoluogo	31.834
10	Resoconto	25.343
11	Pomodoro	25.322
12	Portafoglio	19.602
13	Passaporto	19.141
14	Fine Settimana	18.595

Corpus del Nuovo Vocabolario di Base		
15	Fine Settimana	164
16	Capogruppo	163
17	Banconota	159
18	Buonsenso	152
19	Centrosinistra	149
20	Anno Luce	136
21	Gentiluomo	127
22	Capoluogo	122
23	Buonafede	119
24	Salvaguardia	116
25	Altopiano	108
26	Passatempo	106
27	Pianoterra	105
28	Grattacielo	104
29	Arcobaleno	97
30	Caporedattore	94

ITWAC		
15	Capoverso	18.591
16	Buonsenso	17.914
17	Pianoforte	17.821
18	Mezzanotte	16.874
19	Palcoscenico	15.834
20	Marciapiede	14.547
21	Arcobaleno	12.224
22	Buonafede	12.120
23	Cassintegrazione	10.521
24	Cortometraggio	9.861
25	Altopiano	9.232
26	Pallavolo	8.433
27	Banconota	8.383
28	Pianoterra	8.354
29	Lungometraggio	7.659
30	Bianconero	7.240

Tabella 5. Lista di frequenza dei composti nei due corpora (rango 1-30)

Confrontando le prime trenta posizioni delle liste di frequenza lemmatizzate – relative alle parole composte – dei due corpora (Tabella 5) emergono elementi interessanti circa il loro contenuto e l'influenza che esso determina sui risultati dell'indagine. Osservando le prime posizioni della lista (ranghi 1-5) emergono subito alcune differenze: *in primis*, si nota che nel corpus del NVdB le prime due posizioni sono occupate dalle indicazioni temporali *mezzogiorno* e *mezzanotte*, appartenenti alla fascia d'uso FO del Gradit, laddove in itWaC ai medesimi ranghi si trovano i composti *centrosinistra* e *piattaforma*, parole tipiche del linguaggio della politica, nel primo caso, e del web, nel secondo. In itWaC, in generale, si nota una significativa presenza, ai ranghi più alti, di termini tipici del linguaggio della politica e dei giornali: *centrosinistra* (rango 1), *centrodestra* (rango 6), *capogruppo* (rango 7), *portavoce* (rango 8). Tali forme risultano attestate anche nella lista di frequenza del corpus del NVdB ma a ranghi più bassi: *portavoce* (rango 11), *centrodestra* (rango 12), *capogruppo* (rango 16), *centrosinistra* (rango 19). I ranghi più alti nel corpus del NVdB sono invece occupati da parole di stretto uso quotidiano, appartenenti alla fascia d'uso FO del Gradit, alcuni dei quali assenti dalla lista di itWaC[23]: *pomodoro* (rango 4), *portafoglio* (rango 5), *marciapiede* (rango 6), *asciugamano* (rango 9). Dal confronto tra le due liste di frequenza si può quindi notare come in itWaC risultino sovrastimati alcuni termini particolarmente frequenti nel linguaggio della politica e del giornalismo, e sottostimate le parole di significato più concreto, appartenenti all'orizzonte quotidiano dei parlanti. Tali discrasie sono strettamente collegate al contenuto non bilanciato del corpus, costituito da testi scaricati

[23] In particolare, dei lemmi della lista di frequenza del corpus del NVdB, sono assenti da quella di itWaC le seguenti forme: *asciugamano*, *reggiseno*, *anno luce*, *gentiluomo*, *passatempo*, *grattacielo*, *caporedattore*.

automaticamente dalla rete, tra i quali è molto probabile si trovi un significativo numero di articoli giornalistici o pagine web dedicate all'analisi politica.

Per quanto riguarda il numero delle occorrenze con cui ricorrono i composti, da itWaC si possono estrarre dei dati quantitativi notevolmente più consistenti: l'alto numero di occorrenze registrato, se da un lato costituisce un aspetto positivo, perché fornisce un solido fondamento empirico all'analisi, dall'altro rappresenta un limite non indifferente, perché di fatto impedisce, o quantomeno limita, l'analisi delle concordanze e dei contesti d'uso. Non va inoltre sottovalutato il fatto che, trattandosi di un corpus dal contenuto parzialmente sconosciuto, non si è spesso in grado di verificare l'attendibilità della fonte.

Il corpus del NVdB, pur fornendo dati significativi per un insieme più limitato di forme, garantisce la possibilità di osservarne i contesti d'uso e verificare la distribuzione quantitativa del composto rispetto alla tipologia testuale.

