

Difficoltà con la lemmatizzazione dei binomi lessicali – come potrebbe cavarsi d’impaccio la fraseografia moderna?

Difficulties with the Lemmatization of Lexical Binomials – How Can Modern Phraseography Overcome This Problem?

Damir Mišetić¹

Abstract: The pragmatic turn in the seventies and the growing interest in spoken language in recent decades require modern phraseography and lexicography to develop new and more suitable approaches such as lemmatizing fixed phrases. A class of phrasemes that is particularly interesting for phraseological research are lexical binomials, which are present in all modern languages, even the Romance ones, including Italian. However, their lemmatization involves considerable difficulties for lexicographers and phraseographers. Both in-depth theoretical knowledge and practical know-how on how to recognize and lemmatize them, especially reversible binomials, are required. Unfortunately, Italian phraseography and lexicography do not seem to notice these important developments. Although empirical methods alone will not be able to solve the problems that hinder the development of modern phraseographic research, it will undoubtedly be able to provide us with reliable information on the phraseological heritage of European languages in general, and Italian in particular.

Keywords: lexical binomials, web corpora, lemmatization, irreversibility, Italian

1. Introduzione

La svolta pragmatica negli anni Settanta ha suscitato innanzitutto maggiore interesse per la lingua parlata in generale (cf. Filatkina 2016 : 9), ovverosia per le conversazioni quotidiane, per il parlato spontaneo. Negli ultimi decenni, i cambiamenti spesso svariati e rapidi della lingua standard a livello sincronico richiedono alla fraseografia e alla lessicografia moderna lo sviluppo di approcci nuovi e più adeguati, come la registrazione delle locuzioni fisse nei dizionari fraseologici e in quelli d’uso, onnipresenti sia nella lingua letteraria e

¹ Università di Mostar; damir.misetic@ff.sum.ba.

nello scritto in generale, sia soprattutto negli scambi meno formali, nelle conversazioni quotidiane più aperte ad accogliere le ultime novità rispetto alla lingua letteraria. Questo fatto diventa più interessante in particolare rispetto allo sviluppo e alle esigenze della fraseodidattica come disciplina che si occupa dell'insegnamento dei frasemi e delle collocazioni italiane agli stranieri ed analizza il loro apprendimento in generale (cf. Konecny, Hallsteinsdóttir & Kacjan 2013: *passim*; Autelli & Konecny 2015: 190).

In questo contesto occorre menzionare una classe di locuzioni fisse, ossia combinazioni fraseologiche particolarmente interessanti e al contempo complesse per le ricerche fraseologiche e fraseografiche moderne: i binomi lessicali. Secondo il fraseologo tedesco Harald Burger (Burger, Buhofer & Salm 1982: § 2.3.6) si parla di un binomio lessicale “quando si ha a che fare con due parole diverse della stessa categoria, collegate da una congiunzione o da Ø, con un ordine più o meno fisso; quando si osserva una sequenza (più o meno) irreversibile; oppure quando due parole identiche, unite da congiunzioni o preposizioni, formano un legamento fisso”². I binomi lessicali sono collegati prevalentemente dalla congiunzione copulativa *e* (v. Ljubičić & Mišetić 2021 : 111). Diversi criteri semantici, pragmatici, metrici (v. Mišetić 2018) e fonologici (v. Budimir & Mišetić, 2020) possono influenzare la sequenza dei loro costituenti, che è spesso irreversibile o almeno preferenziale; tuttavia, un numero consistente di binomi reversibili compare in entrambe le sequenze (Budimir & Mišetić 2023). I costituenti dei binomi lessicali appartengono prevalentemente alla stessa categoria lessicale, pur riscontrandosi qualche esempio di forme miste (Ljubičić & Mišetić, 2021 : 108). Oltre a ciò, la loro semantica presenta notevole complessità e si nota subito che esiste “un rapporto semantico particolare tra i due costituenti del binomio e che non si tratta di qualunque combinazione di parole: la loro invariabilità e spesso la non modificabilità interna mostrano che si tratta di un lessema complesso, il cui significato di solito non è desumibile da quello delle parole che lo compongono” (Ljubičić & Mišetić 2023 : 29). Sono presenti in tutte le lingue moderne e rappresentano un universale linguistico (Wälchli 2005). Si riscontrano nelle lingue romanze, italiano compreso, seppur con una frequenza inferiore rispetto alle lingue germaniche e slave (Luque Nadal 2017 : 151):

in italiano: *a uso e consumo; spendere e spandere; né carne né pesce*
 in romeno: *urcând și coborând; din când în când; între ciocan și nicovală*
 in francese: *corps et biens; haut et fort; de temps en temps*
 in spagnolo: *con uñas y dientes; sin ton ni son; ni siente ni padece*
 in portogese: *do pau e da cenoura; alto e claro; a bem ou a mal*
 in catalano: *sa i estalvi; viu o mort; tard o d' hora.*

² Tutte le traduzioni dal tedesco, dal francese e dal croato sono di chi scrive.

Il fenomeno è noto fin dall'antichità classica ed era presente già in latino, come dimostrano esempi quali *sine ira et studio*, *ora et labora*, *panem et circenses*.

La lemmatizzazione di tali unità rappresenta una sfida considerevole per la lessicografia e la fraseografia odierne, specialmente alla luce delle ricerche nei web corpora che sono collezioni di testi orali e scritti che contano miliardi di token. Il problema della lemmatizzazione più adatta delle locuzioni fisse nelle lingue europee occupa già da decenni un posto di primaria importanza in letteratura (v. Pilz 2002; Hallsteinsdóttir 2006; Korhonen 2011; Stumpf 2019; Budimir & Mišetić 2022 ecc.). Per registrarli nei dizionari in maniera più pertinente occorre sia la conoscenza teorica approfondita sia il know-how pratico per riconoscerli nei web corpora e lemmatizzarli, specialmente quelli reversibili, ovverosia i binomi che appaiono in ambedue le sequenze (Mišetić & Budimir 2022 : 67). In questo contesto, la linguistica quantitativa potrebbe offrire un contributo rilevante, come evidenziano le ricerche recenti sullo sloveno (Kelih 2017, 2020), sul croato e sull'italiano (Mišetić & Budimir 2022). Si delineano, infatti, nuove prospettive teorico-metodologiche per l'elaborazione di dizionari fraseologici e d'uso, capaci di offrire una rappresentazione più adeguata del lessico e del patrimonio fraseologico e paremiologico dell'italiano contemporaneo. Eppure, la fraseografia e la lessicografia in ambito italiano, purtroppo, sembrano non accorgersi di questi sviluppi importanti³.

