DATI EMPIRICI E RISORSE LESSICALI

Elisabetta JEZEK*, Isabella* CHIARI[1]

1. Corpora e Risorse Lessicali

La disponibilità di dati linguistici in formato digitale è cresciuta in modo esponenziale negli ultimi 20 anni, stimolando lo sviluppo di modelli per la loro annotazione e di tecniche per la loro analisi statistica, al fine di condurre ricerca linguistica quantitativa e qualitativa e potenziare applicazioni computazionali che prevedono fasi di *machine learning* a partire da dati.

Ciò nonostante, la discussione fondamentale riguardo all'utilizzo di dati linguistici nella costruzione di risorse linguistiche, nella pratica lessicografica (tradizionale e computazionale, cfr. Hanks 2013) e in particolare nella elaborazione di teorie del linguaggio resta aperta e controversa (De Marneffe and Potts 2016), così come spesso sono insufficientemente esplicitati i limiti dell'uso di specifici corpora per la ricerca linguistica, in particolare quella lessicale.

Il termine "risorse lessicali" comprende oggi una vastissima gamma di oggetti: tra questi, versioni online di dizionari cartacei (alcuni tra questi *corpus-based* o *corpus-driven*); *dizionari elettronici* direttamente creati per essere distribuiti esclusivamente online; *dizionari collaborativi* creati da utenti ordinari in progetti volontari; *aggregatori di fonti lessicografiche* (come dictionary.com e thefreedictionary.com); corpora annotati (di lingua scritta, parlata, mista, di dominio specifico, multilingui); lessici computazionali monolingui e multilingui, pensati come *database* lessicali o basi di conoscenze finalizzate non tanto alla consultazione da parte di utenti, ma all'uso e integrazione in applicazioni computazionali, banche dati terminologiche. Sono diffuse le iniziative di standardizzazione degli schemi di annotazione e dei metadati (dati di alto livello, categorie generali volte a favorire interoperabilità e riusabilità delle risorse), metodi innovativi per l'acquisizione di dati (*crowdsourcing*, *gamification*), e iniziative di valutazione e validazione di metodi e risorse.

2. Il dato linguistico

Il dato ha da sempre costituito un elemento centrale nel disegno di opere lessicografiche e nella ricerca linguistica. Come è noto, sono stati tuttavia di volta in volta privilegiati diversi tipi di dati. Per quanto riguarda la sola pratica lessicografica, si possono individuare ad es. (Chiari 2012: 97): 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.

[1] Il paragrafo 1.1 è stato scritto da E. Jezek, il paragrafo 1.2 è invece elaborato da I. Chiari, mentre il paragrafo 1.3 è stato scritto congiuntamente da I. Chiari ed E. Jezek.

Nella lessicografia contemporanea emergono alcuni nodi teorici e applicativi comuni quando si tenti di mettere in comunicazione dati estratti da corpora e *treebanks* e le risorse lessicali. Tra le questioni che si pongono più crucialmente al centro della discussione teorica e applicativa vi sono da una parte le diverse caratteristiche dei corpora disponibili che, per dimensione e composizione qualitativa, non sempre sono adeguati a rispondere ai singoli problemi di ricerca. Fenomeni relativamente rari (come la attestazione dei plurali dei composti italiani) necessitano di corpora molto ampi che tuttavia frequentemente non sono internamente organizzati in modo da permettere una osservazione delle diverse distribuzioni delle forme in competizione in tipologie testuali scritte e parlate diverse. L'uso del web per l'osservazione delle forme ha inoltre numerosi limiti di affidabilità, ripetibilità, verificabilità delle forme in contesto e di estrazione dei dati linguistici.

