Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto

di *Isabella Chiari*

1
Il fatto linguistico in lessicografia

Il fatto linguistico e il dato empirico in lessicografia hanno spesso uno statuto mediato da esigenze applicative e didattiche per le quali ci si muove in una linea sottile tra normatività, prescrittività, modelli su aspetti diversi dell'oggetto rappresentato (e delle sue dimensioni di variazione) e un approccio puramente descrittivo più vicino alla ricerca linguistica. Tale mediazione è spesso dovuta all'essenza applicativa e orientata sull'utente e ai diversi modelli linguistici che emergono nella vita sociale di una comunità linguistica. Mentre le diverse aree della ricerca linguistica sono state toccate e profondamente influenzate dal dibattito sulla natura del dato linguistico, sul suo posarsi su attestazione ed evidenza osservabile o su competenze e intuizioni del parlante (ideale o reale), la lessicografia sembra aver attraversato i decenni dominati dal punto di vista generativo relativamente indenne[1]. Il dibattito sul valore del giudizio dei parlanti e del suo rapporto con il dato attestato è emerso, invece, tardivamente, con la nascita da una parte della lessicografia basata su corpora negli anni Ottanta e dall'altra soprattutto con la comparsa dei primi progetti di lessicografia collaborativa online nell'ultimo decennio.

In lessicografia tuttavia alla dicotomia intuizione/evidenza empirica si aggiunge una esigenza terza che riguarda il ruolo giocato da un modello linguistico, solitamente coincidente con una varietà o una gamma di varietà, fortemente legato a fattori di tipo sociale, culturale, politico e spesso anche socio-economico.

Dal punto di visto storico i primissimi "dizionari", ossia le liste lessicali mono e multilingui di Uruk (dal IV millennio a.C.), di Ebla o Ugarit (dal III millennio a.C.), pur mancando della struttura tipica micro e macro-strutturale dei dizionari attuali, condividevano un obiettivo concreto e descrittivo più che normativo, contenendo liste multilingui delle parole meno comuni o straniere, liste ad uso pratico e senza pretesa di descrizione scientifica, organizzate per forma

1. Cfr. P. Hanks, *Evidence and intuition in lexicography*, in *Meaning and lexicography*, a cura di J. Tomaszczyk, John Benjamins Publishing Company, Amsterdam 1990, pp. 31-41.

o per argomento. Accanto a liste descrittive compaiono, soprattutto in chiave didattica, anche liste prescrittive: basti pensare all'*Appendix Probi*. La normatività è inoltre effettivamente presente nella storia della lessicografia occidentale in senso stretto (esempi sono i numerosi casi di lessicografia accademica: *Vocabolario degli Accademici della Crusca* (1612), *Dictionnaire de l'Académie française* (1694), *Diccionario de la lengua castellana* (1726-39), ed è ben presente anche nelle aspettative degli utenti dei dizionari d'uso di impostazione contemporanea). Mentre nella normatività si riflette tuttavia una idealizzazione della lingua e delle sue manifestazioni, tale idealizzazione tuttavia può esser fatta riposare o su un principio introspettivo (la competenza di un parlante o di un insieme di parlanti definito socio-culturalmente) oppure su un principio empirico fondato sulla raccolta di attestazioni appartenenti a una data varietà.

Anche i dizionari a impostazione più descrittiva, a partire dal celebre dizionario di Samuel Johnson, tuttavia, non si sottraggono allo stesso gioco. Se da una parte tentano spesso di cogliere la lingua nelle sue varietà dandone conto in diversi modi (ordinamento delle accezioni, composizione della glossa, selezione degli esempi, aggiunta di marche d'uso di natura quantitativa e qualitativa) dall'altra inevitabilmente ricadono in una forma più sottile di idealizzazione che consiste nella creazione e selezione di esempi fittizi, nella individuazione di schemi astratti costituiti da porzioni brevi di frasi, nella selezione qualitativa delle voci secondo criteri non sempre ben definiti ed espliciti.

Se la normatività in quanto tale è stata abbandonata, almeno nella sua forma più esplicitamente censoria e purista, pressoché in tutti i dizionari contemporanei della lingua italiana e delle lingue occidentali[2], tuttavia una esigenza normativa[3] rimane spesso nella scelta di includere o non includere certe varianti lessicali e certe forme colloquiali e, soprattutto, nella gerarchia con le quali le forme in competizione vengono presentate al lettore. Nonostante una grandissima attenzione al fatto linguistico empirico anche nella tradizione lessicografica anglosassone è emersa negli ultimi anni una idea del compito lessicografico che costituisce una sorta di mediazione tra descrizione empirica e interpretazione/giudizio come risulta ad esempio nella posizione assunta da Patrick Hanks che definisce il compito del lessicografo quello di «catturare a parole ciò che è *convenzionale* in una lingua»[4], tale convenzione è certo pragmatica e si applica a un modello dinamico.

2. Non si può dire lo stesso, ad esempio, della lessicografia araba attuale, nella quale permane una tendenza verso il modello dell'arabo classico, motivato da ragioni storiche e culturali come la forza dell'esempio ancora vivo del Corano, e da ragioni sociolinguistiche legate all'alto tasso di analfabetismo e alla forte diglossia della grande maggioranza dei paesi arabi.

3. Si sta affermando nella ricerca lessicografica anglosassone il termine "lessicografia proscrittiva" (*proscriptive*) da intendere come il caso in cui si attua una descrizione selettiva e il rapporto con la base empirica mira a raccomandare alcune varietà rispetto ad altre. La "lessicografia prescrittiva", invece, vieta alcune varianti e ne prescrive altre e non ha un rapporto diretto con la base empirica. Cfr. H. Bergenholtz, *User-oriented understanding of descriptive, proscriptive and prescriptive lexicography*, in "Lexikos", XIII, 2003, 1, pp. 65-80.

4. Cfr. Hanks, *Evidence and intuition in lexicography*, cit., p. 32.

In questa mediazione, dovuta in parte al fatto che il dizionario è un'opera descrittiva ma finita, e in parte al suo essere *arte* nel senso etimologico, giocano fattori diversi con i quali il lessicografo è chiamato a identificare e selezionare il fatto linguistico che ritiene rilevante. Hanks individua principalmente due fattori, che spesso contrastano nelle valutazioni di opportunità: un fattore di salienza psicologica (per cui colpisce l'inusuale) e un fattore di salienza sociale (per cui emerge ciò che è comune)[5]. Ma tra i due fattori ne esiste un terzo che si colloca in uno spazio intermedio e che ha a che fare con la modalità di apprendimento del nucleo della lingua e ciò che stereotipicamente nel sentire comune si considera centrale per descrivere il significato delle parole.

Numerosi studi, a partire dagli anni Ottanta, si sono concentrati sull'osservazione degli usi che gli utenti fanno di un dizionario, mediante diverse tipologie di strumenti: questionari, protocolli di sperimentazione e test, analisi dei dati di registro dei dizionari elettronici online[6]. Quest'area può essere considerata oggi una delle principali aree di ricerca della lessicografia contemporanea. La stessa comparsa di questo settore di ricerca è indice di un nuovo modo di misurarsi con il dizionario in cui il centro è l'utente (il "noto sconosciuto" secondo Wiegand[7]) e non il modello linguistico scelto. Se da una parte le ricerche sugli usi dei dizionari sono centrate sull'individuazione di bisogni da soddisfare e indagare mediante questionari, è emersa chiaramente una insufficienza di questo tipo di strumento (poiché fondato sul giudizio soggettivo più che su reale comportamento di ricerca) o, se si preferisce, la necessità di integrare a questionari sperimentazioni in contesti naturali e soprattutto la necessità di analisi dei dati estratti dalle ricerche frequenti sui dizionari online. In particolare l'uso dei dati di registro delle ricerche (*log files* o *logs*), come si vedrà nel § 2.3, appare uno strumento promettente per comprendere le tipologie, fasce e caratteristiche dei lemmi più cercati dagli utenti[8]. Un ulteriore elemento di

5. Ivi, p. 35.

6. Per ragioni di spazio segnalerò soltanto lavori di rassegna più recenti: B. Laufer, M. Kimmel, *Bilingualised dictionaries: How learners really use them*, in "System", XXV, 1997, 3, pp. 361-9; J. H. Hulstijn, B. T. S. Atkins, *Empirical research on dictionary use in foreign-language learning: survey and discussion*, in *Using L2 dictionaries: Studies of dictionary use by language learners and translators*, a cura di B. T. S. Atkins, Niemeyer Verlag, Tübingen 1998, pp. 7-19; S. Tarp, *Reflections on lexicographical user research*, in "Lexikos", XIX, 2009, 1, pp. 275-96; S. Verlinde, J. Binon, *Monitoring dictionary use in the electronic age*, in *Proceedings of the XIV Euralex Intenational Congress*, Fryske Akademy, Leuven 2010, pp. 1144-51; H. A. Welker, *Dictionary Use. A general survey of empirical studies*, s.e., s.l. 2010; R. Lew, *Studies in Dictionary Use: Recent Developments*, in "International Journal of Lexicography", XXIV, 2011, 1, pp. 1-4.

7. H. E. Wiegand, *Einige grundlegende semantisch-pragmatische Aspekte von Wörterbucheinträgen. Ein Beitrag zur praktischen Lexikologie*, in "Kopenhagener Beiträge zur Germanistischen Linguistik", XII, 1977, pp. 59-149.

8. Cfr. G. M. De Schryver, D. Joffe, *On how electronic dictionaries are really used*, in *Proceedings of the Eleventh EURALEX International Congress*, a cura di G. Williams, S. Vessier, Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud, Vannes 2004, pp. 187-96; H. Bergenholtz, M. Johnsen, *Log files as a tool for improving Internet dictionaries*, in "Hermes", XXXIV, 2005, pp. 117-41; H. Bergenholtz, M. Johnsen, *Log files can and should be prepared for a functionalistic approach*, in "Lexikos", XVII, 2007, pp. 1-21.

riflessione, ancora non esplorato, e su cui sarebbe opportuno avere indicazioni d'uso, è la eventuale differenza nell'uso (usi) tra dizionario cartaceo ed elettronico online.

Ciò che per il lessicografo costituisce il dato è tuttavia un oggetto sfuggente e multiforme. Nella storia della lessicografia sono stati di volta in volta privilegiati diversi strumenti, tra questi i più comuni sono: *a*) l'introspezione: *a.1*) l'introspezione del lessicografo; *a.2*) l'introspezione dell'utente ordinario; *b*) l'uso del dizionario: *b.1*) l'analisi di questionari sull'uso dei dizionari; *b.2*) l'analisi del comportamento di utenti in situazioni sperimentali; *b.3*) l'analisi del comportamento di utenti in situazioni reali; *c*) i riferimenti: *c.1*) la descrizione proposta in altre opere lessicografiche (mono o plurilingui); *c.2*) la descrizione proposta in opere di riferimento come grammatiche, lavori didattici e pubblicazioni scientifiche descrittivi; *d*) le attestazioni: *d.1*) l'analisi di esempi d'uso selezionati o casuali tratti da selezioni di testi; *d.2*) l'analisi di esempi d'uso estratti da corpora di riferimento esistenti o costruiti *ad hoc* per fornire la base empirica dell'opera lessicografica.

L'uso di uno o più di questi strumenti caratterizza dunque radicalmente l'immagine e la funzione che le singole opere lessicografiche offrono al lettore e, come si vedrà, pongono problemi diversi al lessicografo nel momento della stesura delle diverse componenti della voce.

2
Le prospettive della lessicografia elettronica

La lessicografia elettronica è un oggetto complesso che raccoglie prodotti di natura, provenienza, e applicazione assai diversa: *a*) *dizionari informatizzati* (versioni online di autorevoli dizionari cartacei); *b*) *dizionari elettronici* direttamente creati per essere distribuiti esclusivamente online[9]; *c*) *dizionari collaborativi* creati da utenti ordinari in progetti volontari; *d*) *strumenti di lessicografia computazionale*, detti anche *dizionari macchina*, che sono pensati come database lessicali o basi di conoscenze finalizzate non tanto alla consultazione da parte di utenti[10], ma all'uso e integrazione in applicazioni computazionali; *e*) *aggregatori di fonti lessicografiche* (come dictionary.com e thefreedictionary.com). In queste tipologie il dato empirico è trattato in modo diverso e soggiace ai diversi scopi dell'applicazione stessa. In questo contributo ci si concentrerà sulle prime tre tipologie lessicografiche pensate per fornire un servizio direttamente all'utente e per rispondere a specifici problemi di competenza linguistica in produzione e ricezione.

9. Il più noto esempio è il tesauro dell'inglese Wordnet, cfr. C. Fellbaum, *WordNet: an electronic lexical database*, MIT Press, Cambridge (MA) 1998; C. Fellbaum, *WordNet and wordnets*, in *Encyclopedia of Language and Linguistics*, a cura di K. Brown, Elsevier, Oxford 2005, pp. 665-70.