Nel presente contributo si cercherà di analizzare l'uso effettivo di una selezione di binomi lessicali, estrapolati dai dizionari fraseologici cartacei e digitali, nei web corpora *itWaC*⁴ e *itTenTen20*⁵ e di alcuni binomi, estrapolati dai testi moderni ed autentici, per individuare le difficoltà della loro lemmatizzazione, particolarmente di quelli reversibili. Si tratterà prima il problema dei binomi registrati nei dizionari italiani, innanzitutto in quelli fraseologici, nonché in quelli d'uso, che non sono attestati nell'uso contemporaneo a livello

³ A conferma di ciò, si rileva la scarsità di contributi scientifici sistematici dedicati alla lemmatizzazione dei binomi lessicali e alla loro rappresentazione nei dizionari, sia d'uso che fraseologici. Mancano analisi empiriche basate su corpora moderni, indagini approfondite sull'uso reale delle combinazioni fisse e riflessioni metodologiche aggiornate, che invece sono sempre più presenti nella lessicografia di area germanofona e anglosassone.

⁴ *Italian Web Corpus* (*itWaC*) è il primo web corpus italiano di nuova generazione. È una raccolta di testi scaricati con metodi automatici dalle pagine web italiane che contiene "quasi 2 miliardi di tokens, corrispondenti a oltre un miliardo e mezzo di parole grafiche, lemmatizzate ed etichettate per PoS attraverso TreeTagger" (Cresti & Panunzi 2013: 152).

⁵ Il web corpus della lingua italiana più grande *itTenTen20* è, come ad es. *deTenTen*, uno dei corpora della famiglia *TenTen* alla quale appartengono i corpora generati automaticamente da internet, che si trovano nella piattaforma web di Sketch Engine (cf. Flinz & Mollica 2021: 111) (https://app.sketchengine.eu/#dashboard?corpname=prelo%2Fittenten20_fl1).

sincronico. Si affronterà il problema dei binomi, estrapolati non solo da testi moderni, nonché da altre tipologie di fonti, frequentemente attestati nei web corpora dell’italiano (*itWaC* e *itTenTen20*). Cionondimeno, il loro lemma fraseologico risulta assente dai dizionari. Infine si parlerà di una difficoltà particolare attinente alla lemmatizzazione dei binomi lessicali: della loro variazione topologica o (ir)reversibilità e della lemmatizzazione più adatta alla luce delle ricerche su corpus, cioè grazie all’analisi che si basa sull’osservazione empirica di dati reali nella lingua italiana, il parlato spontaneo incluso.

2. Binomi registrati nei dizionari fraseologici e in quelli d’uso che non si usano nell’italiano contemporaneo

In apertura, si impone l’attenzione su un fenomeno di particolare rilievo, emblematico della fraseografia contemporanea, ben oltre i confini italiani: nei dizionari fraseologici, nonché in quelli d’uso sono registrate le costruzioni binomiali che non si usano a livello sincronico (Mišetić & Budimir 2022 : § 4). La questione non è limitata alla lessicografia e alla fraseografia della lingua italiana, ma riguarda pure tutte le altre lingue europee e, come già sopramenzionato, costituisce un tema oggetto di lunga discussione nella letteratura fraseologica (cf. Hallsteinsdóttir 2006). Secondo le ricerche di Mišetić e Budimir (2022 : § 3) in italiano si riscontrano 311 binomi lessicali registrati nei dizionari fraseologici. Da questi 311 binomi estrapolati dai dizionari fraseologici (v. Mišetić & Budimir 2022 : 69) 75 binomi non sono attestati nell’*Italian web corpus* (*itWaC*): numerosi non hanno alcuna occorrenza nell’*itWaC* o il loro uso è sporadico e tali binomi non soddisfano il criterio della rappresentatività, ad es.: *non essere né cotto né crudo; mescolare ebrei e samaritani; ammazzar bestie e cristiani; dal Campidoglio alla Rupe Tarpea; commenti e contro commenti; mangiare pane e veleno; non metterci né sale né olio; non aver né bocca né orecchie; non essere né guelfo né ghibellino; non farsi né in là né in qua ecc.* (cf. Mišetić & Budimir 2022 : 68).

Per questa ricerca è stato utilizzato il corpus dei binomi lessicali estratti da alcuni dizionari fraseologici italiani⁶. L’uso dei binomi dell’italiano contemporaneo è stato esaminato nel web corpus italiano più grande *ItTenTen20*.

⁶ Gli esempi citati saranno seguiti dall’indicazione della fonte in forma abbreviata, ritrovabile nella bibliografia. Perciò, la dicitura Quartu indicherà il *Dizionario dei modi di dire* di Monica Quartu (<https://dizionari.corriere.it/dizionario-modi-di-dire/a.shtml>). La forma Pitt è l’abbreviazione del *Dizionario dei modi di dire* di Giuseppe Pittano (2014) ed infine Sorge verrà utilizzata per indicare il *Dizionario dei modi di dire della lingua italiana* di Paola Sorge (1997). La dicitura NDM invece indicherà *Il dizionario della lingua italiana De Mauro* (<https://dizionario.internazionale.it/>) e la forma DSC *Il Sabatini Coletti - Dizionario della lingua italiana* (https://dizionari.corriere.it/dizionario_italiano/).