Il fatto linguistico e il dato empirico utilizzati per la costruzione di risorse lessicali e di *database* lessicografici 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. La costruzione di grandi *database* lessicali multilingui utilizzati in linguistica computazionale ha inoltre comportato la diffusione di risorse di dimensioni enormi ma con una verifica della qualità dei materiali pressoché assente. Si tratta infatti di risorse (esempi come BabelNet, gli allineamenti di Wordnet multilingui, ecc.) in cui l'aggregazione di dati linguistici provenienti da fonti diverse è operato in maniera quasi totalmente automatica e in cui la valutazione della consistenza teorica e linguistica dei dati prodotti è quasi inesistente. I vantaggi relativi dalla disponibilità di risorse di enormi dimensioni non è controbilanciata da una qualità che faccia procedere lo sviluppo della risorsa con gli avanzamenti della teoria e della descrizione linguistica.

Mentre infatti 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), lo sviluppo di risorse applicative non sempre ha percorso le stesse tappe.

La grande mole di materiale testuale a disposizione per l'elaborazione di risorse lessicali finisce per richiedere paradossalmente in maniera più evidente il ruolo di filtro, selezione, mediazione e astrazione del linguista. Tale mediazione è peraltro problematica, come emerge dalla complessità e non univocità delle operazioni di annotazione manuale, che richiedono addestramento, esplicitazione delle operazioni richieste e delle scelte – spesso anche delle convenzioni – operate. Diventa dunque cruciale il ruolo dell'annotatore e la valutazione della coerenza delle annotazioni operate da diversi utenti, siano essi utenti comuni della lingua o linguisti. Laddove invece le operazioni di annotazioni siano operate in maniera automatica il problema della coerenza e della valutazione si sposta a livello di verifica del tasso di errore. Il quadro che emerge conferisce centralità a una visione dei fenomeni lessicali e testuali come fenomeni con regolarità che si polarizzano secondo i tipi testuali e che necessitano di modelli flessibili che diano conto delle varietà attestate nello spazio linguistico delle lingue. Emerge dunque la complessità di dare conto di tale variabilità nella descrizione, spesso statica, fornita dalle risorse. La stessa adozione di terminologia come standard e non standard è un segnale di tale problematicità.

I contributi del numero mostrano in modo evidente l'attenzione ai problemi di trattamento (che sono appunto un prodotto dell'inadeguatezza degli strumenti a trattare il materiale autentico nella sua variabilità) e di annotazione, nonché al delicato problema del rendere conto in modo

sintetico e relativamente 'statico' di questioni dinamiche e variabili a livello di uso linguistico. Anche a livello di interrogazione le criticità che emergono dal trattamento del dato diventano evidenti in quanto, soprattutto quando si interroghino base dati di grandi dimensioni, l'annotazione con i suoi limiti costituisce l'unico filtro possibile per fruire del dato. E tutto ciò che non è restituito dalle possibilità di annotazione per via di incoerenze e inadeguatezze degli strumenti, risulta del tutto invisibile all'occhio di chi consulta le risorse o le integra in applicazioni più complesse.

A questi problemi si aggiunge inoltre un problema interno e costitutivo della lingua in uso stessa, ossia il veloce mutamento del lessico delle lingue (sia a livello di insieme di unità lessicali, ma soprattutto a livello di sviluppo e riorganizzazione dei sensi di lessemi esistenti) che rende le risorse relativamente obsolete in pochissimi anni, con la necessità di manutenzione di risorse e strumenti continua.

Rimane uno spazio da colmare che è quello determinato dai diversi obiettivi delle risorse che si mettono in relazione quando ad esempio si usano corpora per 'informare' risorse lessicali o quando si allineano risorse diverse. Questo spazio è il vero territorio di sfida per la costruzione delle risorse lessicali, poiché richiede una consapevolezza critica dei limiti dei diversi oggetti che si confrontano e del modo in cui intendono rappresentare i dati linguistici, che sola può garantire un esito qualitativamente accettabile e non occasionale nella produzione delle risorse stesse.

In questo quadro la costruzione di risorse lessicali si presenta come uno dei settori che necessitano di una profonda riflessione che metta in comunicazione aree molto diverse della linguistica contemporanea dalla teoria alla descrizione linguistica, dall'annotazione al trattamento, fino agli strumenti di interrogazione. Il rispetto della complessità del dato linguistico infatti richiede uno sforzo di rappresentazione, modellizzazione e descrizione che ancora non ha sempre esiti applicativi adeguati. I contributi del numero cercano di mettere in luce alcuni aspetti critici di questo processo.