	Lemma	Freq. Tot.	Stampa	Narrativa	Saggistica	Spettacolo	CMC	Parlato
1	Mezzogiorno	408	78.7	87.8	31.6	37.6	18	160.5
2	Mezzanotte	388	53.1	97.1	36.7	64.1	42.9	118.7
3	Capolavoro	385	95.1	33.6	97.4	44.6	192.6	7.6
4	Pomodoro	356	63.4	90.6	32.9	34.9	115	59.8
5	Portafoglio	352	86.9	75.7	10.1	76.7	42.9	87.4
6	Marciapiede	259	40.9	117.7	16.4	43.2	11.1	38
7	Palcoscenico	246	69.5	17.7	65.8	112.9	12.5	16.1
8	Piattaforma	239	111.4	15.9	79.7	7	44.3	11.4
9	Asciugamano	233	9.2	72.9	6.3	50.2	119.1	16.1
10	Pianoforte	228	30.7	86.9	22.8	19.5	20.8	55.1

Tabella 6. Primi dieci composti in ordine di frequenza attestati nel corpus del NVdB: distribuzione quantitativa rispetto ai sottocorpora

Dai dati normalizzati dei primi dieci composti più frequenti nel corpus del NVdB (Tabella 6) si può osservare il modo in cui essi si distribuiscono tra le diverse tipologie testuali. Come prevedibile, i primi due composti – che designano indicazioni temporali – si concentrano soprattutto nel parlato e in letteratura. Registrano alte frequenze nel sottocorpus di Narrativa anche altri composti i cui *designata* ricorrono frequentemente nell'orizzonte quotidiano dei parlanti – come *pomodoro*, *marciapiede*, *asciugamano*, *pianoforte* – e che in itWaC risultano sottostimati a causa del numero presumibilmente limitato di testi letterari contenuti nel corpus.

7.1 I composti capo + Nome

I composti con *capo-* fanno parte della vasta ed eterogenea categoria dei composti Nome Nome e costituiscono una serie molto numerosa in italiano: il Devoto Oli registra 129 forme; un indizio della vitalità del pattern *capo* + nome in italiano contemporaneo ci è dato dalla presenza di cinque neoformazioni di questo tipo nei repertori di Adamo e Della Valle[24]. Questo insieme di

[24] In particolare, si tratta delle forme *capoazienda*, *capodelegazione*, *caposcafista*, *capo-staff* (Adamo, Della Valle 2003b e Adamo, Della Valle 2006). Va ricordato che la rappresentatività dei due repertori è molto limitata, dal momento che essi sono stati raccolti dai due studiosi attraverso lo spoglio manuale di un corpus di quotidiani nazionali di piccole dimensioni (cfr. Adamo, Della Valle 2003a).

composti, apparentemente omogeneo dal punto di vista formale, si compone in realtà di forme costituite da due sostantivi tra i quali intercorrono relazioni diverse. Riprendendo la classificazione proposta da Serianni (1989: 154) nella sua grammatica, questo insieme di forme può essere classificato in tre tipologie a seconda che *capo* individui: 1) «colui che è a capo di qualcosa ('x è a capo y')»; 2) «colui che è a capo di qualcuno ('x è capo tra x_1, x_2, x_3...')»; 3) «ciò che si segnala tra altri oggetti omogenei come 'preminente', 'eccellente' (= 'un capo-x')». Adottando la classificazione proposta da Scalise, Bisetto (2009), i composti del primo tipo si possono inserire nel gruppo dei subordinativi (es. *capostazione*, *capoclasse*, *etc*.); quelli del secondo e del terzo tipo appartengono al tipo appositivo, con testa a sinistra nel primo caso (es. *caporedattore*, *capocuoco*, etc.), a destra nel secondo (es. *capoluogo*).

Nella Tabella 7 si riportano i dati relativi alla distribuzione quantitativa delle forme plurali e delle occorrenze rispetto alle tre tipologie individuate attraverso il criterio semantico.

