Recensione

ANDREA SANSÒ (2007, *ed.*), *Language Resources and Linguistic Theory*, Franco Angeli, Milano, ISBN 978-88-464-8944-9, pp. 212, € 19,00.

1. Il volume *Language Resources and Linguistic Theory*, curato da Andrea Sansò, raccoglie gli interventi presentati nell'ambito del *workshop* del maggio 2006, organizzato a Pavia dal gruppo di ricerca che si occupa da qualche tempo, con proficui risultati, di risorse linguistiche.

Dall'avvento dell'era informatica, il calcolatore è stato utilizzato o «come uno strumento per isolare le parole di un testo, ordinarle alfabeticamente, contarle, ed acquisire, così, i dati per l'elaborazione linguistico statistica» o come paradigma e congegno «per costruire modelli del processo umano di strutturazione, interpretazione e comprensione delle frasi di una lingua» (Ferrari, 1991: 5). La crisi dell'intelligenza artificiale e la progressiva crescita degli strumenti di calcolo hanno determinato, nei primi anni Novanta, una linea di sintesi tra i due paradigmi, che culmina nella creazione di Risorse Linguistiche. Tale termine è stato introdotto da Antonio Zampolli in occasione del suo contributo *Constitution of a European language technology agency* al cosiddetto "Danzin Report" (1992), che aveva lo scopo di proporre la costituzione di una *infrastruttura* linguistica europea. Per tali ragioni, il settore ha subito un'accelerazione nella direzione della fruibilità di grandi insiemi di dati linguistici digitali, quali ad esempio i *corpora*, grazie all'introduzione di innovazioni tecnologiche importanti e alla realizzazione di tecniche di elaborazione linguistica in grado di cogliere anche aspetti mutevoli della lingua nelle diverse variabili diamesiche.

L'intento che anima il curatore, nell'illustrare quello che ha espresso sin dal titolo del volume, è discutere il rapporto esistente tra risorse e teoria linguistica[1]. La tematica è legata al dibattito internazionale, in cui le posizioni scientifiche variano tra chi sostiene la tesi della completa pre-teoricità del *corpus* (Tognini Bonelli, 2001) e chi scorge una dipendenza e un mutuo rapporto tra strutturazione delle risorse linguistiche e sviluppo della teoria (Sgall, 1996). L'antiteticità degli schieramenti induce Godfrey e Zampolli (1997: 382) a considerare problematico il valore della teoria nella costruzione di risorse lingui-

[1] Il tema, sia pure nell'ambito più ristretto dei *corpora*, era stato oggetto di un Congresso nel 1999 (MAIR e HUNDT, 2000).

stiche, e a porsi la domanda se esse debbano essere *theory neutral*. A livello scientifico, la soluzione ravvisata dagli autori è quella di trovare un equilibrio tra piano teorico e approccio progettuale che si concretizza nella creazione di *standard* che confluiscono in linee guida per la codifica di testi; si vedano ad esempio due iniziative di respiro internazionale quali *Text Encoding Iniziative* (TEI) e *Expert Advisory Group on Language Engineering Standards* (EAGLES). Nonostante gli sforzi verso l'uniformità di trattamento dei dati, la reciproca influenza tra risorse linguistiche e teorie rimane una questione non risolta, nella misura in cui l'incidenza degli aspetti teorici muta con il mutare della risorsa stessa. Ad esempio, se da un lato sono ammissibili *corpora* privi di ogni forma di categorizzazione (*raw data*), dall'altro la scelta di una teoria è determinante nei confronti di una *tree-bank*, in cui la strutturazione dei dati sintattici può variare sensibilmente a seconda che si adotti un modello generativo o funzionale, con tutte le ipostasi di ciascuno.

Il rapporto fra risorse linguistiche e teoria linguistica rappresenta, dunque, il filo conduttore del volume. I contributi, pur nelle diverse sensibilità e scelte epistemologiche che li sorreggono, affrontano un'ampia gamma di temi nell'ambito di una problematica vastissima e molto sfaccettata.

La raccolta si concentra soprattutto su diversi tipi di *corpora* e loro utilizzazioni, su *tree-bank* e in un caso su un *database* lessico-semantico. I livelli di analisi linguistica interessati sono quello morfologico, lessicale, sintattico e semantico. Il lettore viene messo di fronte ad un ampio spettro di punti di vista ed esperienze che documenta il dinamismo di quest'area di ricerca.