10. Sulle risorse lessicali in chiave computazionale, si veda una breve rassegna in I. Chiari, *Linguistic resources and machine translation trends for the Italian language: overview and perspectives*, in *Atti del Convegno Language Translation Automation Conference (LTAC)*, a cura di V. Cannavina, A. Fellet, The Big Wave, Roma 2012, pp. 105-23.

La questione del modo con il quale i dizionari si misurano con il fatto linguistico è divenuta centro di notevole attenzione dalla comparsa, nella lessicografia britannica innanzitutto, dei cosiddetti *corpus-based dictionaries*, dizionari basati su corpora; il dibattito si è successivamente spostato su aspetti relativi a un cambiamento (integrazione) in atto relativo al mezzo di trasmissione dell'opera lessicografica, che nel tempo è diventata da cartacea, anche digitale, prima in forma statica su supporti come CD-ROM o DVD e successivamente in forma anche dinamica nella lessicografia elettronica online. L'aggiunta del canale online non ha tuttavia, in nessun caso finora registrato, prodotto la scomparsa della versione cartacea. Anzi, è noto il caso del Merriam-Webster che nel 1996 ha lanciato la versione online del *Collegiate Dictionary* e *Collegiate Thesaurus* (http://cougar.eb.com/), primo dizionario interamente messo a disposizione degli utenti online, ottenendo milioni di accessi e un aumento delle vendite dell'edizione cartacea che nel solo 2003 è cresciuto del 17%. L'editore si è inoltre fatto promotore nel 2005 di un dizionario aperto[11] con il quale gli utenti segnalano nuove possibili voci per il dizionario e ne indicano caratteristiche e usi.

Così anche per l'italiano uno dei primi dizionari cartacei ad avere una versione gratuita online fu il *Dizionario della lingua italiana* (DMP, 2000)[12], curato da Tullio De Mauro, ora rimosso dal Web, seguito da Treccani, che ha messo a disposizione degli utenti un portale con il *Vocabolario* e l'*Enciclopedia* (www.treccani.it), e, tra gli altri il Sabatini-Coletti[13] tramite il sito del quotidiano "Corriere della Sera" (http://dizionari.corriere.it/dizionario_italiano/), il *Dizionario di italiano* di Aldo Gabrielli[14] tramite il sito della "Repubblica" (http://dizionari.repubblica.it/italiano.php).

2.1. Il dato empirico nella lessicografia basata su corpora

La diffusione di corpora di riferimento per le principali lingue occidentali soprattutto negli anni Novanta ha permesso alla lessicografia di affinare l'uso dell'attestazione in due direzioni nuove. Da una parte si è potuta centrare la rappresentazione della lingua sulle sue varietà e sull'uso concreto, parlato e scritto, dei suoi utenti, estraendo *pattern* sintattici e lessicali, collocazioni e polirematiche, esempi autentici e frequenti. In secondo luogo la disponibilità di corpora di centinaia di milioni di parole (esemplari per la lessicografia britannica il *British National Corpus* e la *Bank of English*) ha permesso di tener conto anche degli effetti della frequenza (e dispersione) nell'uso delle parole in diverse varietà e tipologie, consentendo un nuovo ordinamento degli elementi della voce, delle accezioni, e l'identificazione più certa, fondata e sperimentalmente ripetibile di marche d'uso.

11. Il dizionario aperto di Merriam Webster si raggiunge alla pagina http://www3.merriam-webster.com/opendictionary/.

12. T. De Mauro, *Il dizionario della lingua italiana*, Paravia, Torino 2000.

13. F. Sabatini, V. Coletti, *Il Sabatini Coletti. Dizionario della lingua italiana*, Rizzoli, Milano 2011.

14. A. Gabrielli, *Grande dizionario Hoepli italiano*, Hoepli, Milano 2011.

L'accesso ai corpora ha anche messo in questione e stimolato la riflessione sui criteri stessi di selezione dei lemmi. Il ricorso a grandi corpora di riferimento ha permesso di notare, ad esempio, come una grandissima parte dei lemmi censiti nei dizionari non compaiano quasi mai nei corpora. Questo fatto ha un duplice valore su cui riflettere, da una parte spinge a una razionalizzazione delle scelte relative alla selezione dei lemmari, dall'altra tuttavia mostra limiti metodologici ineludibili che dipendono dalla distribuzione statistica del lessico in corpora di riferimento (sempre finiti per quanto ampi e comunque disegnati con lo scopo di essere in qualche modo rappresentativi di una o più varietà di lingua). Il disegno dei corpora è dunque già una prima intrusione di un criterio selettivo a priori sull'oggetto linguistico da indagare. Questo fatto è peraltro ineliminabile e insito nel metodo stesso. È la base empirica stessa a mostrare una prospettiva inevitabilmente parziale sull'oggetto empirico che intende rappresentare.

Non risolve, anzi aggrava il quadro, l'idea di usare il Web nel suo complesso come base empirica, come l'uso di Google per verificare e quantificare attestazioni, poiché il ricorso al Web (corpus sbilanciato, sporco, ambiguo, ridondante) non permette una valutazione quantitativa e dunque non permette la valutazione sulla relativa frequenza dei fenomeni in varietà diverse (per alcuni è esso stesso specifica varietà): è in sostanza una banca disordinata di esempi, peraltro non casuali[15].

Esiste inoltre anche il problema opposto. Per lemmi di frequenza alta o media, disponendo di corpora di diverse centinaia di milioni di parole, l'operazione di selezione e di elaborazione di una voce a partire da decine di migliaia di esempi estratti da corpora può diventare un'operazione non solo molto laboriosa ma anche intensamente selettiva, fortemente interpretativa e dunque largamente basata sull'interrelazione tra dato e intuizione linguistica. Tale intuizione agisce libera di ordinare e discriminare i *pattern* tipici di un dato lessema in determinati contesti, di identificare le sue caratteristiche socio-pragmatiche (legate ai contesti d'uso, ai registri, alle varietà, alle connotazioni), ma necessita di una esplicitazione e definizione degli aspetti teorico-linguistici soggiacenti all'elaborazione della voce lessicografica (dei principi della cosiddetta *metalessicografia*) proprio per la natura del confronto con il dato grezzo[16]. In questo senso la lessicografia tradizionale ha trascurato, con poche eccezioni come il caso del GRADIT[17], la ne-

15. Sulle potenzialità e i limiti del Web come *corpus* e soprattutto sull'uso dei motori di ricerca per l'estrazione di dati linguistici, cfr. A. Kilgarriff, M. Baroni, *Proceedings of the 2nd International Workshop on Web as Corpus*, EACL, Trento 2001; A. Kilgarriff, G. Grefenstette, *Introduction to the special issue on the web as corpus*, in "Computational linguistics", XXIX, 2003, 3, pp. 333-47; A. Ludeling, S. Evert, M. Baroni, *Using web data for linguistic purposes*, in "Language and Computers", LIX, 2006, 1, pp. 7-24.

16. Un esempio attinente alla relazione tra dato empirico estratto da corpus e mediazione di una teoria linguistica per l'elaborazione di una voce lessicografica (entro il paradigma della teoria dei *frames*) si trova in S. Atkins, C. J. Fillmore, C. R. Johnson, *Lexicographic relevance: selecting information from corpus evidence*, in "International Journal of Lexicography", XVI, 2003, 3, pp. 251-80.

17. T. De Mauro, *Grande dizionario italiano dell'uso*, UTET, Torino 1999.

cessità di fornire al lettore una documentazione esplicita dei criteri, dei materiali, degli strumenti e dei dati applicati per l'elaborazione delle diverse componenti dell'opera lessicografica e del dettaglio delle sue voci. Nel caso della lessicografia basata su corpora tale necessità si fa ancora più stringente poiché la relazione con il "dato linguistico" si fa più diretta e rigorosa, sia nelle procedure *a priori* di definizione dell'architettura del corpus, di raccolta e di composizione quantitativa, di trattamento automatico, sia nelle procedure *a posteriori* che riguardano la selezione, l'ordinamento, l'identificazione dei pattern lessicali, semantici, combinatori e la loro descrizione.

2.2. Il fatto linguistico nella lessicografia collaborativa online

Il Web 2.0 ha cambiato le abitudini di scambio di informazione e soprattutto le modalità di partecipazione degli utenti a una moltitudine di attività comunicative. Tra queste sono emerse le tecnologie collaborative che spaziano dalla scrittura collaborativa alla costruzione di strumenti di riferimento come enciclopedie e dizionari. La lessicografia collaborativa si basa, come la forma enciclopedica sorella, sul principio del *croud sourcing*, ossia di attività partecipative online nelle quali utenti comuni sono chiamati volontariamente (e gratuitamente) alla realizzazione di compiti diversi (redazione di una voce di enciclopedia, revisione, controllo delle fonti, ecc.) il prodotto dei quali è sottoposto direttamente alla fruizione del pubblico, che a sua volta può intervenire modificando, aggiungendo, cancellando, precisando il testo redatto, seguendo procedure più o meno definite entro il progetto collaborativo. Esistono numerosi progetti collaborativi online di interesse linguistico[18], i più noti sono l'enciclopedia Wikipedia (www.wikipedia.org) e i dizionari online Wictionary (Wikizionario[19] in italiano, nato da una costola di Wikipedia, www.wictionary.org), Urban Dictionary (www.urbandictionary.com), OmegaWiki (http://www.omegawiki.org), Wordnik (http://www.wordnik.com) e WordReference (www.wordreference.com). Questo tipo di progetti sono fondati sul principio noto come *wisdom of the crouds*, la saggezza delle folle, ossia sulla competenza di utenti anonimi che condividono e controllano la qualità del lavoro prodotto.

Dal punto di vista del contenuto l'attenzione alla lessicografia collaborativa è soprattutto focalizzata sulla capacità di cogliere gli usi vivi, spesso anche gergali, occasionali, emergenti, tipici del parlato. Non a caso uno dei primi esempi è la creazione dell'Urban Dictionary, dizionario di *slang* in lingua inglese che ha raggiunto circa 6.800.000 definizioni e si propone come un dizionario aperto alternativo, rappresentativo della cultura popolare[20]. L'idea è dunque basata sul

18 Nella letteratura in lingua inglese ci si riferisce spesso a tali progetti come *collaboratively constructed lexical semantic resources* (CLSR).

19. Si userà il nome Wictionary per indicare il progetto multilingue, ossia la serie di dizionari online collaborativi e la loro architettura, mentre si userà Wikizionario per riferirsi specificatamente alle caratteristiche linguistiche della componente italiana.

20. Sulle caratteristiche specifiche che contraddistinguono Urban Dictionary rispetto ad

presupposto che l'utente ordinario volontariamente sottoponga proposte non solo di lemmi da includere ma anche la redazione stessa della voce. Tale utente è ritenuto, in quanto presumibilmente nativo (ma non è sempre vero), competente nell'identificazione di lemmi significativi, e soprattutto nelle procedure di elaborazione della voce stessa: selezione delle accezioni, elaborazione delle glosse, degli esempi d'uso, identificazione dei registri. Il modello di redazione, in nessuna delle opere collaborative di successo citate, si differenzia nella forma e negli elementi che caratterizzano le voci da un dizionario tradizionale, né nelle modalità di interrogazione.

La potenzialità di integrazione della lessicografia collaborativa con approcci *corpus-based* è sfruttata, ad esempio, da Wordnik, dizionario elettronico online nato nel 2009, che, in un'unica interfaccia di consultazione, fonde alcune risorse lessicografiche tradizionali in vecchie edizioni (come l'American Heritage Dictionary e la versione collaborativa che si basa sul Webster's Revised Unabridged Dictionary del 1913), tesauri tradizionali (Roget's II), dizionari collaborativi (Wiktionary), autorevoli risorse elettroniche recenti (Wordnet) con un corpus di miliardi di occorrenze dell'inglese che serve per allineare e individuare esempi d'uso, e l'integrazione in tempo reale con *tweet* che contengono la parola cercata[21].

Mentre l'attenzione per le risorse informative e enciclopediche è focalizzata sul problema dell'autorevolezza e della scientificità ed eticità nell'uso delle fonti, nel caso di Wictionary molto meno dibattito si è creato e meno si è discusso su aspetti di natura più propriamente lessicologica e linguistica. Oltre ai problemi relativi all'uso delle fonti e soprattutto ai rischi piuttosto forti di plagio da fonti tradizionali (tipico problema di Wikipedia)[22], esistono numerosi problemi peculiari dell'oggetto dizionario, rispetto all'enciclopedia, e risiedono nella particolare "saggezza" necessaria per la redazione di una voce di dizionari, significativamente diversa dalla competenza enciclopedica per molti aspetti.