		Corpus del NVDB				itWaC			
Tipologia di composto	Tipo flessione	Numero forme	%	Occorrenze	%	Numero forme	%	Occorrenze	%
'x è il capo di y'	Flessione interna	28	90	89	89.9	83	79	13.700	88.7
	Flessione esterna	2	6.5	9	9.1	14	13.3	1.644	10.6
	Doppia flessione	1	3.5	1	1	8	7.6	93	0.7
'x è capo tra x_1, x_2, x_3…'	Flessione interna	0	0	0	0	1	23.8	11	1.9
	Flessione esterna	3	50	4	57.1	10	47.6	275	47.9
	Doppia flessione	3	50	3	42.9	10	47.6	288	50.2
'un capo-x'	Flessione interna	0	0	0	0	5	41.6	41	0.2
	Flessione esterna	2	100	130	100	5	41.6	19.009	99.7
	Doppia flessione	0	0	0	0	2	16.8	13	0.07

Tabella 7. Composti con *capo*-: distribuzione quantitativa delle occorrenze rispetto alla classificazione di Serianni (1989)

Nel gruppo dei composti determinativi si può individuare una tendenza molto netta a formare il plurale modificando soltanto il primo costituente, che rappresenta l'elemento testa del composto. La significativa frequenza con cui queste forme ricorrono nell'uso non sembra aver determinato una diminuzione nel grado di trasparenza del composto né lo spostamento della flessione sul margine destro del lessema. Anche per i composti in cui *capo* designa un elemento che si segnala come preminente rispetto ad altri non si registrano oscillazioni significative nella formazione del plurale: nella quasi totalità delle forme la marca di plurale è posta sul costituente testa, posto questa volta a destra.

Le occorrenze dei composti in cui *capo* designa 'colui che è a capo di qualcuno', tipologia meno produttiva delle altre, si distribuiscono quasi equamente tra forme plurali che modificano

soltanto il secondo costituente e forme plurali con entrambi i costituenti modificati: i due tipi di flessione sembrano quindi in concorrenza nell'uso dei parlanti[25]. Del composto più frequente appartenente a tale tipologia, *caporedattore*, itWaC attesta ben quattro forme plurali: *capiredattori* (65 occ.)/*capi redattori* (57 occ.), *caporedattori* (64 occ.) e *capiredattore* (3 occ.).

L'osservazione del contesto in cui le singole forme vengono utilizzate completa il quadro su questo tipo di composti: in relazione a questo aspetto, il corpus del NVDB fornisce un contributo fondamentale perché permette di osservare le concordanze; le enormi dimensioni di itWaC rendono invece di fatto impossibile il suo utilizzo in questo senso. Rispetto alla tipologia dei composti con *capo-*, si può notare che solamente quattro composti, *capofamiglia*, *capolinea*, *capobranco*, *capofila*, in un numero di casi molto limitato (2 occorrenze), vengono utilizzati come invariabili.

Delle tre tipologie, quella in cui si rilevano più discrasie tra i due dizionari, rispetto alla flessione al plurale, è la seconda: nella seguente tabella (Tabella 8) se ne riportano alcuni esempi che permettono di osservare tre casi in cui i dati empirici possono essere utilizzati per migliorare le informazioni lessicografiche. Come si può notare, i dati estratti dal corpus del NVdB per queste forme non sono sufficienti a osservare il fenomeno; in questo caso, quindi, si farà riferimento a quelli estratti da itWaC.

Lemma	Forme plurali attestate	itWaC		Corpus del NVdB		Plurale Devoto Oli	Plurale Gradit
		Freq. plurale	Freq. sing.	Freq. plurale	Freq. sing.		
capocannoniere	capocannonieri	36	1.105	0	12	capicannonieri	capocannonieri
	capicannonieri	1		0			
capocomico	capocomici	31	754	1	4	capocomici, capicomici (*meno comune*)	capocomici, capicomici
	capicomici	3		0			
capocronista	capocronisti	3	99	0	0	capicronisti	capicronisti
	capicronisti	6		0			
capocuoco	capocuochi	3	117	0	18	capocuochi, capicuochi	capocuochi, capicuochi
	capicuochi	0		1			
capomastro	capomastri	53	288	2	12	capomastri, capimastri	capomastri, capimastri
	capimastri	49		0			
capo operaio	capo-operai	1	114	0	0	capi operai	capooperai
	capi operai	37		0			
caporedattore	caporedattori	64	2.244	1	91	capiredattori	caporedattori
	capiredattori	122		2			
	capiredattore	3[26]		0			

[25] In questi casi si potrebbe parlare di «sovrabbondanza», riprendendo il termine proposto da Thornton (2012: 183) in riferimento alla polimorfia nei paradigmi flessivi verbali dell'italiano, definito come il fenomeno per cui «two or more forms are available to realize the same cell in an inflectional paradigm».