Nel Web corpus *ItTenTen20* non si riscontrano ad esempio i seguenti binomi⁷ oppure il loro uso è sporadico e con ciò non soddisfano il criterio della rappresentatività, cioè non sono usati per niente a livello sincronico. Citiamone solo qualche esempio:

- (1) citare testi e pentole [Sorge s.v. *pentola*] (0)
- (2) fare sera e sabato [Sorge s.v. *sera*] (1)
- (3) essere tutto fiori e bacelli [Quartu, Sorge s.v. *fiore*] (2)
- (4) dalla rava alla fava [Quartu s.v. *fava*] (2)
- (5) essere da sella e da soma [Sorge s.v. *sella*] (0)
- (6) arare col bue e l'asino [Quartu, Sorge s.v. *bue*] (1)
- (7) perdere il frutto e il capitale [Sorge s.v. *frutto*] (1)
- (8) non avere né modo né misura [Sorge s.v. *modo*] (1)
- (9) essere tutto voce e penne [Sorge s.v. *voce*] (1)
- (10) mettere qc. fra l'uscio e muro [Quartu, Sorge s.v. *uscio*] (1)
- (11) essere senza puzzoli e senza odori [Sorge s.v. *puzzo*] (1)

Potrebbero essere “fossili dei dizionari” (Sadikaj 2010 : 15) oppure delle modifiche occasionali ossia modificazioni (cf. Mišetić 2022b : § 2.2) che non appartengono alla *langue* e non dovrebbero essere lemmatizzate nei dizionari. L’approccio storico ed areale (cf. Burger 2012; Piirainen 2006), in ambito italiano poco sviluppati, dovrebbero aiutare a chiarire se si abbia a che fare con un arcaismo, un necrotismo ovvero un regionalismo ecc. Questo non esclude l’elaborazione dei dizionari fraseologici dell’italiano letterario, dei dizionari dei regionalismi o degli italiani regionali ecc. Il dizionario fraseologico dell’italiano contemporaneo e la lemmatizzazione più adeguata dei frasemi e dei binomi lessicali nei dizionari d’uso della lingua italiana contemporanea con un approccio sincronico riveste senza dubbio primaria importanza nella lessicografia e nella fraseografia moderna. Il problema dei binomi usati frequentemente a livello sincronico, non lemmatizzati nei dizionari, in particolare nel parlato spontaneo, soprattutto quando il loro uso non si limita a un semplice regionalismo, crea maggiori problemi sia per i dialettofoni che vorrebbero imparare le locuzioni fisse presenti nell’italiano standard sia per tutti gli apprendenti non madrelingua che vorrebbero perfezionare la comprensione, nonché la produzione della lingua italiana. Gli uni e gli altri non riescono a distinguere tra modificazioni e varianti di un binomio lessicale, né a cogliere appieno le condizioni pragmatiche d’uso, dal momento che le locuzioni fisse

⁷ Dov’è necessario, gli esempi estratti dall’Italian Web Corpus *itWaC* o *itTenTen20*, sia le varianti sia le modificazioni, saranno seguiti dal numero delle loro occorrenze, messo tra parentesi tonde.

già lemmatizzate presentano numerose problematiche, in quanto “i significati e le condizioni pragmatiche d’uso risultano sintetizzati e in parte del tutto omessi” (Stumpf 2019 : 115).

3. Binomi lessicali frequenti nell’uso linguistico moderno, non registrati nei dizionari fraseologici né in quelli d’uso

Si manifesta, invece, la tendenza opposta: diversi binomi lessicali trovano ampia conferma d’uso nell’italiano contemporaneo, come evidenziato dalle loro numerose attestazioni nei più estesi corpora digitali. Nel consultare i dizionari fraseologici e d’uso si può constatare con rammarico che questi binomi non hanno il loro lemma fraseologico. Non volendo soltanto criticare i metodi della lessicografia e della fraseografia tradizionale, occorre ciononostante sottolineare che il problema è serio e richiede attenzione e intervento. Per quanto non si intenda respingerli del tutto, occorre ammettere che i metodi tradizionali non sono in grado di descrivere da soli la lingua italiana a livello sincronico. L’ausilio dei metodi della linguistica dei corpora non va più escluso nella compilazione dei dizionari. Per poter descrivere l’uso dei binomi lessicali nell’italiano contemporaneo, va adottato un rigoroso approccio che si basa su evidenze empiriche⁸. Secondo la nostra esperienza e analisi, il purismo eccessivo, le idee erronee sulla salute della lingua, una stilistica non adeguata per l’italiano moderno e l’arbitrarietà nella compilazione dei dizionari, non limitate all’italiano ma riscontrabili anche in altre lingue europee, in particolare nei dizionari fraseologici, sono cause di crescente confusione. Di conseguenza, l’italiano effettivamente usato dai parlanti (cf. Mišetić 2023 : 87) va allontanandosi sempre di più dai lemmi riscontrabili nei dizionari. Citiamone qualche esempio:

- a) Il binomio lessicale *vivo o morto*, altamente frequente nell’uso linguistico contemporaneo, presenta 2973 occorrenze nel corpus italiano *ItTenTen20*. Appare in diverse costruzioni, alcune delle quali potrebbero essere definite varianti del binomio soprammenzionato come ad es.:
 - (12) prendere vivo o morto (25)
 - (13) volere qualcuno vivo o morto (44)
 - (14) ricercato vivo o morto (55)

Tutte queste costruzioni non si ritrovano in alcun dizionario: né nei dizionari fraseologici né in quelli d’uso esaminati (v. Pitt s.v. *vivo e morto*; Sorge s.v. *vivo e morto*; Quartu s.v. *vivo e morto*; NDM s.v. *vivo e morto*).

⁸ Un’eccezione è solo il dizionario d’uso DSC dove viene registrata solo una variante di questo binomio: *cercare qlcu. o vivo o morto* (cf. DSC s.v. *morto*).