Il vantaggio dei progetti collaborativi online è la loro crescita rapida dato il grande numero di utenti che partecipano all'arricchimento della risorsa e il continuo aggiornamento delle voci cui si accompagna una virtuale infinita espansibilità del lemmario di riferimento poiché si annullano i problemi di

altre opere elettroniche aperte si vedano: A. Peckham, *Urban Dictionary: distributed moderation of an online dictionary*, California Polytechnic State University, San Luis Obispo (CA) 2005; C. Cotter, Q. Mary, J. Damaso, *Online dictionaries as emergent archives of contemporary usage and collaborative codification*, in "Occasional Papers Advancing Linguistics", IX, 2007, pp. 1-11. Da notare che, anche nel caso dell'Urban Dictionary, si è arrivati dalla versione online a numerose edizioni cartacee.

21. «Wordnik is billions of words, over a billion example sentences, 7,385,584 unique words, 238,460 comments, 183,882 tags, 121,572 pronunciations, 107,004 favorites and 1,338,508 words in 36,654 lists created by 101,375 Wordniks», *Wordnik: Community*, http://www.wordnik.com/community, ultimo accesso: 4/11/2012

22. Sul plagio dei dizionari online e sulle conseguenze legali in alcuni casi qualche indicazione si trova in V. J. Docherty, *Dictionaries on the Internet: an Overview*, in *Proceedings of EURALEX 2000*, a cura di U. Heid *et al.*, IMS, Stuttgart 2000, pp. 67-74.

selezione dovuti a limiti di spazio[23] sul cartaceo o sull'elettronico statico (CD, DVD, dispositivi di archiviazione di massa portatili) e la possibilità di mantenere lo storico dei diversi interventi redazionali, molto utile per seguire i principi che governano il processo di elaborazione collettiva della singola voce.

2.3. Il caso del Wikizionario

Il Wikizionario è la versione italiana di Wiktionary, nato nel tardo 2002, sul modello della versione inglese, è un dizionario aperto al contributo dell'utente. La struttura e la presentazione delle voci ricalcano quelle di un dizionario tradizionale, senza l'aggiunta di alcun tipo di funzione supplementare.

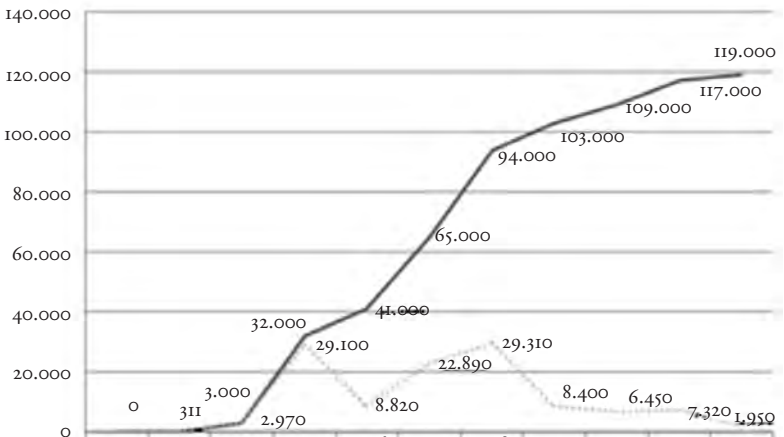
Prima di entrare nel confronto concreto relativo alle caratteristiche linguistiche del Wikizionario, comparato a modelli di dizionario tradizionale, anche informatizzato, è necessario avere un'idea delle caratteristiche che sono considerate le più innovative e vantaggiose dei dizionari collaborativi online: il tasso di accrescimento e la virtuale illimitatezza di spazio, la modificabilità e tracciabilità dell'elaborazione collaborativa, il suo essere *bottom-up* ossia prodotto dalla base degli utenti senza filtri d'autorità.

L'accrescimento della risorsa lessicografica Wikizionario deve essere valutato tenendo in considerazione diversi fattori. Se da una parte il numero di lemmi inseriti nella risorsa sono esponenzialmente aumentati dal 2002 al 2012 giungendo a circa 119.000, il tasso di crescita ha visto uno sviluppo molto significativo soprattutto negli anni 2007 e 2008 con un incremento di 22.890 e 29.310 lemmi rispettivamente. Il tasso di aggiunta una volta toccati circa 100.000 lemmi si assesta negli anni a seguire tra le 6.500 e le 8.500 voci aggiunte all'anno (cfr. figura 1). All'arricchimento del lemmario si accompagna con simile linea di tendenza anche il lavoro di estensione delle singole voci. Come tasso indicativo di crescita si può osservare la grandezza complessiva del Wikizionario inteso come corpus testuale giunge a quasi tre milioni di occorrenze le quali passano da uno a due milioni nei soli anni 2007-08[24].

23. Sulla ridefinizione del problema dello spazio dalla lessicografia su carta a quella elettronica si vedano M. H. Corréard, *Are space-saving strategies relevant in electronic dictionaries?*, in *Proceedings of the Tenth* EURALEX *International Congress*, a cura di A. Povlsen, C. Povlsen, Center for Sprogteknolog, Copenhagen 2002, pp. 12-7; R. Lew, *Space restrictions in paper and electronic dictionaries and their implications for the design of production dictionaries*, in *Issues in Modern Lexicography*, a cura di P. Bański, B. Wójtowicz, Lincom Europa, München 2011.

24. Nel passaggio tra il 2007 e il 2008 la dimensione globale (in numero di occorrenze) del Wikizionario è passata da 955.000 a 1.800.000 parole. La stessa dimensione aggiuntiva coperta nel 2008 (circa un milione di occorrenze) è stata raggiunta nei successivi quattro anni (2009-12) facendo raggiungere alla risorsa la dimensione di 2.900.000 occorrenze circa.

Figura 1
Tasso di accrescimento del Wikizionario (elaborazione su dati stats.wikimedia.org)

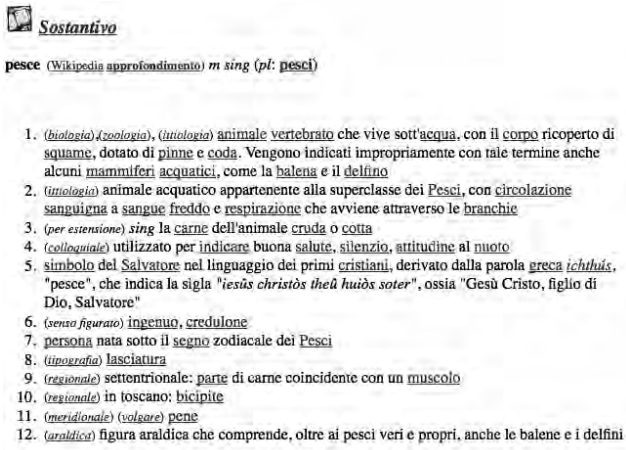


	2002 (dic)	2003	2004	2005	2006	2007	2008	2009	2010	2011	2002 (gen-ago)
Lemmi (ordine di grand.)	1	31	3.000	32.000	41.000	65.000	94.000	103.000	109.000	117.000	119.000
Voci aggiunte	0	0	2.970	29.100	8.820	22.890	29.310	8.400	6.450	7.320	1.950

Tuttavia, è necessario osservare come le indicazioni sull'estensione del lemmario fornite da Wikizionario siano in sostanza difficilmente comparabili con quelle di un dizionario tradizionale compiuto. Wikimedia computa infatti nel lemmario tutte le pagine che abbiano almeno un link interno (e che non siano reindirizzamenti), ossia anche pagine che risultano sostanzialmente non compilate (abbozzi *stub*), composte anche di una sola riga o del solo collegamento interno. Ad esempio se scrivessi la voce GATTO[25] e la compilassi con il solo testo "corpo agile"; se la parola *agile* ha un collegamento ipertestuale alla voce corrispondente, l'articolo sarebbe considerato nel computo perché considerato non vuoto. Non vi sono ulteriori vincoli di dimensione, completezza, numero di revisioni. Il criterio del link interno è attualmente adottato in tutti i progetti Wikimedia (ed è principalmente pensato per Wikipedia), dove il ruolo giocato dai link è tuttavia radicalmente diverso da quello giocato entro Wictionary. Mentre in Wikipedia i link interni sono guidati da un principio enciclopedico parsimonioso e tendenzialmente gerarchico, i link di Wictionary sono pervasivi e spesso semanticamente vuoti (ossia non rimandano a un principio ordinatore del percorso di senso costruito) e governati dalla sola comparsa della forma di citazione nel lemmario complessivo (cfr. figura 2, in cui più di una parola su due è un link interno al dizionario).

25. In questo contributo per riferirsi genericamente a un lessema si usa come di consueto il corsivo (*gatto*), per riferirsi alla voce lessicografica, al lemma in uno o più specifici dizionari, si usa il tondo maiuscoletto (il lemma GATTO). Tutte le altre convenzioni sono quelle solitamente usate nella letteratura linguistica.

Figura 2
La voce PESCE nel Wikizionario (accesso: 23 settembre 2012)



Sostantivo

pesce (Wikipedia approfondimento) *m sing* (*pl*: pesci)

1. (*biologia*),(*zoologia*), (*ittiologia*) animale vertebrato che vive sott'acqua, con il corpo ricoperto di squame, dotato di pinne e coda. Vengono indicati impropriamente con tale termine anche alcuni mammiferi acquatici, come la balena e il delfino
2. (*ittiologia*) animale acquatico appartenente alla superclasse dei Pesci, con circolazione sanguigna a sangue freddo e respirazione che avviene attraverso le branchie
3. (*per estensione*) *sing* la carne dell'animale cruda o cotta
4. (*colloquiale*) utilizzato per indicare buona salute, silenzio, attitudine al nuoto
5. simbolo del Salvatore nel linguaggio dei primi cristiani, derivato dalla parola greca *ichthùs*, "pesce", che indica la sigla "*iesûs christòs theû huiòs soter*", ossia "Gesù Cristo, figlio di Dio, Salvatore"
6. (*senso figurato*) ingenuo, credulone
7. persona nata sotto il segno zodiacale dei Pesci
8. (*tipografia*) lasciatura
9. (*regionale*) settentrionale: parte di carne coincidente con un muscolo
10. (*regionale*) in toscano: bicipite
11. (*meridionale*) (*volgare*) pene
12. (*araldica*) figura araldica che comprende, oltre ai pesci veri e propri, anche le balene e i delfini

Sebbene esistano diversi *Manuali di stile* per Wikipedia e Wictionary, la politica generale relativa all'inserimento di collegamenti interni è definita solamente nelle linee guida per l'enciclopedia e i criteri di diffusione dei dati relativi al lemmario non sono diversificati per le due risorse.

Dal punto di vista redazionale un vantaggio (anche per la metalessicografia) dei Wiktionary è la modalità di interazione collaborativa[26] che prevede interventi sulla pagina stessa (funzione *edit*) da parte di tutti gli utenti con l'obbligo di motivare gli interventi di modifica su singoli aspetti della voce e le pagine di discussione nelle quali questioni di interesse generale o aspetti specifici sono sottoposti a una valutazione del gruppo di utenti. Dunque non solo rimane traccia di ogni discussione aperta e delle soluzioni via via definite, ma vi è anche la possibilità di accedere allo storico delle operazioni di modifica e discussione che gli utenti svolgono su ciascuna singola voce. Alla cronologia di revisione è possibile anche associare una visualizzazione di confronto che mette in parallelo le diverse versioni.

Le caratteristiche dell'architettura collaborativa di Wikizionario (incluso lo storico delle revisioni) sono definite entro il modello Wiki generale che vale per tutti i progetti Wikimedia (Wikipedia, Wiktionary, Wikiquote, Wikibooks, Wikisource, Wikinews, Wikimedia Commons, MediaWiki). Dal punto di vista specificatamente lessicografico, al di là del valore stesso della risorsa intesa come prodotto, le pagine di revisione e discussione offrono una prospettiva del tutto nuova sul modo in cui l'utente vede l'oggetto dizionario e come lo

26. Cfr. I. Gurevych, E. Wolf, *Expert-built and collaboratively constructed lexical semantic resources*, in "Language and Linguistics Compass", IV, 2010, 11, pp. 1074-90; I. Gurevych, T. Zesch, *Collective intelligence and language resources: introduction to the special issue on collaboratively constructed language resources*, in "Language Resources and Evaluation", 2012, pp. 1-7.

intende. Emergono, infatti, nel dibattito e nelle pratiche di modifica innumerevoli spunti di interesse linguistico sugli usi del dizionario e anche, ancora più nettamente, sulla percezione che gli utenti hanno delle caratteristiche della lingua e del problemi di modello e normatività rappresentati dall'opera lessicografica.

La collaboratività, inoltre, garantendo un arricchimento della risorsa piuttosto rapido, ha spinto soprattutto chi si occupa di sviluppo di risorse linguistiche ad avvicinarsi ai progetti di tipo Wiki per estrarre automaticamente informazione linguistica da aggregare ad uso computazionale o da integrare e arricchire con risorse linguistiche *expert-based*[27]. Sono stati fatti alcuni tentativi di usare Wiktionary in questa direzione, ad esempio per la costruzione di un database mediante l'estrazione automatica dei dati[28].