3. I contributi del numero

Sono di seguito raccolti quattro contributi che focalizzano l'attenzione sul rapporto tra dati empirici e risorse lessicali e che affrontano e problematizzano dal punto di vista metodologico il valore empirico e i limiti delle diverse fonti che possono essere considerate 'dato' in lessicologia, oltre al modo in cui queste fonti contribuiscono a definire la rappresentazione del lessico di una lingua. In che modo il dato linguistico, inteso in senso ampio, contribuisce a dar forma a diverse rappresentazioni e modelli del significato delle parole e delle loro relazioni semantiche e lessicali, della connessione tra dimensione semantica, comportamento sintattico e collocazionale e dimensione pragmatica (Jezek 2006)?

Dei quattro contributi, il primo focalizza l'attenzione sullo studio di un singolo fenomeno linguistico attraverso l'utilizzo di corpora, mentre i restanti tre descrivono nello specifico problematiche legate alla costruzione di lessici utilizzando dati empirici.

Il contributo di Silvia Micheli prende in esame il problema del modo in cui i dizionari rendono conto di fenomeni caratterizzati da ampia mutevolezza e in quale maniera le soluzioni proposte dai dizionari tengano in considerazione dati empirici estratti da corpora. Il problema viene affrontato mediante alcuni *case studies* relativi alla formazione del plurale dei composti italiani. Il saggio pone alcune questioni critiche rispetto all'uso dei corpora per rispondere a specifiche domande di ricerca a scopo lessicografico e prende in esame i limiti e i vantaggi dell'uso dei motori di ricerca sul web come strumento per la estrazione di dati linguistici e inoltre si prendono in esame due corpora che rispondono a domande di ricerca molto diverse: il corpus del Nuovo Vocabolario di Base e ItWac, mentre le risorse lessicografiche messe a confronto sono

il Gradit e il Devoto-Oli 2014. Si discute in particolare del modo in cui i diversi corpora (diversi per dimensione e per composizione) siano capaci di rendere conto di fenomeni linguistici rari e di come i dati da essi estratti siano generalizzabili per una presentazione lessicografica.

Il secondo contributo riguarda questioni teoriche e applicative nella costruzione del Lessico dell'Italiano Scritto della Svizzera Italiana in Contesto Scolastico. In questo saggio Luca Cignetti e Silvia De Martini presentano un corpus di italiano scritto da bambini e ragazzi della Svizzera Italiana in contesto scolastico. Emergono dunque problemi di trattamento di fenomeni non standardizzati tipici della scrittura giovanile e spesso influenzati da forme di comunicazione mediata dalle tecnologie. Il saggio tratta sia di problemi applicativi e di trattamento sia di possibili approcci didattici all'uso dei materiali raccolti.

Gli ultimi due contributi rivolgono l'attenzione a risorse lessicali *corpus-driven* che uniscono scopi di ricerca linguistica con scopi computazionali. Il contributo di Berta González Saavedra e Marco Passarotti "Verso un lessico di valenza del latino empiricamente motivato" presenta un lessico in cui l'attenzione è volta alle strutture sintattiche associate alle parole dotate di proprietà argomentali. Dopo una rapida introduzione ai lessici di valenza e alle modalità di costruzione (distinte in *intuition-based* e *corpus-driven*) gli autori presentano Latin Vallex, un lessico di valenza per il latino realizzato in stretta connessione con l'annotazione semantico-pragmatica di due *treebank* latine comprensive di testi di epoche e generi diversi, descrivendone la struttura delle entrate lessicali. Il contributo si sofferma in particolare sulle modalità di codifica di tre costruzioni sintattiche: le proposizioni passive, le proposizioni infinitive, l'ablativo assoluto. Il contributo si conclude con una illustrazione delle modalità di interrogazione che mettono ancora una volta in luce lo stretto legame tra il lessico valenziale e la banca dati da cui esso è estratto: il modo cioè in cui una *frame entry* è connessa alle sue occorrenze testuali. Il contributo si conclude con una interessante considerazione relativa ai limiti dell'indagine *corpus-driven*.