- b) Il binomio lessicale *grosso e grasso* (405) non si riscontra in alcun dizionario, né in quelli fraseologici (v. Pitt s.v. *grosso e grasso*; Sorge s.v. *grosso e grasso*; Quartu s.v. *grosso e grasso*) né in quelli d'uso (v. NDM s.v. *grosso e grasso*; DSC s.v. *grosso e grasso*). In termini di numero di occorrenze, il binomio è usato assiduamente a livello sincronico dai parlanti della lingua italiana.

- (15) A Sinistra c'è grossa e grassa confusione.
- (16) Scegliamo, non a caso, i capi di alcune tra le più grosse e grasse aziende farmaceutiche mondiali.
- (17) Uhh che solita sequenza di parole grosse e grasse già sentita da troll e simili...

Questo binomio presenta inoltre carattere reversibile⁹: accanto alla sequenza A e B, si attesta altresì l'ordine inverso B e A (cf. § 4). Ne consegue che la sua lemmatizzazione risulta più complessa rispetto a quella dei binomi irreversibili. Occorre inoltre osservare che esistono numerosi altri esempi, pienamente attestati nell'uso sincronico con centinaia o migliaia di occorrenze nei principali web corpora, ma tuttora assenti dalla lemmatizzazione sia nei dizionari fraseologici sia in quelli d'uso: *uso e abuso* (2.582), *qui e ora* (16.068), *luci e ombre* (15.530), *errori e(d) orrori* (640), *colpito e affondato* (596), *sempre e ovunque* (6.077), *primo ed unico* (7.994), *umiliato e offeso* (296) (cf. Pitt, Sorge, Quartu, NDM, DSC s.v. *uso, abuso, sempre, ovunque, primo, qui, ora, luce, ombra, errore, orrore, umiliato, offeso*, ecc.)¹⁰.

All'inizio, per aggiornare e completare i dizionari fraseologici e d'uso della lingua italiana contemporanea, in particolar modo i lemmi attinenti alle locuzioni fisse, sarebbero d'aiuto le ricerche *corpus based*. Pur in assenza di liste o raccolte dei binomi lessicali finora pubblicate in italiano, si potrebbe incominciare con la verifica su corpora dei binomi raccolti in diverse tesi di dottorato, tesi di laurea, articoli scientifici o nelle raccolte effettuate in tutte le aree d'Italia e fuori Italia. I test psicolinguistici con i parlanti madrelingua competenti e colti potrebbero confermare l'uso dei binomi. In assenza di una metodologia comunemente accettata e di criteri di ricerca omogenei nell'ambito della linguistica dei corpora, la rappresentatività di una locuzione fissa con centinaia o migliaia di occorrenze – in questo caso, di un binomio non registrato nei dizionari – conferma il suo uso a livello sincronico. L'uso sporadico evidenzia, invece, una modificazione, un errore o un lapsus (cf. Mišetić 2023 : § 1). Lo studio dei binomi lessicali implica un lavoro intenso, considerando

⁹ Purtroppo, le ricerche sulla reversibilità dei binomi reversibili risultano ancora scarse, tanto nell'ambito italiano quanto in quello delle altre lingue europee (cf. Kelih 2017; Kelih 2020; Budimir & Mišetić 2023).

¹⁰ Questi esempi sono stati estratti prevalentemente dalla tesi di dottorato di Pietrzak (2016).

che si tratta di una classe fraseologica consistente, ma non illimitata. Si potrebbe ipotizzare che nell'uso contemporaneo esistano numerosi binomi lessicali rappresentativi che meriterebbero una voce autonoma nei dizionari fraseologici e in quelli d'uso.

Occorre altresì sottolineare come, recentemente, sia stata dedicata particolare attenzione ai cosiddetti frasemi schematici, tra cui i binomi lessicali, che possono essere sia indipendenti sia parte di specifiche costruzioni fraseologiche (cf. Ljubičić & Mišetić 2021: 114). Queste costruzioni possono essere citazioni celebri che in questo caso diventano i cosiddetti “modelli di citazione” (Schütz 2023 : 118), ovvero “i frasemi intertestuali riconducibili a una fonte o a un autore e che tendono a funzionare come modelli” (Schütz 2023 : 110). Finora non esiste un modo accettato all'unanimità dai lessicografi per rappresentarli nei dizionari. I frasemi schematici in italiano non hanno un lemma adatto e accettabile che sia “user friendly” e descriva il loro uso. In questo contesto, si evidenzia il binomio *essere o non essere* che fa parte della celebre citazione shakespeariana *essere o non essere, questo è il dilemma* (ingl. *To be, or not to be, that is the question.*)¹¹. Secondo la nostra ricerca nel web corpus *itTenTen20* questo frasema schematico, rappresentato dalla formula *X o non X*, *questo è il dilemma* oppure *X o Y*, *questo è il dilemma* ha nel web corpus *itTenTen20* 1367 occorrenze. Dall'analisi delle prime cento occorrenze emerge con chiarezza che in italiano esso costituisce un modello di citazione, ovvero un frasema schematico¹². La forma canonica *essere o non essere, questo è il dilemma* appare solo 4 volte, mentre lo slot *X o non X* nonché *X o Y* viene riempito 96 volte da diversi esempi, la maggior parte dei quali appare solo un'unica volta¹³, ad esempio:

- (18) **X o non X:** guardare o non guardare, questo è il dilemma; bio o non bio, questo è il dilemma; botox o non botox, questo è il dilemma.
- (19) **X o Y:** bellezza o espressività, questo è il dilemma; doccia o vasca da bagno, questo è il dilemma; emozioni o algoritmi, questo è il dilemma.

La lemmatizzazione dei frasemi schematici, nonché dei binomi lessicali che servono da modelli di citazione, sarà sicuramente una sfida per la fraseografia moderna.