Anche editori tradizionali come Merriam-Webster e Macmillan, oltre ad aver pubblicato parzialmente o *in toto* dizionari gratuiti online, accanto alle edizioni cartacee, hanno anche iniziato progetti di *dizionari aperti*, ossia dizionari collaborativi nei quali gli utenti possono proporre per l'inserimento lemmi e definizioni. In entrambi i casi raramente i materiali suggeriti hanno esibito i requisiti di uso e convenzionalità necessari a permettere a tali voci di raggiungere le edizioni cartacee.

Inoltre da più parti è stata sottolineata la cattiva qualità e accuratezza del materiale, specialmente nel caso in cui non sia chiaro come verificare il trattamento e le fonti[29], tale constatazione ha ridimensionato alcune delle iniziali prospettive di applicazione computazionali delle risorse collaborative. Bisogna inoltre sot-

27. Tra i numerosi esempi e sperimentazioni applicative si veda il lavoro di raccordo I. Gurevych, J. Kim, *The people's Web meets NLP: collaboratively constructed language resources*, Springer Verlag, Berlin, in corso di stampa.

28. Cfr. T. Zesch, C. Müller, I. Gurevych, *Extracting lexical semantic knowledge from Wikipedia and Wiktionary*, in *Proceedings of the Conference on Language Resources and Evaluation* (LREC) 2008, pp. 1646-52; A. Krizhanovsky, *Transformation of Wiktionary entry structure into tables and relations in a relational database schema*, in "preprint arXiv", 2010; A. A. Krizhanovsky, *The comparison of Wiktionary thesauri transformed into the machine-readable format*, in "preprint arXiv", 2010; F. Sajous, E. Navarro, B. Gaume, L. Prévot, Y. Chudy, *Semi-automatic enrichment of crowdsourced synonymy networks: the WISIGOTH system applied to Wiktionary*, in "Language Resources and Evaluation", LII, 2011, 1, pp. 1-34; J. McCrae, E. Montiel-Ponsoda, P. Cimiano, *Collaborative semantic editing of linked data lexica*, in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LRE'12)*, a cura di N. Calzolari K. Choukri, T. Declerck, M. Ugür Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, ELRA, ISTAMBUL 2012, pp. 23-5; G. Sérasset, *Dbnary: Wiktionary as a LMF based Multilingual RDF network*, in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC-2012)*, cit., pp. 2466-72.

29. Nel caso di Wikipedia esiste una pratica consolidata e delle linee guida esplicite riguardo all'uso delle fonti, alla obbligatorietà dell'esplicitazione, alle modalità di citazione e alle politiche di correzione dove questi requisiti non siano soddisfatti. Nonostante ciò, più volte è stata denunciata la mancanza di affidabilità del dato dovuta da una parte all'anonimato dei collaboratori dall'altro ai rischi di informazioni pilotate, vandalismo e superficialità nel riportare dati secondari. Non esiste per Wikizionario alcuna politica di gestione del contenuto né di revisione delle fonti. Peraltro i problemi di accuratezza e qualità dell'informazione linguistica è di natura empirica in un senso ben diverso rispetto all'informazione enciclopedica per cui linee guida in proposito finirebbero per coincidere con un manuale di lessicografia e avrebbero difficoltà ad essere impostati per un uso libero, volontario, di non professionisti.

tolineare come l'accrescimento dei progetti collaborativi stia lentamente diminuendo, come si può notare sia per Wikizionario nei dati riportati sopra, ma anche nel caso della popolarissima Wikipedia.

Dal punto di vista del trattamento strettamente linguistico Wiktionary si pone come progetto lessicografico multilingue da due diversi punti di vista: da un lato si è dotato di versioni in 170 lingue. Tali versioni tuttavia non sono a loro volta monolingui, ma contengono definizioni nella lingua di arrivo (lingua della versione) di parole prevalentemente della lingua stessa, ma anche appartenenti a lingue diverse, sempre glossate nella lingua di arrivo. Il Wiktionary nella versione inglese contiene definizioni in inglese di lemmi appartenenti a circa 450 lingue diverse.

La multilinguisticità di Wiktionary è dunque del tutto atipica nel panorama lessicografico tradizionale, trattandosi di opera mono-e bilingue (o meglio plurilingue monodirezionale).

Per quanto riguarda l'elaborazione e l'impalcatura generale dell'opera, rispetto ad altri progetti simili, come Urban Dictionary, Wiktionary prevede linee guida per la compilazione e non lascia l'utente totalmente sprovvisto di indicazioni normative[30].

Un esempio è costituito dalla formulazione delle norme di selezione dei lemmi da includere nel dizionario. Il criterio d'uso adottato è basato su un insieme di criteri molto generali basati sull'attestazione, per i quali tuttavia non è obbligo fornire fonti:

> – Clearly widespread use, or
> – Usage in a well-known work, or
> – Usage in permanently recorded media, conveying meaning, in at least three independent instances spanning at least a year, or
> – For terms in extinct languages: usage in at least one contemporaneous source[31].

Ma quali sono i vantaggi e gli svantaggi di spostare la responsabilità della compilazione di un dizionario da un piccolo gruppo di esperti a un larghissimo gruppo di utenti ordinari della lingua? L'individuazione e descrizione del fatto linguistico, così come presentato e ordinato in un dizionario, è operazione che per essere portata a termine necessita solo della condizione dell'essere parlanti nativi della lingua descritta? Vi è analogia tra competenze linguistiche produttive e ricettive non specialistiche, competenze metalinguistiche di base, ed eventualmente, competenze nella espressione di giudizi di grammaticalità e ciò che è richiesto invece in termini di elaborazione linguistica e concettuale, al lessicografo per la

30. Cfr. P. A. Fuertes-Olivera, *The function theory of lexicography and electronic dictionaries: Wiktionary as a prototype of collective free multiple-language Internet dictionary*, in "Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow. Linguistic Insights-Studies in Language and Communication", XC, 2009, pp. 99-134.

31. Cfr. *Wiktionary*, http://en.wikipedia.org/wiki/Wiktionary; ultimo accesso: 3 novembre 2012.

costruzione di uno strumento applicativo dagli usi molteplici, ben definito storicamente e culturalmente?

Tale questione può essere affrontata in modi molto diversi. Nel paragrafo successivo vedremo descrittivamente e comparativamente in che modo Wikizionario e i dizionari tradizionali rendono conto delle diverse componenti che costituiscono il lavoro e la competenza specifica del lessicografo.

3
Confronto tra dizionari tradizionali e dizionari elettronici collaborativi: il caso dell'italiano

In questo paragrafo si cercherà di mostrare quali siano i punti in comune e le principali divergenze, allo stato attuale, tra lessicografia redatta da esperti e lessicografia collaborativa per la lingua italiana. Per far ciò si sono presi in esame alcuni lavori che possano essere considerati come comparabili. Come esempi di lessicografia tradizionale, principalmente cartacea (ma dotata anche di versione informatizzata), sono stati presi ad esempio i dizionari monovolume *Dizionario della lingua italiana* (DMP, 2000)[32], curato da Tullio De Mauro e il *Sabatini Coletti Dizionario della lingua italiana* (SC, 2012)[33]. Tra le risorse elettroniche di tipo lessicografico si è scelto invece di includere Wikizionario. Si sono invece escluse le risorse in formato di rete semantica o tesauri, come i due *wordnet* italiani (ItalWordNet e Multiwordnet)[34], perché non comparabili – per obiettivi e funzioni, struttura generale delle voci e delle relazioni, utenti – alle opere lessicografiche citate[35].

3.1. Il lemmario e la copertura

Il primo aspetto di confronto riguarda la pura estensione delle opere lessicografiche, esibita principalmente dal numero di lemmi inseriti nell'opera e dalla sua

32. De Mauro, *Il dizionario della lingua italiana*, cit.

33. F. Sabatini, V. Coletti, *Il Sabatini Coletti. Dizionario della lingua italiana*, Sansoni, Milano 2012.

34. Cfr. L. Bentivogli, E. Pianta, C. Girardi, *Multiwordnet: developing an aligned multilingual database*, in *Proceedings of the First International Conference on Global WordNet*, Mysore, India, 2002, pp. 21-5; A. Roventini, A. Alonge, F. Bertagna, N. Calzolari, J. Cancila, C. Girardi, B. Magnini, R. Marinelli, M. Speranza, A. Zampolli, *Italwordnet: building a large semantic database for the automatic treatment of Italian*, in "Computational Linguistics in Pisa – Special Issue", XVIII-XIX, 2003, 2, pp. 745-91.

35. Altri lavori comparano per l'inglese, il tedesco e il russo dizionari tradizionali con Wictionary e Wordnet, cfr. Krizhanovsky, *The comparison of Wiktionary thesauri transformed into the machine-readable format*, cit.; C. M. Meyer, I. Gurevych, *Worth its weight in gold or yet another resource-a comparative study of Wiktionary, Openthesaurus and Germanet*, in *Computational Linguistics and Intelligent Text Processing*, a cura di A. Gelbukh, springer, Berlino 2010, pp. 38-49; C. M. Meyer, I. Gurevych, *How web communities analyze human language: word senses in Wiktionary*, in *Proceedings of the Web Sci10: Extending the Frontiers of Society On-Line*, a cura di N. C. Raleigh, 2010, pp. 349; C. M. Meyer, I. Gurevych, *Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography*, in *Electronic Lexicography*, a cura di S. Granger, M. Paquot, Oxford University Press, Oxford 2012.

dimensione globale espressa in numero di parole (occorrenze totali). In secondo luogo il lemmario va valutato qualitativamente, osservando quali lemmi sono inclusi, secondo quale criterio, e quale sia il grado di copertura del lessico offerto dal dizionario.

Il lemmario, se osservato puramente dal numero dei lemmi dichiarati per ciascuna opera, risulta pressoché analogo (circa 120.000 lemmi). La dimensione del lemmario attuale del Wikizionario tuttavia risulta meno accurata rispetto alle dichiarazioni dei dizionari cartacei poiché, come si è detto sopra, è considerata non provvisoria qualunque voce contenga un link interno, non tenendo conto della articolazione interna della voce, della sua completezza o di una valutazione sulla possibilità di dichiararla compiuta[36]. Per comprendere meglio il dato numerico si possono osservare da una parte la dimensione totale dell'opera, misurata in numero di parole, dall'altra la corrispondente lunghezza media di ciascuna voce.

L'estensione del lemmario di Wikizionario è inoltre non rappresentativa della reale risorsa lessicografica per l'italiano per due motivi: dei circa 118.000 lemmi indicati come attualmente presenti nel dizionario, solo un terzo è costituito da lemmi italiani (parole italiane trattate come in un dizionario monolingue tradizionale) ossia circa 40.000, i restanti due terzi sono invece rappresentati da lemmi in lingue straniere (arabo, cinese, russo, tedesco, inglese, ecc.) che contengono nella voce del Wikizionario solo un'indicazione (anche piuttosto sommaria) del traducente in italiano, o una brevissima glossa (ad esempio *soccer* italiano: calcio). D'altra parte una peculiarità di Wictionary è quella di voler costituire un dizionario multilingue in cui lo scopo è «descrivere tutti i lemmi di tutte le lingue»[37] in ciascuna delle lingue del progetto. Dunque se è vero che per Wikizionario l'interesse centrale è la copertura dei lemmi italiani, è anche vero che esiste una esigenza di descrizione (in italiano) anche di lemmi appartenenti a una delle 170 lingue del progetto. Del resto l'ambiguità dell'obiettivo è anche affermata dall'ulteriore progetto prototipo di dizionario multilingue *community driven*, Wikabulary, di cui si nota la differenza rispetto a Wictionary come «focus on translation to other language (and on definition in the other language) rather than on definition in the same language»[38].

Una seconda discrepanza nel confronto tra i lemmari dipende dalla scelta di Wikizionario di non separare in pagine diverse gli omografi (dunque *riso* come "pianta e seme commestibile" e come "il ridere" che si trovano censiti entro la

36. Mentre, nel caso di Wikipedia, alcune voci sono a tutti gli effetti considerate compiute e vengono "chiuse" dalla redazione, ossia risultano sostanzialmente non modificabili dagli utenti perché hanno raggiunto un livello di completezza e condivisione riconosciuto dai redattori, non esiste tale prassi per il Wikizionario, per cui tutte le voci risultano sempre aperte. Questa, se da un lato è da ritenersi una buona prassi poiché consente potenzialmente la segnalazione di usi innovativi o non previsti, dall'altra non consente una valutazione ragionevole del livello di elaborazione di ciascuna voce.

37. *Wikizionario: Aiuto: Aggiungere una sezione in lingua straniera*, http://it.wiktionary.org/wiki/Aiuto:Aggiungere_una_sezione_in_lingua_straniera, ultimo accesso: 11 ottobre 2012.

38. *Wikabulary*, http://meta.wikimedia.org/wiki/Wikabulary, ultimo accesso: 10 ottobre 2012.

medesima voce), mentre nei dizionari tradizionali le voci sono separate e segnalate da un preponente. Questa differenza cambia significativamente l'idea stessa che Wikizionario propone di lemma, che finisce per coincidere con la forma inserita da tastiera ed essere fortemente condizionata dalle esigenze di ricerca da tastiera dell'utente, tanto da includere anche alcune forme flesse. Il lemmario dichiarato da Wikizionario più che essere un lemmario in senso stretto è un indice delle forme di citazione dei lemmi (o ancora più precisamente un indice degli articoli/pagine).