2. Ad aprire il volume è Sansò con una breve e chiara introduzione intitolata *Language resources and linguistic theory. By way of introduction*, in cui sintetizza il senso del progetto. Il telaio di questa silloge in lingua inglese raccoglie tredici articoli di ventuno studiosi (alcuni dei quali partecipano a più di un saggio) e si suddivide in quattro sezioni: *Corpora and annotation* (E. Hajičová; P. Sgall; F. Dell'Orletta, A. Lenci, S. Montemagni e V. Pirrelli); *Resources for language typology* (M. Cysouw; F. Da Milano, C. Mauri e A. Sansò); *Language resources and second language acquisition* (C. Andorno e S. Rastelli; S. Rastelli); *Applications, corpus-based studies, ongoing projects* (M. Menchi; S. Quaglia; T. Caselli; F. Strik Lievers; M. Formentelli; S. Cerini, V. Compagnoni, A. Demontis, M. Formentelli e C. Gandini).

Visto che alcuni argomenti emergono trasversalmente in più di una sezione, è mia preferenza presentare gli articoli ordinati per linee di ricerca.

3. La tematica trattata nel maggior numero di articoli è quella dell'annotazione. Uno degli aspetti fondamentali della preparazione di un *corpus* è lo

schema di annotazione, ovvero l'insieme di etichette, *tags*, che evidenziano nel testo la categoria dei singoli segmenti linguistici di diverso livello. L'annotazione, pertanto, è una delle interfacce tra teoria astratta ed applicazione classificatoria.

Su questa materia si esprimono inizialmente due studiosi della scuola di Praga, Hajičová e Sgall, che nei rispettivi articoli presentano un progetto già ampiamente documentato in altre sedi, la *Prague Dependency Treebank* (PDT). Una *tree-bank*[2] è un *corpus* in cui il testo viene annotato mediante parentesi che evidenziano le strutture sintattiche ad albero. La *Treebank* di Praga per il cèco si pone come alternativa, nei metodi e nel tipo di annotazione, alla *Penn Treebank*. La principale differenza rispetto a quest'ultima consiste nel tipo di rappresentazione sintattica utilizzata, conforme alle strutture di dipendenza elaborate dalla teoria di Tesnière (1959), sia pure rivista con numerose integrazioni mutuate dalla tradizione linguistica praghese. Come la grammatica di dipendenza, la PDT pone le parole della frase in relazione tra loro mediante *funtori,* evitando l'utilizzo, proprio della grammatica generativa, di simboli intermedi come *noun phrase* o *verb phrase*.

L'originalità della PTD consiste nell'organizzazione dell'annotazione in tre livelli. Il primo è quello morfologico, che corrisponde ad un *POS tagging* esteso in modo da esprimere tutte le dimensioni morfologiche di una lingua altamente flessiva come il cèco. Il secondo, detto analitico, costruisce un albero di dipendenza in cui vengono inseriti tutti gli elementi della frase. Il terzo livello, *tectogrammaticale*, è un albero di dipendenza semplificato in cui appaiono solo gli elementi semanticamente rilevanti, cui vengono associate etichette funzionali come *attore*, *beneficiario* o *paziente*. Il fondamento teorico di questo schema di annotazione è la *Functional Generative Description* (FGD), le cui principali caratteristiche sono la distinzione tra centro e periferia del verbo e il principio di proiettività[3] come collegamento tra la struttura argomentale e distribuzionale.

Nel saggio *Corpus annotation as a test of linguistic theory. The case of Prague Dependency Treebank*, Haijčová considera tre fenomeni linguistici: la struttura dell'informazione data nella distinzione *topic/focus*, la rilevanza del principio di proiettività sull'ordine delle parole e la coreferenza testuale. L'autrice mostra in tutti e tre i casi come il fondamento teorico dell'annotazione faciliti la formulazione di ipotesi scientificamente fondate.

Sgall, invece, con l'articolo *Issues of verb valency in syntactic annotation*

[2] La prima banca dati ad albero creata è la *Penn Treebank* per l'inglese (cfr. MARCUS *et al.*, 1993).