¹¹ Questo binomio è considerato un potenziale candidato per i modelli di citazione da Schütz (2023 : 110). Il suo uso viene esaminato e descritto come frasema schematico (ted. *Modellbildung*) nel web corpus tedesco più grande *DeReCo* da Stumpf (2016). È una celebre frase di William Shakespeare (*Amleto*, Atto 3, scena 1).

¹² Questo frasema schematico è stato trattato estesamente da Schütz (2024 : 11-15).

¹³ Ancorché non esistano criteri omogenei per definire un frasema schematico, in questa sede lo si intende secondo la proposta di Stumpf (2015 : 393), che si rivela di rilevante utilità ai fini del suo riconoscimento, pur senza risolvere tutti i problemi relativi alla demarcazione tra frasemi schematici e modifiche occasionali o modificazioni di un frasema. Per la critica di questo metodo v. Flinz e Mollica (2021: 123) o Mišetić (2023 : 91).

4. (Ir)reversibilità dei binomi lessicali

I binomi lessicali in italiano sono riconoscibili principalmente da una sequenza di componenti rigorosamente definita; tuttavia, circa un terzo di essi compare in due sequenze, risultando reversibili (Mišetić & Budimir 2022 : 70). Nei dizionari esaminati quasi tutti i binomi sono registrati come irreversibili. Il binomio irreversibile avrebbe la struttura generale che permette solo la sequenza $[(C)[A]_x C [B]_x]_y$ ¹⁴, mentre quello reversibile consentirebbe altresì la sequenza $[(C)[B]_x C [A]_x]_y$.

Eccone alcuni esempi dei quali nel Web corpus *ItTenTen20* sono attestate entrambe le sequenze:

- (20) giorno e notte (34.303) – notte e giorno (8.353)
- (21) baci e abbracci (321) – abbracci e baci (72)
- (22) forte e chiaro (764) – chiaro e forte (351)
- (23) gioie e dolori (541) – dolori e gioie (22)
- (24) fare tuoni e fulmini (723) – fare fulmini e tuoni (100)
- (25) felici e contenti (10.207) – contenti e felici (336)
- (26) per mare e per terra (1.029) – per terra e per mare (1.508)
- (27) a torto o a ragione (6128) – a ragione o a torto (752)
- (28) vincitori e vinti (3.084) – vinti e vincitori (735)

Mentre per gli esempi 20, 22, 25, 26, 27, 28 si può affermare con sicurezza che rappresentano una variante topologica¹⁵, per gli esempi 21 e 24 lo si può affermare con un grado minore di certezza, l'esempio 23 *gioie e dolori*, invece, è un binomio quasi irreversibile (v. § 5). Tutti i binomi sopramenzionati sono registrati come irreversibili nei dizionari fraseologici (v. Sorge s.v. *tuono*) e in quelli d'uso (v. NDM s.v. *giorno e notte*, *notte*, *abbraccio*; DCS s.v. *notte*)¹⁶.

Il numero di occorrenze può essere il risultato di errori e lapsus fraseologici (cf. Mišetić 2023 : § 1), di giochi di parole o di qualche modificazione fatta intenzionalmente. I metodi tradizionali e i test con i parlanti nativi potrebbero aiutare a risolvere il problema delle varianti

¹⁴ Questa formula significa che “i due membri del binomio sono rappresentati da A e B (appartenenti a una medesima categoria lessicale X), mentre C sta per l’elemento di congiunzione (che può comparire anche in forma discontinua) e Y rappresenta la categoria in uscita” v. [https://www.treccani.it/enciclopedia/binomi-irreversibili_\(Enciclopedia-dell'Italiano\)/](https://www.treccani.it/enciclopedia/binomi-irreversibili_(Enciclopedia-dell'Italiano)/) (16 / 1 / 2025).

¹⁵ Rispetto al termine *variante topologica* si veda Mišetić (2022: 127) che, nel caso dei binomi, è praticamente il sinonimo della reversibilità o dell’irreversibilità, con il quale si mette in rilievo il fatto della variabilità, finora trascurato nelle ricerche fraseologiche e osservato poco nella compilazione dei dizionari (cf. Mišetić 2022a: § 2).

¹⁶ Alcuni dei binomi esaminati non sono neppure lemmatizzati nei due dizionari d’uso.

diatopiche: la linguistica areale, poco sviluppata in ambito italiano, dovrebbe chiarire se certe sequenze sono usate solo in una varietà regionale della Penisola oppure sono rappresentative per l’italiano contemporaneo. In questo contesto le discussioni su quello che è degno o indegno¹⁷ di fare parte della lingua italiana dovrebbero essere rigettate, dal momento che derivano da una comprensione del tutto errata della funzione dei dizionari fraseologici. Ciò vale anzitutto per la fraseologia, per il fatto che la normazione – ossia “la codificazione nell’ambito delle combinazioni fraseologiche ha avuto meno influsso sulla riduzione delle varianti e sulla formazione della fissità rispetto a quanto accaduto nell’ambito della pronuncia dell’ortografia o della grammatica” (Filatkina 2016 : 15).

5. Lemmatizzazione delle varianti e delle modificazioni, in particolar modo della variazione topologica

Il fraseologo austriaco Emmerich Kelih (2017) propone un approccio che, secondo le ricerche di Mišetić e Budimir (2022), si è dimostrato efficace nel fornire una lemmatizzazione migliorata – seppur non del tutto impeccabile – dei binomi reversibili e irreversibili. Nella loro indagine nel web corpus *itWaC*, Budimir e Mišetić (2022 : § 5) riprendono le proposte di Kelih (2020 : § 2.4) per distinguere tra binomi irreversibili e reversibili. In questa sede, si discute tale teoria con un’attenzione particolare all’esistenza di altri web corpora della lingua italiana, in particolare il web corpus *ItTenTen20*, significativamente più ampio rispetto al già citato *itWaC*, nel quale risulta più complesso stabilire se una combinazione fraseologica possa considerarsi rappresentativa o meno¹⁸.