[26] Effettuando su itWaC una query relativa alla forma *capiredattore*, si trovano inizialmente 11 risultati: tuttavia, osservando le concordanze, si può notare che in nove casi si tratta dello stesso testo copiato in diverse pagine internet; in realtà, la forma ha quindi solo 3 occorrenze. Questo caso permette di riflettere

capotecnico	capotecnici	7	316	0	0	capotecnici, capitecnici	capitecnici, capotecnici
	capitecnici	66		0			

Tabella 8. Composti con *capo-*: casi di discrasia tra dati empirici e dizionari

Il primo caso è quello di *capocannoniere*, del quale i due dizionari riportano due forme di plurale differenti: il Gradit registra la forma con flessione esterna *capocannonieri*, il DO quella con doppia flessione *capicannonieri*; i dati estratti da itWaC mostrano invece un quadro molto chiaro in cui la forma con flessione esterna è la più frequente nell'uso dei parlanti.

Il secondo caso è quello di *capocomico* e *capotecnico*: per entrambi i composti i due dizionari registrano due plurali, laddove i dati di itWaC mostrano la netta prevalenza di una delle due forme, quella con flessione esterna nel primo caso (*capocomici*), quella con doppia flessione nel secondo (*capitecnici*).

Diverso è il caso di *caporedattore*, del quale itWaC attesta ben tre forme di plurale in concorrenza nell'uso dei parlanti[27]: tale oscillazione non è colta dai dizionari che si limitano a registrare una sola forma, su cui peraltro non concordano (per il Gradit il plurale di *caporedattore* è *caporedattori*, per il DO è *capiredattori*).

7.2 I composti basso- + Nome

I composti *basso*+Nome sembrano costituire una serie non più produttiva in italiano contemporaneo: nei repertori di neologismi di Adamo e Della Valle (2003b, 2006) non è registrata nessuna parola composta con *basso* come primo costituente. Insieme alle forme con *alto*, questo tipo di composti viene segnalato dalle grammatiche[28] perché ammette delle oscillazioni nella formazione del plurale.

Lemma	Forme plurali attestate	itWaC		Corpus del NVdB		Plurale Devoto Oli	Plurale Gradit
		Freq. plurale	Freq. singolare	Freq. plurale	Freq. singolare		
bassadanza	bassedanze	2	27	0	0	bassadanze, bassedanze	bassadanze
bassofondo	bassofondi	6	198	0	0	bassifondi	bassifondi, bassofondi (ant)
	bassifondi	1.513		13			
bassopiano	bassopiani	44	157	0	2	bassopiani, bassipiani	bassipiani
	bassipiani	71		0			

su uno dei problemi con cui si deve fare i conti quando si utilizza itWaC: essendo costituito da pagine web scaricate automaticamente dalla rete, esso contiene molte copie di testi che possono falsare anche in modo significativo il dato quantitativo sulle occorrenze di una forma.

[27] Questi dati riflettono, e quindi confermano, la situazione delineata dal grafico prodotto da GNV (paragrafo 3).

[28] È ad esempio segnalato in Serianni (1989: 155).

bassorilievo	bassorilievi	1.272	1.327	10	13	bassorilievi	bassorilievi, bassirilievi (*ant.*)
	bassirilievi	42		0			

Tabella 9. Composti con *basso-*: casi di discrasia tra dati empirici e dizionari

I dati riportati in Tabella 9 permettono di osservare come, anche per questo tipo di forme, le indicazioni dei dizionari non sempre coincidano e i dati estratti dai corpora possano aiutare a chiarire la situazione.

Della forma *bassadanza* itWaC attesta una sola forma plurale con doppia flessione, *bassedanze*; i due dizionari si comportano in modo diverso: il Gradit registra la forma con flessione esterna, il DO riporta due plurali.

Bassopiano presenta nell'uso due forme plurali concorrenti, entrambe attestate in itWaC con un significativo numero di occorrenze; tale oscillazione è segnalata dal DO, ma non dal Gradit, che si limita a registrare la forma con doppia flessione.

Nei casi di *bassofondo* e *bassorilievo*, i dati di itWaC e del Corpus del NVdB indicano che i parlanti utilizzano una forma plurale in particolare, nettamente più frequente rispetto alla concorrente e registrata dal DO come unico plurale; in questi casi il Gradit sceglie di riportare anche l'altra variante, specificando però che si tratta di una forma più antica.