[3] Se un nodo A dipende da B e tra A e B c'è C nell'ordine lineare, C è subordinato di B.

of a large corpus, descrive in modo sistematico l'annotazione della PDT. Lo studioso si focalizza sulla problematicità della valenza verbale, cercando di fornire dei criteri operativi per distinguere gli argomenti dagli aggiunti. La principale regola di discriminazione consiste nel fatto che gli argomenti non possono apparire più di una volta nella frase, mentre gli aggiunti possono cumularsi nella stessa funzione. Sgall verifica il sistema annotativo su un campione di 80.000 frasi del corpus nazionale cèco, e segnala i casi che, rimanendo aperti, richiedono un'integrazione della teoria. In ogni caso, la FGD si realizza in modo adeguato nel sistema annotativo della PDT, offrendo una descrizione del nucleo di una lingua secondo principi generali che, a differenza del modello chomskiano, trovano la loro conferma nel dato empirico.

Entrambi gli articoli sopra descritti si schierano a favore dell'interdipendenza tra la risorsa linguistica e una «*reliable theoretical backbone*», come evidenziato in Hajičová *et al.* (1998: 713).

Diverso è il paradigma a cui si ispirano due articoli della terza sezione sull'annotazione finalizzata agli studi di L2. Si tratta di un progetto che, sia pure meno avanzato del precedente, è ugualmente ambizioso. Già nel primo articolo, *The road not taken. On the way to a parser of learning Italian*, Andorno e Rastelli evidenziano l'inadeguatezza di un'annotazione orientata all'errore, pur ampiamente praticata nell'ambito degli studi sull'apprendimento della seconda lingua. La volontà degli autori è quella di adottare un sistema di annotazione flessibile, tale da permettere una classificazione dinamica degli errori di apprendimento. Il contributo parte da una rassegna dei *corpora* di apprendenti di italiano, classificando le discrepanze tra l'interlingua e la lingua obiettivo. In particolare, l'interlingua contiene forme non presenti nella lingua *target* o forme che non hanno la stessa funzione. Lo strumento proposto per evidenziare l'evolvere di queste differenze è un doppio sistema annotativo: uno orientato alla lingua *target* e uno orientato all'interlingua.

La duplice annotazione è ripresa nel secondo articolo firmato da Rastelli, intitolato *Going beyond errors: position and tendency tags in a learner corpus*. Qui l'autore distingue etichette di *posizione* ed etichette di *tendenza*: le prime definiscono la posizione degli elementi nell'ambito del contesto verbale, mentre le seconde sono funzione della proiettabilità dei tratti del verbo. I principi adottati nell'annotazione sono: la sottospecificazione, ossia l'applicazione del numero minimo dei tratti per ogni posizione, così da non predeterminare la classificazione laddove essa risulti dubbia; la ridondanza, che permette che un elemento sia annotato più di una volta con etichette distinte; l'insensibilità dell'annotazione alla diagnostica dell'errore; infine, il principio di rilevanza, che consente, in caso di ambiguità, di fondarsi sul contesto e sul referente.

Ritornando al primo articolo, con un insolito impianto, si passa immediatamente a descrivere PIl2, un *parser* sintattico e discorsivo per i *corpora* di ap-

prendenti. Il *parser*, definito nel secondo articolo *flat parser* (il cui *output* sembra quello di un *chunker*), compie analisi che prendono come perno il verbo e ne determina tre posizioni precedenti e tre successive, senza distinguere argomenti da aggiunti. L'esempio di pp. 91ss. aiuta a comprendere il processo di analisi dei testi, anche se la chiarezza è compromessa dal fatto che l'ultimo passo (il documento XML) riporta un frammento di testo diverso da quello presentato nelle tabelle precedenti di pp. 92 e 93.

La ricerca illustrata in questi due ultimi contributi appare ancora in una fase di instabilità tecnica, anche se le premesse e gli sviluppi teorici sembrano molto promettenti.

Nella quarta sezione, dedicata a progetti in corso, sono presenti ancora due discussioni sui sistemi annotativi.