Per risolvere il problema della reversibilità o dell’irreversibilità di un binomio lessicale nell’uso contemporaneo della lingua italiana potrebbe essere utile il calcolo dell’indicatore dell’irreversibilità proposto da Kelih (2017 : 197) che offre alcuni metodi statistici appropriati.

5.1. Un metodo semplice per calcolare l’indicatore della reversibilità (p_1)

Il primo metodo sarebbe un metodo semplice per calcolare l’indicatore della reversibilità (p_1). La variazione topologica, ovvero (ir)reversibilità dei binomi lessicali:

¹⁷ Il rifiuto di criteri puramente prescrittivi non implica, ovviamente, che i dizionari fraseologici debbano accogliere indiscriminatamente ogni tipo di sequenza linguistica; l’esclusione di espressioni apertamente offensive, volgari o blasfeme – in contrasto con le norme etiche, culturali o religiose largamente condivise – risulta anzi coerente con tali presupposti.

¹⁸ V. Moon (2008); Parizoska (2022: § 3).

Indicatore della reversibilità (Kelih 2017 : 197):

$$\begin{aligned} n_1 &= \text{frequenza del binomio con la sequenza A e B} \\ n_2 &= \text{frequenza del binomio reversibile con la sequenza B e A} \\ n &= n_1 + n_2 \\ p_1 &= n_1 / n \end{aligned}$$

L'indicatore p_1 nel caso $n_1 = n_2$ ha un valore minimo pari a 0,5, e valore 1 se $n_1 = n$.

a) Binomi irreversibili

Il binomio lessicale (*andare*) *d'amore e d'accordo* (itTenTen20)

$$\begin{aligned} n_1 &= 2232 \\ n_2 &= 9 \\ n &= n_1 + n_2 \\ n &= 2241 \\ p_1 &= n_1 / n \\ p_1 &= 0,995 \end{aligned}$$

In questo caso si può parlare con sicurezza di un binomio irreversibile. Pochi esempi della sequenza B e A possono essere delle modificazioni, giochi di parole o lapsus fraseologici.

b) Binomi reversibili

Il binomio lessicale *per terra e per mare* (itTenTen20)

$$\begin{aligned} n_1 &= 1508 \\ n_2 &= 1029 \\ n &= n_1 + n_2 \\ n &= 2537 \\ p_1 &= n_1 / n \\ p_1 &= 0,594 \end{aligned}$$

A nostro avviso, in questo caso si potrebbe parlare di un binomio reversibile, nonostante manchi una classificazione comunemente accettata in fraseologia e fraseografia che definisca i criteri per distinguere binomi reversibili da irreversibili.

5.2. Uso del cosiddetto test asintotico

Il secondo metodo, più complesso rispetto a quello presentato nel paragrafo precedente, viene definito *test asintotico* ed è stato preso dalla statistica matematica (Kelih 2017 : 199):

$$u = \frac{p_1 - 0.5}{\sqrt{p_1(1 - p_1)/n}}$$

Secondo questa formula, è possibile verificare se l'indicatore dell'irreversibilità si allontana dal valore 0,5.

Se il valore u è maggiore di 1,96, l'irreversibilità è minima. Se, invece, u è inferiore a 1,96, si potrebbe parlare di una sequenza significativamente irreversibile (cf. Mišetić & Budimir 2022: 74).

5.3. Proposte per la lemmatizzazione adatta della reversibilità dei binomi a livello sincronico

In considerazione di quanto precedentemente menzionato, emerge chiaramente che la lemmatizzazione della variazione topologica è necessaria nei dizionari moderni; tuttavia, rappresenta una vera e propria sfida per la lessicografia e la fraseografia contemporanee. Ancorché nelle lingue europee esistano tradizioni diverse attinenti alle marche d'uso nei dizionari (cf. Mišetić 2025: 141-142), finora non sono state introdotte neppure marche d'uso specifiche per i binomi lessicali irreversibili o reversibili. Grazie all'ausilio delle tecnologie moderne e delle ricerche su corpus questo tentativo potrebbe essere realizzato. Kelih (2020) propone la classificazione di binomi irreversibili, raramente e altamente reversibili (ted. *irreversibel*, *geringfügig reversibel*, *hochgradig reversibel*), mentre Mišetić e Budimir (2022 : 74) discutono di binomi completamente irreversibili, parzialmente irreversibili, reversibili che danno preferenza ad una sequenza e completamente reversibili. Queste marche d'uso, inoltre, sarebbero utili per le analisi diacroniche in futuro, dal momento che "tutti i sistemi linguistici, quelli fraseologici inclusi, sono sottoposti a modifiche minori o maggiori e non si può parlare di stabilità assoluta" (Budimir 2020 : 80). A questo proposito, le ricerche diacroniche in merito all'irreversibilità dei binomi lessicali dimostrano che, almeno nel tedesco medievale, questa non è una caratteristica tipica delle costruzioni binomiali (cf. Burger 2012 : § 4.2). In tal senso, sarà interessante osservare lo sviluppo dei binomi reversibili in futuro. La questione se sia una caratteristica tipica dei binomi lessicali oppure se il principio di economia, spingendo ad una riduzione delle differenze in una lingua moderna a livello sincronico, porterà alla scomparsa delle varianti topologiche, potrebbe essere approfondita e grazie all'approccio empirico potrebbero essere tratte conclusioni convincenti basate sull'osservazione diretta dei dati.