7.3 I composti Verbo + Nome

I composti Verbo+Nome rappresentano la tipologia su cui per più tempo e in modo più significativo si è concentrata l'attenzione degli studiosi: l'aspetto su cui si è più a lungo dibattuto è la natura del primo costituente (cfr. Bisetto 1999; Floricic 2008); i contributi più recenti ne hanno invece analizzato la produttività e le proprietà semantiche e morfosintattiche (cfr. Magni 2010; Ricca 2005, 2010, 2015; Von Heusinger, Schwarze 2011).

Per analizzare la flessione di questo tipo di forme è fondamentale poter disporre di un corpus che consenta lo spoglio delle concordanze per osservare i contesti d'uso. Questo tipo di composti può infatti essere variabile, ossia presentare una forma per il singolare e una per il plurale, o invariabile, quando compare nell'uso in un'unica forma: per discriminare tra i due casi è quindi necessario osservare il contesto sintattico in cui il composto è utilizzato. La tabella 10 riporta un campione di composti Verbo+Nome[29] che, sulla base dei dati estratti dal corpus del NVdB, possono essere classificati come "variabili": si tratta di composti che al singolare sono costituiti da un elemento verbale e un sostantivo singolare (ad esempio *grattacielo*); al plurale essi vengono sempre flessi modificando il secondo elemento (quindi *grattacieli*). Queste forme non sono quindi mai attestate nel corpus come invariabili: il composto *passaporto*, ad esempio, è sempre attestato come singolare; la forma *passaporti* indica sempre un plurale.

Lemma	Forme plurali attestate	Corpus del NVdB	itWaC		Plurale Gradit

[29] Per questioni di spazio si riportano solo i primi otto composti VN variabili in ordine di frequenza; per un quadro più ampio rimando a Micheli (2016: 245-52).

		Occ. singolare	Occ. plurale	Occ. Tot lemma	Occ. tot lemma	Plurale Devoto Oli	
battibecco	battibecco	20	0	39	1.391	Variabile	Variabile
	battibecchi	0	19				
batticuore	batticuore	18	0	21	699	Invariabile	Variabile
	batticuori	0	3				
giravolta	giravolta	11	0	16	786	Variabile	Variabile
	giravolte	0	5				
grattacielo	grattacielo	55	0	104	6.978	Variabile	Variabile
	grattacieli	0	49				
parapetto	parapetto	39	0	44	2.534	Variabile	Variabile
	parapetti	0	5				
parasole	parasole	12	0	14	539	Invariabile	Variabile
	parasoli	0	2				
passaporto	passaporto	158	0	181	19.141	Variabile	Variabile
	passaporti	0	23				
portabandiera	portabandiera	13	0	16	917	Invariabile	Invariabile
	portabandiere	0	3				

Tabella 10. Composti Verbo+Nome: forme variabili più frequenti nel corpus del NVdB

Come si può notare osservando i dati sulla frequenza totale dei lemmi, queste forme registrano in itWaC un altissimo numero di occorrenze, che rende di fatto impossibile controllarne tutti i contesti: in questo caso, le dimensioni di itWaC costituiscono quindi un forte limite a tale risorsa. Diversa è la situazione del corpus del NVdB, sulla base del quale è invece possibile analizzare il fenomeno e confrontarlo con le due opere lessicografiche.

In questo caso è soprattutto il DO a riportare informazioni discordanti con i dati empirici: *batticuore*, *parasole* e *portabandiera* vengono infatti registrati come invariabili, nonostante nell'uso siano attestate le rispettive forme flesse; più preciso sembra essere il Gradit che sbaglia solo nel caso di *portabandiera*.

8. Conclusioni

Con questo lavoro si è cercato di contribuire alla riflessione sul rapporto tra dati empirici e risorse lessicografiche, con particolare riferimento a un fenomeno morfologico spesso oggetto di dubbi da parte dei parlanti: la formazione del plurale delle parole composte.

Nella prima parte del lavoro (§ 1-3) si è discussa la possibilità di utilizzare il web come fonte di dati, mettendo in luce i numerosi problemi metodologici che ne inficiano la scientificità e ne scoraggiano quindi l'uso nell'ambito di studi linguistici, nei quali d'altra parte il contributo dei corpora tradizionali risulta insostituibile. Sono stati poi analizzati alcuni aspetti quantitativi e