Tabella 1
Lemmario, dimensione ed estensione delle voci di DMP e Wikizionario

Dizionario	Lemmi	Dimensione (in occorrenze)	Estensione media delle voci (in occorrenze)
DMP	129.432	5.244.679	40,52 parole
Wikizionario	118.441[a]	ca. 2.900.000[b]	24,48 parole

[a] Dato aggiornato a settembre 2012 (http://it.wiktionary.org). Il conteggio dei lemmi in Wictionary (e Wikizionario) include tutte le pagine che contengano almeno un link interno. Il dato reale è di fatto inferiore dato lo stato incompleto e di puro riferimento di molte pagine interne in fase di lavorazione.
[b] La dimensione di Wikizionario è aggiornata a febbraio 2012 e stima il numero di occorrenze testuali. Sono esclusi i reindirizzamenti, il codice HTML e Wiki e i link nascosti.

Considerati DMP e Wikizionario come corpora in cui siano considerati "testo" tutti gli elementi che costituiscono il contenuto delle singole voci (informazioni grammaticali, glosse, esempi, sillabazione, ecc.), escludendo l'eventuale architettura dell'impaginazione (nel caso di Wikizionario escludendo il codice HTML, le funzioni Wiki e i link nascosti) emerge una forte sproporzione quantitativa. Mentre il Wikizionario non raggiunge i tre milioni di occorrenze, il DMP supera i cinque milioni. Se il dato si riporta in termini, seppure solamente indicativi, di lunghezza media di ciascuna voce si può chiaramente intravedere lo statuto non compiuto delle voci del Wikizionario[39] poiché tale lunghezza è di circa 24 parole, contro le 40 del DMP (cfr. tabella 1).

Al dato grezzo relativo all'estensione del lemmario va associato un confronto relativo alla qualità dei lemmi inclusi nel lemmario e alla considerazione relativa alle fasce d'uso e al tipo di lemmi che ciascuna opera seleziona. Per fare ciò è tuttavia necessario uniformare i lemmari secondo uno stesso principio. Poiché, come si è detto, il Wikizionario accorpa gli omografi si è proceduto alla semplificazione del lemmario del DMP e del SC schiacciando gli omografi nella stessa voce. In questo modo i lemmi del DMP si riducono a 122.262. Dei rispettivi lemmari sono confrontate solo le voci monorematiche escludendo polirematiche, suffissi, numeri e sigle.

39. Nel computo sono incluse tutte le voci dichiarate come non abbozzate secondo il criterio formale della presenza di un link interno.

Tabella 2
Estensione dei lemmari confrontati (DMP, SC, Wikizionario)

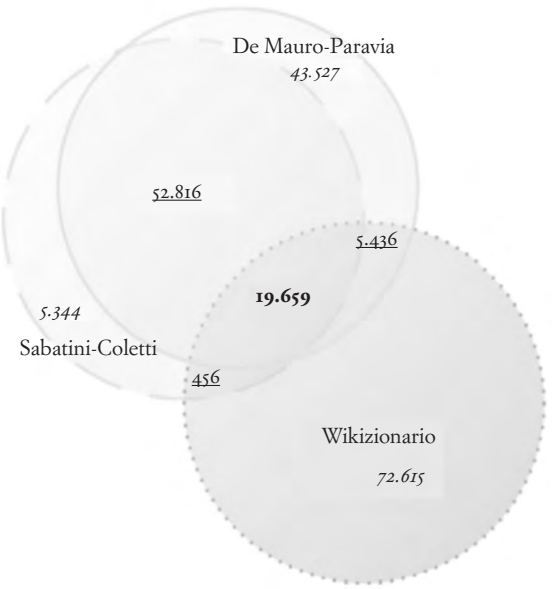
Dizionario	Lemmi monorematici
DMP (2000)	121.438
SC (2012)	77.819
Wikizionario (2012)	98.166[a]

[a] Il lemmario del Wikizionario, come si è detto, contiene anche un gran numero di lemmi appartenenti a lingue straniere, poiché non è stato possibile filtrarli a monte, il lemmario è stato trattato come i lemmari dei dizionari tradizionali estraendo solo le monorematiche, eliminando le sigle, numeri e i suffissi ed eliminando tutti i lemmi scritti in alfabeti diversi da quello latino.

Confrontando i lemmari uniformati (tabella 2) si ottiene un quadro molto sbilanciato relativo alle scelte di inclusione ed esclusione delle tre opere. Di un totale di 199.853 lemmi diversi che si ottengono cumulando i dati delle tre opere (ed eliminando i duplicati) solamente 19.659 sono i lemmi comuni a tutti. Potremmo pensare allora che la selezione della restante porzione di ciascuna opera risponda a criteri diversi per ciascuna opera e dunque che le relazioni tra i lemmi restanti tra loro sia egualmente casuale. Tuttavia osservando le intersezioni tra i lemmari attraverso un diagramma di Venn è possibile notare alcune caratteristiche peculiari (cfr. figura 4). Si osserva infatti che dei lemmi non comuni a tutti una grandissima porzione è invece solamente in comune a DMP e SC (altri 52.816 lemmi). In altre parole i lemmi completamente in comune tra DMP e SC coprono quasi il 60% del DMP e più del 93% del SC, con differenze dovute alla maggiore estensione del lemmario del DMP, che conta 43.527 lemmi esclusivi.

Se invece osserviamo la sovrapposizione tra DMP e Wikizionario la situazione cambia radicalmente poiché i lemmi totalmente comuni (comuni a tutti, e comuni solo a DMP e Wikizionario) sono 25.095, ossia solo il 20,6% del lemmario del DMP e 25,5% del lemmario del Wikizionario. Sembra dunque possibile osservare una maggiore omogeneità nelle scelte complessive di DMP e SC, e una effettiva peculiarissima scelta dei lemmi del Wikizionario. Il Wikizionario infatti possiede in comune con le altre fonti solo il 26% del suo lemmario, mentre il restante 74% (72.615 lemmi) è costituito da scelte di selezione esclusive che includono oltre a un gran numero di lemmi appartenenti a lingue straniere – si badi non forestierismi che si possono trovare in testi italiani, ma lessemi in uso esclusivamente in lingue straniere con glosse in italiano o traducenti come *aale* (tedesco), *barber* (inglese), *comète* (francese), *øy* (norvegese) – forme dialettali come *barbuttiari* (siciliano) o *colôr* (friulano), e anche lessemi tecnico-specialistici come *barbitta*, *turoniano* o *smarino*, letterari, rari o di basso uso come *abbalconato,* ma anche a lemma si trovano alcune forme femminili (*neutra*, *incestuosa*, *mattutina*) o plurali (*barboncini*, *coltelli*) senza sovrappiù semantico e perfino forme coniugate di verbi (*prendo*) e forme molto innovative con attestazioni di dubbia significatività (*turnicare* da uno spettacolo di Aldo, Giovanni e Giacomo).

Figura 3
Relazioni tra i lemmari di DMP, SC e Wikizionario (in corsivo i lemmi esclusivi di ciascuna fonte, in sottolineato i lemmi comuni a due fonti, in grassetto i lemmi comuni a tutte le fonti)



Ricerche precedenti hanno esaminato e confrontato i lemmari del Wiktionary inglese e di Wordnet, trovando dati analoghi a quelli osservati qui per l'italiano, ossia una sovrapposizione che riguarda solamente il 23% del lemmario di Wiktionary[40]. Per il tedesco invece la sovrapposizione tra Wiktionary tedesco e Germanet raggiunge il 51,7% dei lemmi di Wiktionary e per il russo la sovrapposizione tra la versione russa di Wiktionary e RussianWordNet è di circa il 15,8%[41]. Bisogna tuttavia sottolineare che la comparazione non pare del tutto legittima poiché Wiktionary ha una logica fondata sui singoli lessemi, mentre Wordnet, Germanet e RussianWordNet hanno una logica basata su insiemi sinonimici. Stupisce molto più notare che, per l'italiano, comparando Wikizionario a dizionari che presentano lemmari di dimensione paragonabile, la sovrapposizione riguardi al massimo il 26% del lemmario della risorsa collaborativa.

40. Cfr. Meyer, Gurevych, *How web communities analyze human language*, cit. Da notare inoltre come il numero complessivo dei lessemi coperti da Wordnet fosse circa la metà rispetto a Wiktionary.

41. Va notato, tuttavia, che i dati sul tedesco e sul russo riguardano i lessemi, disambiguando l'omografia relativa delle forme di citazione. Cfr. Meyer, Gurevych, *Wiktionary: a new rival for expert-built lexicons?*, cit. I dati sull'italiano e sull'inglese riguardano invece le pure forme di citazione (senza abbinamento di classe grammaticale). È presumibile ritenere che il dato sulle forme di citazione non disambiguate risulti più alto rispetto a quello che si ottiene aggiungendo il discrimine della classe grammaticale.

3.2. La copertura del vocabolario di base e i bisogni dell'utente

Per interpretare meglio il dato è necessario tuttavia focalizzare l'attenzione sulle diverse fasce del lessico coperte da ciascuna delle fonti. In particolar modo è necessario chiedersi se: *a*) le scelte di inclusione nelle tre fonti siano al loro interno omogenee, ossia coprano almeno la fascia che è considerata il cuore del lessico della lingua italiana, ossia il vocabolario di base (VDB)[42] o comunque un nucleo di lessemi considerati basici per un locutore italiano; *b*) le scelte individuali relative all'inclusione nel lemmario di ciascuna fonte siano basate su qualche criterio empirico o intuitivo oppure siano casuali.

La questione relativa alla copertura del lemmario di base può essere affrontata in modi diversi. L'uso del VDB (circa 7.000 lessemi appartenenti a tre fasce: fondamentale, alto uso e alta disponibilità) come esclusivo parametro di basicità risulta, per certi versi, scelta di parte poiché si tratta di una lista elaborata dall'autore di uno dei dizionari messi a confronto, anche se la "basicità" del VDB è stata riconosciuta come più tipicamente caratterizzante del nucleo del lessico italiano nel suo complesso anche dal punto di vista diacronico[43]. Per evitare di interpretare il dato in maniera parziale si è fatto ricorso anche ad altre due liste di riferimento: la lista delle circa 9.000 parole classificate come di "alta disponibilità" dal SC (SC_AD) e una lista di circa 5.000 lessemi con maggiore uso estratti da un corpus bilanciato di 20 milioni di occorrenze del Nuovo Vocabolario di Base (NVDB)[44]. Le tre liste mostrano aspetti diversi di basicità: il VDB è composto da tre fasce (fondamentale e alta frequenza, basate su un criterio di uso; alta disponibilità, basata su sperimentazione su lessemi che occorrono raramente nei testi ma sono molto conosciuti dai parlanti), i lessemi di "alta disponibilità" del Sabatini-Coletti sono basati su una intuizione di basicità[45]; mentre la lista dei lessemi del NVDB è una pura lista d'uso (basata sull'occorrenza dispersa in un corpus).

Se si osserva la tabella 3 ancora una volta le due opere tradizionali si distinguono da Wikizionario per la capacità di copertura dei lemmari di base. Mentre il DMP registra complessivamente la copertura maggiore di tutte e tre le liste, il Wikizionario registra coperture significativamente più basse: copre il 78,2% del

42. Cfr. T. De Mauro, *Guida all'uso delle parole: parlare e scrivere semplice e preciso per capire e farsi capire*, Editori Riuniti, Roma 1980.

43. Cfr. A. Giuliani, C. Iacobini, A. M. Thornton, *La nozione di vocabolario di base alla luce della stratificazione diacronica del lessico dell'italiano*, in *Parole e numeri. Analisi quantitative dei fatti di lingua*, a cura di T. De Mauro, I. Chiari, Aracne, Roma 2005, pp. 193-213.

44. Per una prima presentazione cfr. I. Chiari, T. De Mauro, *The new basic vocabulary of Italian: problems and methods*, in "Rivista di statistica applicata – Italian Journal of Applied Statistics", XXII, 2012, 2, pp. 21-35.

45. «Uno scopo eminentemente pratico: fornire a particolari categorie di utenti della lingua – tutti coloro che comunicano ampiamente con il pubblico: giornalisti, redattori di testi in genere – una generica indicazione sui *vocaboli che si presumono conosciuti e ben compresi da un parlante odierno di cultura media*, per suggerire di preferirli, in determinate circostanze, ad altri meno noti, oppure di dotare questi ultimi di spiegazione» F. Sabatini, V. Coletti, *DISC : dizionario italiano Sabatini Coletti*, Giunti, Firenze 1997, p. XIII.

VDB, il 72,2% del lessico di "alta disponibilità" individuato da Sabatini-Coletti, e raggiunge l'85,3% del NVDB, distanziandosi tuttavia per difetto di ben dieci punti dalle altre due opere.