6. Osservazioni conclusive

Secondo le nostre ricerche sia nei dizionari fraseologici sia in quelli d'uso italiani sono registrati numerosi binomi lessicali non ritrovabili nei web corpora *itWaC* e *ItTenTen20* (cioè non usati nell'italiano contemporaneo), oppure il loro uso è sporadico e con ciò non rappresentativo. Il numero delle loro occorrenze non è sufficiente per includerli nella lista dei binomi usati a livello sincronico. Si riscontrano, invece, binomi frequenti, usati a livello sincronico e attestati nei web corpora *itWaC* e *itTenTen20*, che non sono lemmatizzati nei dizionari. La lemmatizzazione delle varianti e la distinzione tra varianti e modificazioni nelle lingue europee (cf. Mišetić 2022a: § 2.2) non di rado è insufficiente dal punto di vista fraseografico moderno: difatti, l'elaborazione fraseografica dei binomi reversibili in ambito italiano finora non è neppure stata oggetto di ricerca. La lemmatizzazione della variazione topologica, cioè della reversibilità dei binomi lessicali, non viene nemmeno discussa nella letteratura fraseologica e la loro elaborazione fraseografica presenta nella lessicografia e nella fraseografia italiana un'eccezione assoluta. Pur non risolvendo, a nostro avviso, tutti i problemi legati alla lemmatizzazione dei frasemi nell'italiano contemporaneo, la linguistica dei corpora potrebbe comunque contribuire a colmare il divario tra l'uso sincronico dei binomi lessicali e la loro codificazione nei dizionari.

Bibliografia

- Antonelli, G. (2016), *L'italiano nella società della comunicazione 2.0*, Società editrice il Mulino, Bologna.
- Autelli, E., Konecny, C. (2015), “Combining Lexicography with Second-Language Didactics : The Case of the Bilingual Collocations Dictionary Kollokationen Italienisch-Deutsch”, in Karpova, O. M., Kartashkova, F. I. (eds), *Life Beyond Dictionaries*, Cambridge Scholars Publishing, Cambridge, p. 185-198.
- Budimir, I. (2020), *Frazemi prve hercegovačke hrvatske periodike na prijelazu iz 19. u 20. stoljeće i njihova leksikografska obradba*, Fram Ziral, Mostar.
- Budimir, I., Mišetić, D. (2020), “Utjecaj ograničenja prominentnosti slogova na slijed sastavnica binoma u talijanskome i hrvatskome jeziku”, *Identiteti – kulture – jezici*, Filozofski fakultet Sveučilišta u Mostaru, Mostar, p. 64-77.
- Budimir, I., Mišetić, D. (2022), “O problemima leksikografske obradbe frazema u hrvatskome jeziku”, in Marković, I., Nazalević Čučević, I., Gligorić, I. M. (eds), *Rijeći o riječi i Rijeći: zbornik u čast Zrinki Jelaska*, Disput, Zagreb, p. 639-652.
- Budimir, I., Mišetić D. (2023), “Reveribilni binomi u talijanskome i hrvatskome jeziku u svjetlu multifaktorske perspektive” (usmeno izlaganje), Književnost, umjetnost, kultura između dviju obala Jadrana / Letteratura, arte, cultura tra le due sponde dell'Adriatico, Sveučilište u Zadru, Zadar 5- 6 listopada 2023. godine. <https://www.croris.hr/crosbi/publikacija/prilog-skup/803072>.

- Burger H. (2012), "Alte und neue Fragen, alte und neue Methoden der historischen Phraseologie", in Filatkina, N. et al. (eds), *Aspekte der historischen Phraseologie und Phraseographie*, Universitätsverlag Winter, Heidelberg, p. 1-21.
- Burger, H. (2015), *Phraseologie-Eine Einführung am Beispiel des Deutschen*, Erich Schmidt Verlag, Berlin.
- Burger, H., Buhofer A., Salm, A. (1982), *Handbuch der Phraseologie*, De Gruyter, Berlin-New York.
- Cresti, E., Panunzi, A. (2013), *Introduzione ai corpora dell'italiano*, Il Mulino, Bologna.
- Filatkina, N. (2016), "Wie fest sind feste Strukturen? Beobachtungen zu Varianz in historischen Wörterbüchern und Texten", in Borek L., Rapp, A. (eds), OPAL 2, Institut für Deutsche Sprache, Mannheim, p. 7-27.
- Flinz, C., Mollica, F. (2021), "Keine Antwort ist auch eine Antwort: analisi intra- e interlinguistica", *Lingue e Linguaggi*, 46, p. 109-140.
- Hallsteinsdóttir, E. (2006), "Phraseographie", *Hermes – Journal of Language and Communication Studies*, 36, p. 91-128.
- Kelih, E. (2017), "Reversibilität von Paarformeln und Binomialen: Methodologische Aspekte", *Glottotheory*, 8/2, p. 191-201.
- Kelih, E. (2020), "Gebrauchshäufigkeit und Reversibilität von Binomialen (Paarformeln) im Slowenischen", *Wiener Slavistisches Jahrbuch (Neue Folge)*, 8, p. 104-121.
- Konecny, C., Hallsteinsdóttir, E., Kacjan, B. (2013), "Zum Status quo der Phraseodidaktik: Aktuelle Forschungsfragen, Desiderata und Zukunftsperspektiven", in Konecny, C., Hallsteinsdóttir, E., Kacjan, B. (eds), *Phraseologie im Sprachunterricht und in der Sprachendidaktik. / Phraseology in language teaching and in language didactics*, Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta [Zora; 94], Maribor, p. 153-172.
- Korhonen, J. (ed.) (2011), *Phraseologie und Lexikografie. Phraseologismen in ein- und zweisprachigen Wörterbüchern mit Deutsch*, Proverbium (= Supplement series of Proverbium 32), Burlington, Vermont.
- Ljubičić, M., Mišetić D. (2021), "Struttura dei binomi lessicali in italiano e in croato", in Saržoska, A. (ed.), *Atti del convegno internazionale. L'italianistica nel terzo millennio: le nuove sfide nelle ricerche linguistiche, letterarie e culturali. 60 anni di studi italiani all'università SS. Cirillo e Metodio di Skopje*, Facoltà di Filologia "Blaže Koneski", Skopje, p. 105-118.
- Ljubičić, M., Mišetić, D. (2023), "Analisi semantica dei binomi lessicali in italiano e in croato", in Spahić, E., Radeljković, I., Osmanović, L. (eds), *Zbornik radova: 70 godina Odsjeka za romanistiku*, Univerzitet u Sarajevu – Filozofski fakultet, Sarajevo, p. 28-38.
- Luque Nadal, L. (2017), "Aspectos fraseológicos y culturales de los co-compuestos o binomios léxicos", *Language Design*, 19, p. 194-204.
- Mišetić, D. (2018), "Slijed sastavnica binoma u talijanskome i hrvatskome jeziku", *Hum*, XIII,/20, p. 308-325.
- Mišetić, D. (2022a), "Variazione dei frasemi nelle lingue europee alla luce delle ricerche attuali", *Studia Romanica et Anglicana Zagabiensia*, 67, p. 123-135, DOI: <https://doi.org/10.17234/SRAZ.67.9>
- Mišetić, D. (2022b), "Modifiche occasionali delle combinazioni fraseologiche nei web corpora: le modificazioni come uso tipico dei frasemi", *Phrasis: Rivista di studi fraseologici e paremiologici*, 6, p. 82-92.