Il primo saggio, "*Eco*" *parallel corpus* di Menchi, intende dimostrare la rilevanza dell'allineamento di un corpus parallelo[4] per ricerche cross-linguistiche. L'autrice adotta come campione sette capitoli del romanzo *Il nome della rosa* di Umberto Eco, direttamente collegati con le relative traduzioni in undici lingue indoeuropee, dove la versione italiana è l'elemento cardine per accedere alle traduzioni. I passi della ricerca sono descritti con chiarezza distinguendo la fase di raccolta dati, il *pre-processing* e allineamento delle frasi. Attualmente il *corpus* è allineato solo a livello di frase, anche se nell'articolo è presente un paragrafo dedicato ai problemi posti dall'allineamento a livello di parola. I *corpora* paralleli costituiscono un utile strumento per gli studi sulla traduzione e sono una risorsa importante per l'ambito tecnologico della traduzione automatica, dove la *Example Based Machine Traslation* (EBMT) utilizza esempi tratti da campioni reali per migliorare il lavoro del traduttore. Un altro settore di utilizzo si situa nel campo degli studi di tipologia linguistica (cfr. Wälchili, 2007; Stolz, 2007). Per entrambi gli ambiti sussiste, tuttavia, il rischio che la lingua studiata sia una sorta di *traduzionese*, in quanto la traduzione potrebbe risentire dell'influenza della lingua di partenza.

Nel secondo saggio, *An annotation scheme for bridging anaphors and its evaluation*, Caselli elabora una propria ipotesi di classificazione della *bridging anaphora*, fenomeno connesso con l'uso di sintagmi nominali definiti per introdurre un nuovo referente come anaforico «*not of but via the referent of the antecedent expression*» (Kleiber, 1999: 339). L'annotazione delle catene anaforiche consiste nel collegare tra loro, mediante un indice, le espressioni che coriferiscono[5] con una prima entità introdotta nel testo. Il caso della *bridging*

[4] Sono stati utilizzati programmi *open-source*.

[5] L'espressione "coriferire con" è un termine che nasce dalla posizione teorica secondo la quale sia la prima testa nominale che le successive espressioni di riferimento evocano tutte lo stesso *referente*, perciò tutte "coriferiscono".

anaphora è riconoscibile in frasi come *ho visto una casa, ma il giardino era troppo piccolo,* dove la definitezza del sintagma *il giardino* è motivata dal suo coriferire con un'entità introdotta dal sintagma nominale *una casa.* Dopo una presentazione della teoria di Kleiber, l'autore propone un sistema di annotazione *pragma-cognitivo* ed evidenzia tre dimensioni di analisi: semantica, cotestuale e contestuale. Per quanto riguarda l'influenza della struttura del discorso nella ricerca degli antecedenti, Caselli adotta la teoria del *centering* di Grosz *et al.* (1995) e Poesio *et al.* (2004), che stabilisce una gerarchia tra centri di attenzione (*backward-looking* > *preferred* > *forward-looking center*)[6]. Il *test* dello schema annotativo utilizza un *corpus* di diciassette articoli del *Sole 24 Ore*, da cui sono estratti 1412 sintagmi nominali definiti. Di essi, solo 299 sono propriamente di *brindging* ma costituiscono quasi il 64% del totale dei sintagmi nominali usati come anaforici. I sintagmi nominali sono stati suddivisi in cinque categorie che segnalano l'importanza dell'aspetto lessico-enciclopedico nella struttura del discorso. Si riscontra che nel 70% dei casi le espressioni coreferenziali si trovano nella stessa frase, o al massimo in quella successiva rispetto alla testa a cui si riferiscono; ciò dimostra l'importanza del *focus* locale. La valutazione di questi dati suggerisce alcune integrazioni al modello, e porta alla formulazione di un articolato sistema di annotazione, presentato dettagliatamente nell'articolo. La metodologia scelta dall'autore per *validare* lo schema di annotazione è il calcolo del coefficiente di accordo tra diversi annotatori che utilizzano lo stesso schema (cfr. Carletta, 1996). Anche se i risultati, a detta di Caselli, seppur migliori di altri studi sull'argomento, sono deludenti, il sistema di annotazione può dirsi *validato.*

4. Uno spazio minore ma ben delineato è riservato alla tipologia linguistica, ambito di ricerca nel quale il *database* sembra costituire la risorsa più largamente impiegata.

Si apre questa tematica con l'articolo *A social layer for typological databases* di Cysouw. La constatazione di alcune discrepanze nell'interpretazione dei dati fornite dagli autori di *database* mostra che esse non dipendono da cattive o insufficienti interpretazioni, ma dalla loro mancata contestualizzazione nel dibattito scientifico. Ad esempio, la classificazione della lingua *drehu* è ambigua: Iggesen (2005) la descrive come priva di un sistema di casi, Comrie (2005) come dotata di marcatore di caso attivo/inattivo. Eppure Cysouw puntualizza che ambedue attingono alla stessa fonte bibliografica di Moyse-Faurie

[6] La teoria del *centering* classifica i sintagmi nominali di una frase in tre categorie: quelli che si connettono agli enunciati precedenti (*backword-looking*); quelli che possono collegarsi con gli enunciati successivi (*forword-looking*); quelli potenzialmente collegabili con l'enunciato che segue (*preferred*).