Tabella 3
Copertura dei lemmari di base nei dizionari

	VDB	% cop.	SC_AD	% cop.	NVDB	% cop.
TDM	6.575	*100,0*	8.921	*99,9*	4886	*98,7*
SC	6.359	*96,7*	8.929	*100,0*	4766	*96,3*
Wikizionario	5.143	*78,2*	6.451	*72,2*	4224	*85,3*

Mentre le differenze che si registrano tra TDM e SC in merito al vocabolario di base (circa 3%) sono principalmente dovute a un diverso criterio di lemmatizzazione (SC non registra a lemma, ad esempio, moltissimi femminili – come *avvocatessa*, *anziana*, *bambina*, *cassiera* – e derivati avverbiali – come *attualmente*, varianti grafiche, che nel VDB e nel DMP figurano nel lemmario, le differenze di copertura di Wikizionario sembrano dovute non a scelte formali, ma a veri e propri vuoti di trattamento (mancano ad esempio i lemmi *manicomio*, *brigadiere*, *bovino*, *palazzina*, *campanello*, *seccare*, *svago*, *telefilm*, tra questi ben 429 lessemi del vocabolario fondamentale).

Una simile analisi condotta sulla copertura di lessici di base dell'inglese (copertura di Wiktionary dei lemmi della lista di Swadesh, delle parole frequenti della BNC e del Basic English), del tedesco (copertura del Wiktionary tedesco dei lemmi della lista di Swadesh e di Gut1 Wortschatz 100 e 500) e del russo (copertura del Wiktionary russo dei lemmi della lista di Swadesh e di Schteinfeldt) registra differenze significative: per l'inglese ci si colloca sempre sopra il 99% di copertura, per il tedesco intorno a 98-99%, per il russo intorno al 97%[46]. L'italiano si colloca invece, come abbiamo visto, tra il 72 e il 78%.

Ci si può domandare tuttavia se effettivamente il vocabolario fondamentale e anche il vocabolario di base nel suo complesso non siano poi di fatto elementi poco cercati in un dizionario online proprio perché contengono parole molto conosciute. In questo caso la copertura non sistematica di Wikizionario potrebbe essere giustificata da un'analisi più attenta dell'uso che della risorsa si fa, osservato mediante il ricorso al registro delle ricerche del sito.

Le ricerche sinora svolte sull'uso da parte degli utenti delle diverse fasce del vocabolario (seppure tutte relative a lingue diverse dall'italiano) forniscono un quadro molto più problematico. In particolare è stato osservato come la frequenza sia un criterio rilevante per le ricerche solamente nella fascia altissima, ossia la fascia delle prime migliaia di parole ordinate per frequenza[47]. Uno studio

46. Cfr. Meyer, Gurevych, *Wiktionary: a new rival for expert-built lexicons?*, cit.
47. Cfr. Dati abbastanza concordi in questa direzione sono presentati su inglese, francese, tede-

sull'uso della *Base lexical du français* (BLF)[48] indica una chiara preferenza nelle ricerche per le parole di altissima frequenza nei corpora (perfino per le parole vuote). Le ricerche più frequenti che vengono fatte online riguardano le parole che afferiscono a poco più del vocabolario fondamentale. Se si osservano invece nel loro complesso tutte le ricerche che vengono fatte in un dizionario online si scopre che, se si esclude appunto la fascia altissima, vi è una debolissima relazione tra parole cercate e frequenza in un corpus, per cui è impredicibile il tipo di lemma che viene sottoposto a ricerca sulla base di evidenze empiriche d'uso testuale. Una ricerca sull'uso del Danske Netordbog[49] (composto da circa 108.000 lemmi e dunque paragonabile ai monovolume analizzati nel presente contributo), dizionario monolingue online, rivela che circa il 20% dei lemmi ricercati non sono presenti nell'opera[50]. Bisogna notare che solo una parte, anche se alta in termini di frequenza, è costituita da ricerche di lessemi in forme ortografiche scorrette.

Non sono disponibili dati recenti e sistematici relativi alle parole cercate nei dizionari italiani online. Esiste tuttavia una lista di parole più cercate nel Wikizionario raccolte sulla base di 31 giorni di accessi per ricerca nell'anno 2008-09[51]. Sulla base di tale dato è possibile verificare i dati menzionati sopra per la situazione della lingua italiana.

Di 1.387 sequenze ricercate nel Wikizionario 173 sono sequenze di parole grafiche: di queste ultime la maggioranza sono polirematiche (*capo di abbigliamento*, *cabina telefonica*, *microscopio elettronico*, *indennità di disoccupazione*, *chiodo di garofano*, *calzoncini da bagno*, *topo di biblioteca*), mentre pochi sono modi di dire (*non avere peli sulla lingua*, *andare ad ingrassare i cavoli*, *nessuno può piacere a tutti*, *la bellezza sta negli occhi di chi guarda*), nomi propri come *Basso Sassone*, *Repubblica Federale di Germania*, *Nuova Zelanda*), frasi (*ti amo*). Nelle sequenze monorematiche vi sono inoltre 95 nomi propri. Rimangono 1.119 lessemi monorematici da sottoporre ad analisi. Poiché il dato non contiene informazione sulla frequenza con la quale i lessemi della lista sono cercati non si possono confrontare i ranghi con una lista di frequenza, si potrà invece verificare quanti dei lessemi cercati appartengono alle fasce del VDB.

sco, sesotho sa leboa (lingua sudafricana), danese, norvegese, in M. S. Johnsen, *Logfiler som leksikografisk analyseinstrument og hjælpeværktøj*, ASB. Institut for Sprog og Erhvervskommunikation, Aarhus Universitet, 2005; G. M. De Schryver, P. Joffe, S. Hillewaert, *Do dictionary users really look up frequent words? On the overestimation of the value of corpus-based lexicography*, in "Lexikos", XVI, 2006, pp. 67-83; Verlinde, Binon, *Monitoring dictionary use in the electronic age*, cit. Risulta esserci correlazione tra rango di ricerca nel dizionario e rango in un corpus di riferimento solo per le parole fino a rango di frequenza 5.000 per l'inglese e fino a rango 3.000 per il sesotho sa leboa nello studio di De Schryver, Joffe e Hillewaert. Bisogna tuttavia dire che i dati estratti da corpora sono stati confrontati nella versione non lemmatizzata, scelta che potrebbe aver disperso alcuni dati correlabili e significativi dal punto di vista linguistico.

48. Cfr. Verlinde, Binon, *Monitoring dictionary use in the electronic age*, cit.

49. Cfr. *Den Danske Netordbog (Danish Internet Dictionary)*, http://www.ordbogen.com/, ultimo accesso: 31 ottobre 2012.

50. Bergenholtz, Johnsen, *Log files as a tool for improving Internet dictionaries*, cit., p. 125.

51. Cfr. *Most searched terms [hits per day] (Wiktionary/it/)*, http://wikistics.falsikon.de/latest/wiktionary/it/searchTerms.htm, ultimo accesso: 3 novembre 2012.

Tabella 4
Fasce di frequenza delle parole più cercate in Wikizionario (2008-09)

Fascia d'uso	Lessemi cercati	%
Vocabolario fondamentale (FO)	315	28,2
Vocabolario di alto uso (AU)	183	16,4
Vocabolario di alta disponibilità (AD)	101	9,0
Non VDB	520	46,5

Come si può vedere in tabella 4, anche nel caso dell'italiano più della metà delle parole ricercate appartengono al vocabolario di base (53,5% delle parole cercate). Entro il vocabolario di base le fasce più significativamente presenti, prevedibilmente, sono quella del fondamentale e alto uso raggiungendo circa il 45% delle parole cercate. Per quanto riguarda il vocabolario fondamentale risultano cercate anche le preposizioni e congiunzioni (*ma*, *a*, *di*). I dati confermano sostanzialmente i dati forniti per le altre lingue sulla relazione tra frequenza d'uso e frequenza di ricerca nei dizionari elettronici.

Osservando inoltre più da vicino il 46,5% di parole cercate non appartenenti al vocabolario di base troviamo i lessemi più vari (*litografia*, *ibrido*, *guardrail*, *giocondo*, *emancipare*, *affisso*, *refurtiva*, *dissidente*, *switch*), ma ancora in linea con i dati ottenuti per i dati di registro di dizionari elettronici di altre lingue si trovano tra le parole più cercate espressioni volgari e riferimenti sessuali (23 lessemi su 1.119)[52].

La prima conseguenza dell'analisi sulla relazione tra frequenza e ricerca delle parole nei dizionari è che una risorsa online come Wikizionario, con ancora più evidenza, ha la necessità di coprire il vocabolario fondamentale per esaudire le richieste dei suoi fruitori, mentre può non tener conto dell'uso nei corpora per la predisposizione dei criteri di selezione del lemmario nel suo complesso. Questa osservazione contrasta con i dati di copertura attualmente registrati nella versione italiana online e che vedono il vocabolario di base e le altre liste "basiche" non completamente coperte dal dizionario collaborativo. La seconda conseguenza più interessante dal punto di vista linguistico riguarda l'uso delle liste di frequenza estratte da corpora come criterio base per l'identificazione dei lemmi da selezionare per un dizionario (monolingue o bilingue). Da questo punto di vista i dati attualmente disponibili sulla correlazione tra funzioni di ricerca su dizionari e rango nelle liste di frequenza fa pensare che, almeno per la costruzione del lemmario, una metodologia *corpus-based*, per quanto empiricamente ben fondata, non sia la panacea di ogni bisogno linguistico, anche se risulta certamente *un* criterio esplicitamente proponibile (e certificabile) contro l'assenza di criteri o linee guida.

52. Sono presenti, tra le parole cercate, sia espressioni tipicamente usate come insulto (*stronzo*, *porco*, *culattone*, *cagna*) o a funzione pragmatica (*cazzo*, *minchia*), organi sessuali usati anche in senso metaforico (*fica*, *figa*, *culo*, *capezzolo*) le principali descrizioni di orientamento sessuale (*bisessuale*, *omosessuale*, *eterosessuale*).

3.3. Le glosse e l'articolazione in accezioni

Il confronto che l'opera lessicografica tiene con il dato linguistico empirico però non si esaurisce nella costituzione e selezione del lemmario, ma tipicamente dipende anche dall'elaborazione delle glosse e dalla loro articolazione in accezioni. Poiché esula dagli obiettivi di questo contributo un'analisi puntuale dei diversi modi di articolare le accezioni in opere lessicografiche diverse e l'individuazione dei principi di metalessicografia che determinano di volta in volta l'accorpare o lo scorporare famiglie di sensi e contesti d'uso, si cercherà di dare un'idea complessiva da una parte relativa al valore "empirico" della glossa, dall'altra dalla rete che la glossa intreccia con gli altri elementi della voce lessicografica.

L'articolazione delle accezioni lessicografiche è, in certo senso, il più gravoso e complesso degli oneri del lessicografo, poiché richiede un'organizzazione del dato linguistico empirico o della competenza del parlante non sempre esplicitabile in criteri e linee guida condivise:

> I dizionari sono solitamente parchi di chiarimenti sui criteri adottati per distinguere e giustificare le diverse accezioni di una stessa parola. I criteri sono spesso impliciti e sono diversi. Ciò del resto in parte è necessario per consentire al dizionario di dar conto della sinuosità e variabilità dell'uso effettivo, in parte è il riflesso del carattere a lungo poco sistematico e prevalentemente artigianale e intuitivo del lavoro lessicografico anche più accurato[53].

D'altra parte, non solamente l'uso è sinuoso e variabile, ma gli esempi concreti d'uso, i dati eventualmente estratti da corpora di riferimento, sono una massa multiforme il cui ordinamento è operazione di astrazione su un terreno continuo e *fuzzy*[54] e in cui molti casi risultano del tutto ambigui all'assegnazione[55].

Nel Wikizionario, data la sua natura che non prevede la supervisione di esperti, non sono affrontati nell'elaborazione delle glosse i problemi tradizionali di coerenza e accuratezza globale, come ad esempio, l'uso di modelli descrittivi diversi relativi al contenuto (definizioni per condizioni necessarie e sufficienti, definizioni fenomenologiche, definizioni prototipiche, ecc.) e alla

53. T. De Mauro, *Basi di conoscenze e banche dati lessicali*, in *XXI secolo. Comunicare e rappresentare*, Istituto della Enciclopedia Italiana, Roma 2009, pp. 253-308.

54. Cfr. G. Lakoff, *Lexicography and generative grammar 1: hedges and meaning criteria*, in "Annals of the New York Academy of Sciences", CCXI, 1973, 1, pp. 144-53.