- Mišetić, D. (2023), “C’è del marcio in Danimarca - e non solo là: sui problemi della distinzione tra modificazioni e frasemi schematici nei web corpora”, *Phrasis: Rivista di studi fraseologici e paremiologici*, 7, p. 82-96.
- Mišetić, D. (2025), *Frazemi biblijskoga podrijetla u talijanskome i hrvatskome jeziku*, Synopsis, Sarajevo-Zagreb.
- Mišetić, D., Budimir, I. (2022), “(I)reverzibilnost leksičkih binoma u talijanskome i hrvatskome jeziku”, *Sponde: Rivista di lingue, letterature e culture tra le due sponde dell’Adriatico*, 2/1, p. 65-78, DOI: <https://doi.org/10.15291/sponde.4089>
- Moon, R. (2008), “Conventionalized as-similes in English: A problem case”, *International Journal of Corpus Linguistics*, 13/1, p. 3-37.
- Parizoska, J. (2022), “Idiom modifications: What grammar reveals about conceptual structure”, *Lexis – Journal in English Lexicology*, 19, DOI: <https://doi.org/10.4000/lexis.6293>
- Pietrzak, J. B. (2016), *Binomios fraseológicos en el italiano contemporáneo*, Tesis doctoral, Universidad Complutense de Madrid, Madrid.
- Piirainen, E. (2006), “Phraseologie in arealen Bezügen: ein Problemaufriss”, *Linguistik online*, 27/2, p. 195-218.
- Pilz, K. D. (2002), “Vorschläge für ein Phraseolexikon der deutschen Sprache oder Vorschläge für ein Lexikon der deutschen Phraseme / Phraseologismen”, in Hartmann, D., Wirrer, J. (eds), *Wer A sagt, muss auch B sägen. Beiträge zur Phraseologie und Sprichwortforschung aus dem Westfälischen Arbeitskreis*, Schneider Verlag Hohengehren, Baltmannsweiler, p. 299-311.
- Sadikaj, S. (2010), *Metaphorische Konzepte in somatischen Phraseologismen des Deutschen und Albanischen eine kontrastive Untersuchung anhand von Herz- und Hand-Somatismen*, Univ. Diss., Würzburg.
- Schütz, J. (2023), “Modified winged words from a typological and terminological perspective in English, German and Italian phraseological studies: *Ed è subito caos*”, *Phrasis: Rivista di studi fraseologici e paremiologici*, 7, p. 110-120.
- Schütz, J. (2024), “Phrasemes as prototypical linguistic micro-representations of the world: the role of values in the propagation of Winged Words”, *Lexis* [Online], 24, URL: <http://journals.openedition.org/lexis/8565>, DOI: <https://doi.org/10.4000/12cvs>
- Stumpf, S. (2015), *Formelhafte (Ir-)Regularitäten. Korpuslinguistische Befunde und sprachtheoretische Überlegungen*, Peter Lang, Frankfurt am Main.
- Stumpf, S. (2016), “Modifikation oder Modellbildung? Das ist hier die Frage – Abgrenzungsschwierigkeiten zwischen modifizierten und modellartigen Phrasemen am Beispiel formelhafter (Ir-)Regularitäten”, in Grewendorf, G. et al. (eds), *Linguistische Berichte*, 247, p. 317-342, DOI: https://doi.org/10.46771/2366077500247_3
- Stumpf, S. (2019), “Phraseografie und Korpusanalyse”, *Linguistik Online*, 96/3, p.115-131.
- Wälchli, B. (2005), *Co-Compounds and Natural Coordination*, Oxford.

Dizionari cartecei

- Pitt = Pittano, G. (2014), *Dizionario dei modi di dire*, Zanichelli editore, Bologna.
- Sorge = Sorge, P. (1997), *Dizionario dei modi di dire della lingua italiana*, Newton & Compton editori, Roma.

Dizionari online e altre fonti sitografiche

DSC = *Il Sabatini Coletti – Dizionario della lingua italiana*, https://dizionari.corriere.it/dizionario_italiano/ (11/2/2025).

NDM = *Il nuovo De Mauro*, <https://dizionario.internazionale.it/> (10/2/2025).

Quartu = Quartu, M., *Dizionario dei modi di dire*, <https://dizionari.corriere.it/dizionario-modi-di-dire/a.shtml> (29/10/2024).

[https://www.treccani.it/enciclopedia/binomi-irreversibili_\(Encyclopedia-dell'Italiano\)/](https://www.treccani.it/enciclopedia/binomi-irreversibili_(Encyclopedia-dell'Italiano)/) (16/1/2025).

Web corpora

itTenTen20 = Italian Web 2020, disponibile sulla piattaforma SketchEngine (URL: https://app.sketchengine.eu/#concordance?corpname=preloaded/ittenten20_fl1) (2/11/2024).

itWaC = Italian web corpus (NoSketchEngine) (URL: <http://nlp.ffzg.hr/resources/corpora/hrwac/>) (25/10/2024)