qualitativi che distinguono le parole composte dai lessemi semplici, in relazione ai quali sono stati proposti dei criteri metodologici per la scelta del campione e del corpus da utilizzare per un'indagine che intenda osservarne la formazione del plurale. I principali problemi ancora irrisolti riguardano in particolare due questioni: le caratteristiche che un corpus dovrebbe presentare per permettere lo studio delle parole composte e la loro effettiva applicabilità a una risorsa reale. Dimensioni e contenuto del corpus costituiscono gli aspetti su cui è necessaria una riflessione più ampia rispetto a quanto è stato fatto finora. In particolare, da un lato occorre valutare con più precisione l'estensione del corpus rispetto alla frequenza delle entità lessicali oggetto di studio, dall'altro costituiscono requisiti imprescindibili il bilanciamento del corpus e la possibilità di consultarne le concordanze. Riuscire a garantire l'estensione necessaria e la qualità del contenuto di un corpus rappresenta un problema di non facile soluzione a livello pratico; questo per almeno due motivi: perché la raccolta di corpora molto estesi, come ad esempio itWaC, implica l'utilizzo di una procedura automatica che non permette alcun bilanciamento rispetto alla tipologia testuale e perché le grandi dimensioni di un corpus impediscono, o quantomeno limitano, la possibilità di consultarne le concordanze. Sul rapporto tra dimensioni e qualità del contenuto devono quindi concentrarsi le riflessioni future.

Nella seconda parte del contributo (§ 4-7) sono stati presentati la metodologia di indagine e i risultati di una ricerca condotta su due corpora di italiano contemporaneo, itWaC e il corpus del Nuovo Vocabolario di Base: attraverso la descrizione di alcuni *case studies*, relativi a tre particolari tipologie di composti, si è messo in luce in che modo, e in che misura, i dati estratti da corpora possono contribuire a migliorare le informazioni di natura morfologica fornite da due dizionari, il Devoto Oli 2014 e il Gradit, di cui si sono messi in evidenza i limiti. L'analisi dei composti con *capo-* e *basso-* come primo costituente ha permesso di individuare almeno tre casi in cui quanto riportato dai dizionari non riflette i dati quantitativi estratti da corpora e necessita quindi di integrazioni o correzioni: (a) Gradit e Devoto Oli registrano forme di plurale diverse, di cui una assente, o attestata con valori di frequenza bassissimi, dai corpora; (b) uno dei due dizionari, o entrambi, registra più forme plurali, laddove nell'uso ne risulta attestata una in particolare, con alti valori di frequenza; (c) uno dei due dizionari, o entrambi, registra un solo plurale, nonostante nell'uso vi siano più forme in concorrenza, con valori di frequenza molto vicini. Il caso dei composti Verbo Nome ha permesso di mettere in luce l'importanza di osservare i contesti d'uso per individuare quando la forma è invariabile: anche rispetto alla variabilità/invariabilità di tali composti si sono registrate incongruenze tra dati e dizionari.

Rispetto a tali problemi, il contributo dei due corpora si è dimostrato differente, in una certa misura complementare, e strettamente legato alle caratteristiche strutturali delle due risorse: in particolare, si è potuto osservare che un corpus di medie dimensioni e di contenuto variato, come il corpus del NVdB, permette di estrarre dati quantitativi e qualitativi relativi soltanto a un numero limitato di composti, ma garantisce la possibilità di osservarne i contesti d'uso e coglierne la variabilità/invariabilità; d'altra parte, un corpus di grandi dimensioni come itWaC fornisce dati quantitativi più consistenti e relativi a un maggior numero di forme, ma ne limita l'analisi qualitativa.

Il presente lavoro ha dimostrato che, seppur ancora limitato da questioni teoriche e pratiche, il contributo dei dati empirici in lessicografia appare irrinunciabile anche per quanto riguarda le informazioni morfologiche che accompagnano i lemmi; è d'altra parte necessario portare avanti la riflessione teorica sulla realizzazione di corpora e sull'estrazione di dati rappresentativi e consistenti che permettano di rendere proficuo il dialogo tra dizionari e dati.

BIBLIOGRAFIA

Adamo, G., Della Valle, V. (2003a), *L'Osservatorio neologico della lingua italiana: linee di tendenza nell'innovazione lessicale dell'italiano contemporaneo*, in Idd. (a c. di), *Innovazione lessicale e terminologie specialistiche*, Firenze, Olschki, pp. 83-105.

Adamo, G., Della Valle, V. (2003b), *Neologismi quotidiani: un dizionario a cavallo del millennio (1998-2003)*, Firenze, Olschki.

Adamo, G., Della Valle, V. (2006), *2006 parole nuove. Un dizionario di neologismi dai giornali*, Milano, Sperling & Kupfer.