55. Il compito di associare accezioni di un dizionario a contesti d'uso è compito complesso e per certi versi metalinguisticamente innaturale (per questo si veda la sperimentazione condotta sull'associazione di accezioni lessicografiche a classi ontologiche in I. Chiari, A. Oltramari, G. Vetere, *Di cosa parliamo quando parliamo fondamentale?*, in *Lessico e Lessicologia. Atti del Convegno della Società di Linguistica Italiana (Viterbo 27-29 settembre 2010)*, a cura di S. Ferreri, Bulzoni, Roma 2012, pp. 185-202. Ancor più arduo e spesso indecidibile è il compito di ricondurre esempi concreti a specifiche accezioni d'uso di un dizionario e questo non perché il dizionario sia mal articolato, ma perché è presumibile che esistano circostanze in cui si applica un insieme di famiglie di sensi collettivamente in un contesto d'uso e non obbligatoriamente una sola sua articolazione astratta.

forma (glosse autonome e costituire da frasi complete o glosse sintetiche per parole chiave, l'evitamento della circolarità delle glosse, l'individuazione di esempi d'uso adeguati e formati in modo non definitorio[56]. In generale l'architettura poco supervisionata non permette la verifica della coerenza delle scelte e della loro applicazione, risultando in voci difformi tra loro e a volte circolari.

L'ordinamento delle accezioni ha una grandissima importanza nel disegno generale di una voce lessicografica. Studi recenti sulle strategie di consultazione delle voci di dizionario sottolineano ad esempio quanto sia determinante la prima accezione (e l'ultima) e il meccanismo che permette la navigazione nella voce per consentire l'identificazione dell'accezione appropriata[57]. La navigazione può essere guidata in vari modi, oltre alla lista bruta delle accezioni numerate, esistono oggi diversi sistemi che servono strategicamente a ottenere migliori risultati nella selezione del senso ricercato: menu della voce, segnaposti, marche d'uso. I primi due, menu e segnaposti, sostanzialmente assenti nella lessicografia tradizionale, sono presenti oggi in alcuni dizionari monolingui per apprendenti dell'inglese[58]. Le marche d'uso invece, oggi piuttosto comuni nella lessicografia italiana e internazionale, sono da considerarsi un indiretto indicatore della centralità di una accezione nel complesso di un lemma, fornendo dunque una guida e un orientamento di tipo metalinguistico alla fruizione della voce (di questo di parlerà più avanti vedi § 3.3). Nella letteratura sul tema, per quanto ancora circoscritta e sperimentata solo sulla lessicografia anglosassone, si sottolinea come l'effetto della prima accezione sia determinante, trattandosi spesso dell'unica di cui l'utente fruisce nella sua ricerca. L'importanza dei sistemi di orientamento si dimostra specialmente ma non esclusivamente nel caso di voci molto lunghe e soprattutto nel caso di fruitori con livelli più bassi di competenza linguistica rispetto a gruppi più esperti per i quali la presenza o assenza di elementi di guida risulta non significativa per l'accuratezza e il tempo di identificazione dell'accezione ricercata.

Per quanto riguarda invece più direttamente l'ordinamento in genere nella lessicografia tradizionale si osserva un criterio di tipo cronologico che mette al

56. Per una panoramica sintetica ed esauriente dei problemi di metalessicografia più significativi si veda B. T. S. Atkins, M. Rundell, *The Oxford guide to practical lexicography*, Oxford University Press, Oxford 2008.

57. Cfr. R. Lew, J. Pajkowska, *The effect of signposts on access speed and lookup task success in long and short entries*, in "Horizontes de Lingüística Aplicada", VI, 2007, 2, pp. 235-52; R. Lew, *Users take shortcuts: navigating dictionary entries*, in *Proceedings of the* XIV *Euralex International Congress*, a cura di A. Dykstra, Ljouwert, Afûk, 2010, pp. 1121-32; H. Nesi, K. H. Tan, *The effect of menus and signposting on the speed and accuracy of sense selection*, in "International Journal of Lexicography", XXIV, 2011, 1, pp. 79-96.

58. Il *Longman Dictionary of Contemporary English* (LDOCE), terza edizione 1995, e il *Cambridge International Dictionary of English* (CIDE), 1995, usano sistemi di segnaposti costituiti da una parola o una frase identificativa del senso messa in evidenza per le accezioni lunghe. Ad esempio nel CIDE il lemma CRANE ha tre parole-guida o segnaposti: *machine*, *bird*, *stretch* che contrassegnano i principali sensi. Mentre il sistema dei menu è adoperato dal *Macmillan English Dictionary for Advanced Learners* (MED2), seconda edizione 2007.

primo posto la prima accezione attestata ove possibile determinarla, nella lessicografia spiccatamente basata su corpora (di cui tuttavia non abbiamo ancora esempi per l'italiano) l'ordinamento è frutto di una analisi dell'uso delle accezioni e mette al primo posto l'uso più frequente. Nella lessicografia britannica l'esempio più caratteristico di ordinamento cronologico è l'Oxford English Dictionary (OED), mentre il primo dizionario che ordina le accezioni secondo la frequenza in un corpus di riferimento è considerato il Collins Cobuild English Language Dictionary.

Anche questo aspetto dipende dall'approccio generale dell'opera, dai suoi obiettivi e dalla disponibilità di informazioni etimologiche, cronologiche e di frequenza. In entrambi i casi si tratta comunque di dati empirici desumibili da fonti esterne di natura testuale. Non sempre tuttavia il criterio di ordinamento è reso esplicito anche se nel lavoro di redazione non può non esser fornita indicazione in proposito trattandosi di un elemento tutt'altro che marginale della voce.

Con ordinamento secondo la frequenza è ad esempio, per la lingua inglese, l'impianto generale di ordinamento delle accezioni del *Collins Cobuild Dictionary for Advanced Users*. Nel caso del GRADIT e del monovolume DMP il criterio, reso esplicito nella documentazione di accompagnamento, è reso in maniera integrata: «le accezioni sono state ordinate, dove ciò non fosse troppo in contrasto con il loro uso, secondo un criterio cronologico, a partire da quella più anticamente attestata»[59]. In alcuni casi «il criterio della successione cronologica delle accezioni è stato abbandonato a favore di un ordinamento a grappolo che privilegia ai primi posti le accezioni avvertite come più importanti e frequenti nell'uso»[60]. Le accezioni tendono inoltre ad essere raggruppate per classi grammaticali.

Nel caso del Wikizionario invece, nonostante la presenza di un documento di linee guida per la redazione delle voci[61] che riguarda anche le caratteristiche delle definizioni, non vi è traccia di alcuna indicazione relativa all'ordine delle accezioni.

A titolo esemplificativo si prenda il lemma *portale* (s.m.) in Wikizionario, nel DMP e nelle due opere citate a riferimento della voce di Wikizionario (il Sabatini-Coletti e il Vocabolario Treccani online). Tutti e quattro i dizionari articolano la voce in tre accezioni. Tuttavia l'organizzazione semantica interna e l'ordinamento risultano diversi.

I tre dizionari tradizionali articolato le accezioni nello stesso modo. L'ordinamento non segue un criterio cronologico (poiché l'accezione più antica è la seconda), ma secondo un criterio legato all'uso più corrente e comune. Infatti il DMP riporta la marca d'uso CO (comune) per la prima accezione. Seguono le due accezioni tecnico-specialistiche. La voce di Wikizionario invece mette al primo posto l'accezione informatica, al secondo posto quella generale di "porta prin-

59. T. De Mauro, *La fabbrica delle parole: il lessico e problemi di lessicologia*, UTET, Torino 2005, pp. 81-2.

60. *Ibid*.

61. *Wikizionario: Manuale di stile*, http://it.wiktionary.org/wiki/Wikizionario:Manuale_di_stile, ultimo accesso: 6 ottobre 2012.

cipale di una chiesa…", evitando qualunque riferimento all'accezione di tecnica costruttiva. Riporta invece una accezione non menzionata in nessuna delle altre opere che è quella legata all'uso nella fantascienza.

Tabella 5
La voce *portale* in quattro dizionari

Wikizionario (2012)	DMP (2000)
1. (*informatica*) sito web che costituisce un punto di partenza, una porta di ingresso ad un gruppo consistente di risorse e criteri di navigazione di Internet o di una Intranet 2. (*architettura*) porta principale monumentale di un palazzo o di una chiesa, spesso ornata di sculture, bassorilievi e fregi 3. nella narrativa dei generi fantascienza e fantasy, dispositivo immaginario che funge da via di passaggio fra due siti fra loro distanti nello spazio e/o nel tempo	1. CO Porta monumentale di chiese e palazzi, sovente artisticamente decorata 2. TS tecn. Struttura statica ottenuta collegando rigidamente due piedritti verticali o inclinati con una trave orizzontale rettilinea o ricurva 3. TS inform. Sito di grandi dimensioni in cui gli utenti abbonati possono trovare la maggior parte dei servizi senza navigare in rete
SC (2011)	**Vocabolario Treccani (2012)**
1. Porta principale di una chiesa o di un palazzo, generalmente di forme monumentali 2. mecc. Struttura statica formata da una trave orizzontale che collega due piedritti verticali 3. inform. Sito internet che indirizza l'utente verso il reperimento di informazioni e servizi all'interno del sito stesso o in generale sul Web	1. Porta esterna d'ingresso a un edificio, artisticamente decorata e di grandi dimensioni; il termine si adopera con riferimento a edifici monumentali (ma in questo caso è più com. *portone*) e soprattutto a chiese: *un p. gotico*, *romanico*; *il p. del duomo di Pisa*; *il p. ligneo della chiesa di S. Sabina in Roma*. 2. Struttura a telaio costituita da due piedritti, ad asse verticale o anche inclinato, e da una traversa superiore ad asse rettilineo o curvo, solidale ai piedritti, dotata eventualmente di sbalzi laterali: *p. di acciaio*, *di cemento armato*; in partic., *gru a p.*, gru del tipo mobile, inferiormente munita di una struttura a portale doppio. 3. Per calco dell'ingl. *portal*, pagina iniziale di un sito internet che mette a disposizione dell'utente informazioni e servizî del sito stesso oppure collegamenti ad altri siti, che rinviano ad altrettanti servizi.

Se raffrontiamo l'articolazione della voce e la confrontiamo con il corpus di riferimento del NVDB[62], organizzato in sottocorpora bilanciati (stampa, narrativa, saggistica, spettacolo, comunicazione mediata dal computer, parlato) possiamo avere un quadro empirico sulla attestazione del termine nelle diverse accezioni.

62. Cfr. Chiari, De Mauro, *The new basic vocabulary of Italian*, cit.

Tabella 6
Il lemma *portale* (N) nel corpus del NVDB (dati normalizzati)

	CAT	Uso	Freq. tot.	Stampa	Narrat.	Sagg.	Spett.	CMC	Parlato
portale	N	204	553	39	2	466	9	32	4

Le attestazioni nel corpus (cfr. tabella 6) si polarizzano nei sottocorpora di saggistica, stampa e comunicazione mediata dal computer, ma in realtà distribuiscono le accezioni secondo due principali linee di tendenza. L'accezione che concentra il maggior numero di occorrenze è l'accezione informatica con l'87% delle occorrenze del lemma abbastanza ben disperse (assenti solo in narrativa e spettacolo), la seconda accezione in termine di frequenza è quella generale ("portale di chiesa") che raccoglie il 12,6%, anche questa ben dispersa (assente solo in parlato), mentre una sola occorrenza è registrata per l'accezione di tecnica costruttiva o meccanica, rilevata nel sottocorpus stampa, mentre nessuna occorrenza è ricavata nel corpus per l'accezione più innovativa nel dominio della fantascienza.

Dal corpus emergono, inoltre, alcuni modelli d'uso, in particolare per l'accezione informatica, le polirematiche *portale informatico*, *portale aziendale*, *portale di impresa*, *portale generalista*, *portale orizzontale*, *portale verticale* che non sono segnalate in nessuno dei quattro dizionari. Per l'accezione tecnica si trova traccia della locuzione *fresatrice a portale*[63].

Il rapporto tra le diverse articolazioni dei dizionari e il corpus analizzato non è di omologia. Nessuno dei dizionari rappresenta tutte le accezioni attestate nell'ordine di frequenza. In particolare mentre tutti i dizionari d'autore descrivono tutte le accezioni attestate (seppure in ordine diverso rispetto alla frequenza d'uso poiché pongono il senso più frequente all'ultimo posto), il dizionario collaborativo si avvicina a cogliere l'ordine della frequenza d'uso, ma perde la descrizione di una delle accezioni attestate (quella tecnica).

Poiché non è possibile confrontare e valutare in modo puramente quantitativo la somiglianza complessiva nella articolazione dei sensi proposta da ciascun dizionario, si è pensato di ricorrere all'uso di un algoritmo che misura la similarità testuale, simile a quelli usati per l'identificazione automatica del plagio[64]. La misura di similarità testuale scelta usa un algoritmo che determina la differenza tra due file di testo determinando la percentuale di cambiamento introdotta tra

63. Nel GRADIT sono segnalate tre polirematiche per *portale*, per l'accezione tecnica *gru a portale*, per l'accezione informatica *portale orizzontale* e *portale verticale*. Nel corpus del NVDB sono ampiamente attestate le ultime due, rispettivamente con 18 e 17 occorrenze, tutte registrate nel sottocorpus di saggistica.