Il contributo di Anna Feltracco "T-PAS: costruire una risorsa per l'italiano basata sull'analisi di un corpus" riporta la costruzione di una raccolta di strutture predicato-argomenti con informazione semantica sul tipo semantico delle posizioni argomentali, associata a un corpus di occorrenze e a un repertorio di tipi semantici. La risorsa T-PAS rappresenta un interessante modello complementare ai lessici di valenza sopra descritti, in cui l'attenzione è sulla semantica delle costruzioni anziché sulla sintassi. Nel contributo, l'autrice descrive la procedura di acquisizione delle T-PAS e lo stato dell'arte della risorsa alla quale ha contribuito per la parte lessicografica. Oltre alle specifiche dell'estrazione, che mettono in luce la metodologia utilizzata nella generalizzazione delle strutture a partire dalle occorrenze del corpus, sono riportati dati e caratteristiche salienti della risorsa, e i risultati dei primi esperimenti in cui la risorsa è stata utilizzata. Questi includono un esercizio di disambiguazione del significato del verbo in contesto. Sono infine riportati dati relativi alla valutazione di una porzione della risorsa e le problematicità legate alla annotazione delle relazioni tra T-PAS. Il contributo si chiede con osservazioni relative alle possibilità di espansione semi-automatica della risorsa.

Si ringraziano i membri del Comitato Scientifico del laboratorio "Dati Empirici e Risorse Lessicali" organizzato in occasione del XLIX Congresso internazionale di Studi della SLI (Università di Malta, 24-26 settembre 2015), dove cui i contributi sono stati presentati: Silvia Bernardini, Marco Biffi, Federica Casadei, Ulrich Heid, Alessandro Lenci, Francesco Urzì. Si ringrazia inoltre Elisa Corino per la cura redazionale e Carla Marello per la proposta di accogliere i contributi nella rivista RiCognizioni.

BIBLIOGRAFIA

De Marneffe, M-C., Potts, C. (2016), *Developing linguistic theories using annotated corpora*, in I. Nancy, J. Pustejovsky (eds.), *The Handbook of Linguistic Annotation*, Berlin, Springer, in corso di stampa.

Chiari, I. (2012), *Il dato empirico in lessicografia: dizionari tradizionali e collaborativi a confronto*, in Bollettino di Italianistica. Per Tullio De Mauro, II: 94-125.

Hanks, P. (2013), *Lexical Analysis: Norms and Exploitations*, Cambridge - MA, The MIT Press.

Jezek, E. (2006), *Argument Structure, Verb Patterns and Dictionaries*, in: C. Marello, E. Corino, C. Onesti (a c. di), *Euralex XII Proceedings*, Torino, Edizioni dell'Orso: 1169-1180.

ELISABETTA JEZEK • E. Jezek is Associate Professor in Linguistics at Università di Pavia where she has taught syntax and semantics and applied linguistics since 2001. Her research interests include lexical semantics, verb classification, theory of argument structure, event structure in syntax and semantics, lexicon/ontology interplay, word class systems, and computational lexicography. She has edited a number of major works in lexicography and published contributions focusing on the interplay between corpus analysis, research methodology, and linguistic theory.

E-MAIL • jezek@unipv.it

ISABELLA CHIARI • is Researcher (Assistant professor) in "La Sapienza" University of Rome, where she is holding courses in General linguistics and Computational linguistics for Bachelor, Master and PhD degrees since 2000. She is also Associate at the Institute of Cognitive Sciences and Technologies (ISTC-CNR) and President of Amal for Education, a non profit organization focused on education for refugees. She wrote several essays and articles, directed and participated in numerous research projects. Her research interest focus on statistical and corpus linguistics, Italian lexicography, methodology of linguistic research.

E-MAIL • isabella.chiari@uniroma1.it