Baroni, M., Bernardini S., Ferraresi A., Zanchetta E. (2009), *The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora*, in "Journal of Language Resources and Evaluation", 43(3), pp. 209-226.

Baroni, M., Guevara, E., Pirrelli, V. (2007), *NN compounds in Italian: Modelling category induction and analogical extension*, in V. Pirelli (a cura di), *Psycho-Computational Issues in Morphology Learning and Processing* (Special Issue of "Lingue e Linguaggio", 6.2), Bologna, Il Mulino, pp. 263-290.

Baroni, M., Kilgarriff, A. (2006), *Large linguistically-processed web corpora for multiple languages*, in *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Posters & Demonstration*s. Association for Computational Linguistics, pp. 87-90.

Baroni, M., Guevara, E., Pirrelli, V. (2006), *Sulla tipologia dei composti NN in italiano: principi categoriali ed evidenza distribuzionale a confronto*, in G. Ferrari, R. Benatti, M. Mosca (a cura di), *Linguistica e Modelli tecnologici della ricerca. Atti del XL Congresso Internazionale di Studi della Società di Linguistica Italiana*, Roma, Bulzoni, pp. 21-38.

Beißwenger, M., Chanier, T., Chiari, I., Erjavec, T., Fisser, D., et al.. (2016) *Integrating corpora of computer-mediated communication into the language resources landscape: Initiatives and best practices from French, German, Italian and Slovenian projects. CLARIN Annual Conference 2016*, Oct 2016, Aix-en-Provence, France. https://www.clarin.eu/content/programme-clarin-annual-conference-2016.

Bisetto, A. (1999), *Note sui composti VN dell'italiano*, in P. Benincà, A. Mioni, L. Vanelli (a c. di), *Fonologia e morfologia dell'italiano e dei dialetti d'Italia*, Roma, Bulzoni, pp. 503-38.

Brysbaert, M., New B. (2009), *Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English*, in "Behavior research methods" 41.4, pp. 977-90.

Chiari, I., (2016) (a c. di), *Capirsi e fraintendersi al computer. La negoziazione del senso nella conversazione sui nuovi media*, Roma, Carocci (in stampa).

Chiari, I., De Mauro, T. (2012), *The new basic vocabulary of Italian: problems and methods*, in "Statistica Applicata. Italian Journal of Applied Statistics", 22 (1), pp. 21-35.

Crystal, D. (2006), *Language and the internet*, Cambridge, Cambridge University Press.

D'Achille, P., Grossmann, M. (2009), *Stabilità e instabilità dei composti aggettivo + aggettivo in italiano*, in E. Lombardi Vallauri, L. Mereu (a c. di), *Spazi linguistici. Studi in onore di Raffaele Simone*, Roma, Bulzoni, pp. 143-71.

Devoto G., Oli G. C. (2014), *Il Devoto Oli. Vocabolario della lingua italiana*, a cura di Luca Serianni e M. Trifone, Firenze, Le Monnier.

Favretti, R. Rossini, Tamburini F., De Santis C. (2002), *CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model*, in *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Munich: Lincom-Europa".

Floricic, F. (2008), *The Italian Verb-Noun anthroponymic Compounds at the Syntax / Morphology Interface*, in "Morphology", 18.2, pp. 167-93.

Gomez Gane, Y. (2008), *Google ricerca libri e la linguistica italiana: vademecum per l'uso di un nuovo strumento di lavoro*, in "Studi linguistici italiani", 2, pp. 1000-19.

Gradit = *Grande dizionario italiano dell'uso*, diretto da Tullio De Mauro, Torino, Utet, 2007.

Grandi, N. (2015) (a c. di), *La grammatica e l'eяrore. Le lingue naturali tra regole, loro violazioni ed eccezioni*, Bologna, Bononia U.P.

Grossmann, M., Rainer F. (2009), *Italian adjective-adjective compounds: Between morphology and syntax*, in "Italian Journal of Linguistics", 21.1, pp. 71-96.

Hathout, N., Montermini, F., Tanguy, L. (2008), *Extensive data for morphology: using the World Wide Web*, in "French Language Studies", 18, pp. 67-85.