64. Cfr. R. Lukashenko, V. Graudina, J. Grundspenkis, *Computer-based plagiarism detection methods and tools: an overview*, in "Proceedings of the 2007 international conference on Computer systems and technologies", CCLXXXV, 2007, pp. 36-8.

una versione e un'altra[65]. Un testo confrontato con se stesso ha dunque valore 100, testi che non hanno nessun elemento in comune danno valore 0.

Tabella 7
Similarità tra voci del Wikizionario, DMP e SC

Lemma	DMP/SC	Wikizionario/DMP	Wikizionario/SC
accesso	44,7	41,3	39,9
bar	41,9	38,4	50,4
cabaret	45,4	39,3	38,6
contatto	39,8	16,9	25,7
diocesi	39,4	39,2	42,6
discussione	45,3	27,9	39,8
fischio	39,9	39,9	41,8
gratis	36,9	38,9	50,9
hippie	44,9	39,8	39,8
internazionale	40,0	8,0	15,1
modifica	49,8	37,2	48,7
notizia	44,4	38,9	39,6
opera	37,7	34,3	39,9
portale	53,8	42,4	43,4
portare	40,0	14,0	34,8
proposta	39,8	37,9	39,8
risposta	32,5	22,4	38,9
schiavitù	49,9	38,0	39,4
strumento	44,7	17,4	37,6
tavolo	50,2	39,8	41,6
media arit.	*43,0*	*32,6*	*39,4*

Su 20 lemmi estratti casualmente dal lemmario di Wikizionario e confrontati con le rispettive voci del DMP e SC (cfr. tabella 7) emerge ancora una volta un quadro abbastanza chiaro, anche se più graduale rispetto ai dati sul lemmario, che distanzia moltissimo le voci di Wikizionario da quelle del DMP e soprattutto che indica una similarità globale maggiore tra DMP e SC rispetto a Wikizionario con le altre opere. La media della similarità tra DMP e SC è infatti di circa 43% su ciascuna voce, mentre scende a 39,4% per Wikizionario e SC, e ancora a 32.6% tra Wikizionario e DMP.

La misura della similarità mostra tuttavia solo un aspetto globale sull'elaborazione della voce e riguarda sia la scelta di specifiche forme per descrivere la glossa (le misure di similarità sono indipendenti dalla lingua e valutano il testo come

65. Cfr. E. W. Myers, *An O (ND) difference algorithm and its variations*, in "Algorithmica", 1, 1986, 1, pp. 251-66.

sequenza formale di caratteri), sia l'articolazione in frasi, la lunghezza della voce. Gli aspetti più direttamente connessi alla valutazione qualitativa richiedono invece un'operazione di confronto per certi versi priva di senso, poiché nel nostro caso il dizionario collaborativo non adotta disposizioni esplicite né condivise di gestione di ciascuna voce, che costituisce un *unicum* a sé, a differenza delle voci delle opere autoriali in cui il principio di coerenza e omogeneità nel trattamento delle voci costituisce la vera differenza caratterizzante di un'opera lessicografica rispetto a un'altra.

3.4. Le fonti di attestazione e le fonti lessicografiche

Un discorso a parte richiedono le fonti. La nozione di fonte stessa ha una natura per il dizionario profondamente diversa dalla nozione di fonte di una enciclopedia o di altre tipologie di pubblicazione. Le fonti in lessicografia infatti nella tradizione corrispondono alle attestazioni, tipicamente rappresentate dagli esempi d'uso soprattutto letterari e, concretamente, dai *citation slips*, i foglietti che schedavano e raccoglievano appunto attestazioni rilevanti di ambito letterario, saggistico, pubblicistico. I *citation slips* costituivano il corpus di riferimento dell'opera lessicografica e ne determinavano l'aderenza al dato empirico d'uso. Per l'utilizzo estensivo del sistema della citazione attestata, soprattutto dell'uso contemporaneo, il dizionario di Samuel Johnson viene spesso considerato uno degli antesignani della lessicografia basata su corpora. La fonte empirica in questo senso è intesa come testimonianza o documento utile per la ricostruzione storica. La fonte infatti è dato primario e contiene un elemento determinante per interpretare il dato linguistico, ossia la sua storicità e cronologia d'uso.

In questo senso la linguistica dei corpora (sincronica e diacronica) sistematizza metodologicamente l'estrazione delle fonti di attestazione scritta e parlata per l'elaborazione della voce nella lessicografia storica e in quella d'uso[66]. L'insieme delle fonti di attestazione è dunque la fonte empirica del dato linguistico ed è costituita da un insieme finito di occorrenze parlate e scritte.

Su un piano diverso si colloca la *fonte lessicografica.* Il GRADIT, ad esempio, fa uso di fonti lessicografiche come il Battaglia, il Devoto-Oli, lo Zingarelli[67], rende precisamente conto dell'uso delle opere di riferimento nella selezione e nell'uso che di esse si fa nell'opera. Le fonti lessicografiche vengono usate per la validazione del lemmario e per individuare criteri di condivisione dell'insieme dei lessemi da sottoporre a descrizione. Non è tuttavia la fonte lessicografica il dato empirico, ne costituisce invece una forma di riconoscimento e certificazione e ne testimonia la tradizione.

Dunque il valore empirico di questi due tipi di fonti è radicalmente diverso e la trasparenza nell'esplicitazione dell'uso dell'una o dell'altra tipologia di fonte

66. Per una sintetica introduzione alla lessicografia italiana si veda V. Della Valle, *Dizionari italiani: storia, tipi, struttura*, Carocci, Roma 2005.

67. Cfr. De Mauro, *La fabbrica delle parole*, cit., pp. 42-3.

determina il nesso unico e peculiare che mostra la relazione tra dizionario e fatto linguistico empirico (e storico).

Se si prende il caso di Wikizionario, ciascuna voce è accompagnata da una rubrica di fonti. Tuttavia le fonti del Wikizionario sono esclusivamente del secondo tipo sopra descritto (e tra l'altro tutte fonti autorevoli, *expert-based*, in versione cartacea o informatizzata), ossia fonti puramente lessicografiche. Non risultano mai dichiarate fonti di attestazione. In effetti, nonostante la possibilità effettiva di introdurre nuovi sensi e accezioni non registrati in opere lessicografiche precedenti (come l'uso di *portale* nella narrativa fantascientifica), la principale base empirica dichiarata dal Wikizionario è il ricorso a dizionari per lo più online come Treccani e Sabatini-Coletti.

4
Conclusioni: il dato empirico e il suo uso da parte del lessicografo

Ritorniamo alle quattro sorgenti di evidenza empirica, di natura assai diversa, individuate in apertura (vedi § 1, intuizione, usi dell'opera, fonti di riferimento, attestazioni). L'intuizione, pur messa in questione da approcci empirici fondati sull'analisi del comportamento linguistico, rimane un elemento imprescindibile dell'operazione metalinguistica di elaborazione della voce lessicografica. L'attestazione stessa, che rimane il cardine del lavoro lessicografico, mostra infatti alcuni limiti costitutivi e ineliminabili e richiede oggi con maggiore consapevolezza teorica metodi e modelli da proiettare sul dato.

I dizionari collaborativi potenzialmente offrono l'opportunità di cogliere suggerimenti su usi periferici e innovativi ancora non descritti o rappresentati dalla lessicografia d'autore. Questa opportunità è tuttavia nei fatti ancora ben poco realizzata per carenza di criteri pubblici di attestazione e per il ricorso radicale, e purtroppo spesso mal applicato, alle sole fonti di riferimento esistenti. L'opportunità di affiancare all'intuizione informata del lessicografo quella *naïf* dell'utente ordinario non deve tuttavia essere sottovalutata. Questa infatti non solo getta luce sul modo in cui un parlante concettualizza e concepisce la sua stessa lingua, ma indirettamente fornisce indicazioni sulle attese e i modelli che l'utente stesso ha nei confronti di un'operazione sempre mediata e applicativa come quella dell'uso del dizionario.

In questa direzione si muove un'area rilevante della lessicografia contemporanea che studia appunto l'uso/gli usi che si fanno del dizionario. Si tratta di una forma di evidenza empirica di natura profondamente diversa dalle altre poiché non riguarda il dato che serve alla descrizione linguistica in senso diretto, ma riguarda l'uso applicativo che si fa dell'oggetto concreto dizionario. Tale uso risponde a bisogni linguistici concreti dell'utente e non può non proiettare dunque modelli di descrizione che non siano solo scientificamente adeguati ma anche applicativamente efficienti. In questo senso l'incontro tra ricerca sugli usi del dizionario e le altre fonti empiriche si configura come un processo di forte mediazione, conciliazione che impone a volte sacrificio e

semplificazione della rappresentazione della realtà linguistica nella sua continua variazione.

Gli strumenti di riferimento preesistenti, le fonti lessicografiche, tengono la traccia di questa mediazione nel tempo e cristallizzano la tradizione di questo continuo scambio, senza tuttavia costituire una gabbia interpretativa. La lessicografia italiana come quella britannica danno infatti ampia testimonianza di prospettive, modelli anche radicalmente opposti che storicizzano in maniera diversa il rapporto tra utente e produttore del dizionario nel corso del tempo.

Dal punto di vista della descrizione linguistica non v'è dubbio che la questione delle attestazioni rimanga la fonte empirica principale e costitutiva della lessicografia moderna e contemporanea. L'avvento della linguistica dei corpora non ha fatto altro che accentuare un bisogno ben presente nella tradizione italiana dal *Vocabolario degli Accademici della Crusca*, passando per Tommaseo, Manzoni, Ascoli, tra i tanti. Oggi non manca la disponibilità del dato, ma paradossalmente proprio tale larga disponibilità rende più incerto l'incedere.

In un contributo del 1987 John Sinclair considerava la costituzione di un dizionario completamente automatizzato come in fase avanzata di disegno[68]. L'incontro tra corpora, sistemi di trattamento automatico del materiale linguistico e lessicografia era nelle prime fasi di sperimentazione. Oggi, con una mole di dati significativamente più ampia sul versante delle risorse linguistiche e con numerose esperienze di lessicografia *corpus-based*, il rapporto tra testi scritti e parlati, usi linguistici e metodologie di elaborazione lessicografica si è fatto più stretto mettendo in luce la problematicità nella gestione del dato empirico che esibisce variazioni d'uso secondo una moltitudine di dimensioni difficilmente afferrabili in modo esclusivamente automatico. Questo perché in tali variazioni si intrecciano fattori testuali e linguistici con fattori sociolinguistici, culturali e pragmatici che l'opera lessicografica deve usare come indicatori e guide metalinguistiche.

La grande mole di materiale testuale a disposizione definisce paradossalmente in maniera più evidente il ruolo di filtro, selezione, mediazione e astrazione richiesta al lessicografo[69]. La quantità di occorrenze individuali estratta da corpora in forma di concordanze o di profili lessicali fa emergere regolarità e tendenze solo quando interpretata con strumenti critici complessi. In un certo senso non si può parlare di procedura puramente *bottom-up* per la lessicografia contemporanea (ma nemmeno per la linguistica dei corpora in genere), né in qualche modo di circolarità come sostiene Michael Rundell, ma di una forma di abduzione e di astrazione guidata dall'intuizione e dalla consapevolezza dei fattori sociali e pragmatici che governano le lingue e i loro usi. L'attenzione che la linguistica

68. Cfr. J. Sinclair, *The nature of the evidence*, in *Looking up: account of the cobuild project in lexical computing*, a cura di J. Sinclair, Collins, Glasgow 1987, pp. 150-9.

69. Sulla necessità per il lessicografo di disporre di maggiore acume linguistico nell'era dei corpora si esprime criticamente M. Rundell, *Good old-fashioned lexicography: human Judgment and the limits of automation*, in *Lexicography and natural language processing. A festschrift in honour of* BTS *Atkins*, a cura di M. H. Corréard, EURALEX, Grenoble 2002, pp. 138-55.

contemporanea ha dedicato alla fraseologia inoltre ha permesso di condividere una visione più granulare e continua dei fenomeni di lessico e grammatica[70], e ha sottolineato da prospettive e modelli anche molto diversi la natura *fuzzy* dei confini tra sensi e loro organizzazione interna e ha spinto numerosi lessicografi ad assumere una visione *prototipica* delle accezioni lessicali («word meaning can be regarded as (at best) yet another form of prototype»[71]). E da questo punto di vista è opportuno domandarsi quanta astrazione metalinguistica occorra nella distinzione di tali sensi e quali siano le somiglianze di famiglia che si osservano nelle fisionomie e nei profili che ne emergono.

70. Cfr. T. De Mauro, *On Lexicon and Grammar*, in *Atti del XII Congresso Internazionale di Lessicografia*, a cura di E. Corino, C. Marello, C. Onesti, Edizioni dell'Orso, Alessandria 2006, pp. 19-20.

71. Cfr. Rundell, *Good old-fashioned lexicography*, cit., p. 147.