Iacobini, C., Thornton, A. M. (2016), *Morfologia e formazione delle parole*, in S. Lubello (a c. di), *Manuale di linguistica italiana*, Berlin/Boston, De Gruyter Mouton, pp. 190-220.
Iannàccaro, G. (2000), *Per una semantica più puntuale del concetto di 'dato linguistico': un tentativo di sistematizzazione epistemologica*, in "Quaderni di semantica: rivista internazionale di semantica teorica e applicata", 21.1, pp. 51-80.
Kilgarriff, A. (2007), *Googleology is bad science*, in "Computational linguistics", 33.1, pp. 147-51.
Kilgarriff, A., Grafenstette, G. (2003), *Introduction to the special issue on the web as corpus*, in "Computational linguistics", 29.3, pp. 333-47.
Lehmann, C. (2004), *Data in linguistics*, in "The Linguistic Review", 21.3-4, pp. 175-210.
Libber, G., Jarema, G. (2007) (a c. di), *The Representation and Processing of Compound Words*, Oxford University Press.
Lüdeling, A., Evert, S., Baroni, M. (2007), *Using web data for linguistic purposes*, in M. Hundt, N. Nesselhauf, C. Biewer (a c. di), *Corpus linguistics and the Web*, Amsterdam, Rodopi, pp. 7-24.
Magni, E. (2010), *From the periphery to the core of Romance [VN] compounds*, in "Lingue e Linguaggio", 9.1, pp. 3-39.
Masini, F., Scalise, S. (2012), *Italian compounds*, «PROBUS», 24, pp. 61-91.
Michel, J. B., Shen, Y. K., Aiden, A. P. Veres, A., Gray, M. K., Brockman, W. (2012), The Google Books Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. Lieberman Aiden, *Quantitative Analysis of Culture Using Millions of Digitized Books*, in "Science", pp. 176-82.
Micheli, M. S., (2016), *Sul plurale delle parole composte nell'italiano contemporaneo*, in "Studi di lessicografia italiana", XXXIII, pp. 227-55 (in stampa).
Montermini, F. (2015), *Regole (e irregolarità) nella formazione delle parole*, in Grandi (2015), pp. 63-83.
Montermini, F. (2008), *La composition en italien dans un cadre de morphologie lexématique*, in D. Amiot, *La composition dans une perspective typologique*, Artois Presses Université, pp. 161-87.
Ricca, D. (2015), *Verb-noun compounds in Romance*, in Peter O. Müller, I. Ohnheiser, S. Olsen, F. Rainer (a c. di), *Word-formation. An International Handbook of the Languages of Europe*, Berlin/Boston, De Gruyter Mouton, pp. 688-707.
Ricca, D. (2010), C*orpus data and theoretical implications: With special reference to Italian VN compounds*, in S. Scalise, I. Vogel (a c. di), *Cross-disciplinary Issues in Compounding*, Amsterdam/Philadelphia, John Benjamins, pp. 237-54.
Ricca, D. (2005), *Al limite tra sintassi e morfologia; i composti aggettivali VN nell'italiano contemporaneo*, in M. Grossmann, A. Thornton, (a c. di) *La formazione delle parole. Atti del XXXVII Congresso di Studi della Società di Linguistica Italiana*, Roma: Bulzoni, pp. 1000-22.
Scalise, S., Bisetto, A. (2009), *The classification of compounds*, in R. Lieber e P. Stekauer (a c. di), "The Oxford handbook of compounding", OUP Oxford, pp. 49-82.
Serianni, L. (2014), *Giusto e sbagliato: dove comincia il territorio dell'errore?*, in S. Lubello (a c. di), *Lezioni d'italiano. Riflessioni sulla lingua del nuovo millennio*, Bologna, Il Mulino, pp. 235-46.
Serianni, L. (2004), *Il sentimento della norma linguistica nell'Italia di oggi*, in "Studi Linguistici Italiani", XXX, pp. 85-103.
Serianni, L. (1989), *Grammatica italiana*, Torino, UTET.
Sgroi, S. C. (2016), *Grammatica "clericale" vs. grammatica "laica"*, in "Rivista Italiana di Dialettologia. Lingue dialetti società", XXXIX, pp. 169-85.
Sgroi, S. C. (2010), *Per una grammatica «laica». Esercizi di analisi linguistica dalla parte del parlante*, Torino, UTET.
Thornton, A. M. (2012), *Reduction and maintenance of overabundance. A case study on Italian verb paradigms*, in "Word Structure", 5, pp. 183-207.
Von Heusinger, K., Schwartze, C. (2011), *Italian V+N compounds, inflectional features and conceptual structure*, in "Morphology", 23.3, pp. 325-50.

M. SILVIA MICHELI • PhD Student in Linguistic Science (University of Pavia, University of Bergamo, Italy); scientific interests: morphology, word-formation, corpus linguistics.

E-MAIL • s.micheli@outlook.it