(1993). La contrapposizione si deve al fatto che Comrie considera come marcatore di caso una particella agentiva. Questo parere, anche se difforme da quello di Iggesen, è in linea con una sua osservazione secondo cui in lingue prive di caso morfologico le relazioni possono essere espresse da parole funzionali indipendenti. L'ambiguità di classificazione non si deve ad insufficienza di fonti linguistiche ma dalla discordante categorizzazione di alcune parole funzionali. Per ovviare a casi simili, l'autore propone di aggiungere un livello *sociale* in cui le diverse posizioni adottate dagli studiosi vengano esplicitate e poste a confronto.

Un deciso mutamento di prospettiva è dato dal contributo *Documenting linguistic variation in Europe. The Pavia Typological Database* di Da Milano, Mauri e Sansò, che descrive il *Pavia Typological Database* (PTD), in cui confluiscono dati tipologici provenienti da diverse fonti ma rilevati nell'ambito dei progetti *EuroTyp* e *MedTyp*. La copertura (*coverage*) investe un discreto numero di lingue europee e *circum*-mediterranee, con l'inclusione di alcune lingue caucasiche come il georgiano. L'articolo prende in considerazione la frase relativa, le costruzioni coordinative e gli elementi deittici. I dati relativi ai tre settori sembrerebbero memorizzati e strutturati in archivi separati e diversi. Il modulo delle frasi relative registra esempi da venti lingue, secondo la classificazione di Keenan e Comrie (1977). Nell'ambito di questo archivio vengono classificate le variazioni linguistiche in relazione alla testa e alla funzione sintattica del pronome relativo. La coordinazione presenta esempi da trentun lingue europee, elicitati per mezzo di un questionario e classificati in base alle categorie di combinazione, contrasto e alternativa. L'archivio degli elementi deittici è costituito da un consistente numero di traduzioni di uno dei romanzi che hanno per protagonista *Harry Potter*. Gli esempi sono ordinati in base ad alcuni parametri semantici e le traduzioni si possono visualizzare comparativamente partendo dall'originale inglese. L'articolo, inoltre, si sofferma sui dettagli tecnici della struttura XML. Allo scopo di evitare di far ricorso a formati *proprietari*[7], gli autori preferiscono un'annotazione XML piuttosto che l'uso di un vero *database management system*.

5. Un'altra tematica di grande respiro è quella relativa all'uso dei *corpora* impiegati come base empirica, senza mettere in discussione la struttura dell'informazione in essi contenuta.

Tra i contributi più significati annovererei quello di Dell'Orletta, Lenci, Montemagni e Pirrelli, *Corpus-based modelling of grammar variation*, che costituisce un elegante procedimento di prova a sostegno di una teoria sulla varia-

[7] Formati di archiviazione interni a uno specifico *software*.

zione grammaticale in italiano e cèco. Lo studio utilizza una metodologia statistica basata sul principio della massima entropia di Ratnaparkhi (1998).

L'osservazione di una diversa distribuzione di soggetto e oggetto in lingua cèca e italiana permette di formulare due ipotesi alternative: nella prima, le restrizioni grammaticali delle due lingue sono ordinate diversamente; nella seconda, hanno lo stesso ordinamento ma con pesi diversi. La seconda ipotesi è preferita dagli autori e si colloca nella linea di Manning (2003), che considera le restrizioni sintattiche inerentemente probabilistiche. Il procedimento di prova affrontato riduce il problema ad un compito di identificazione di soggetto e oggetto, verificabile empiricamente su due *corpora*: il PDT per il cèco e *l'Italian Syntactic Semantic Treebank* (ISST) per l'italiano. Le due lingue condividono, a livello strutturale, la libertà delle relazioni grammaticali rispetto al verbo e la possibilità di un'assenza del soggetto (*pro-drop*), mentre differiscono nella marcatura dei casi in quanto il cèco li marca morfologicamente. Riducendo il problema in termini trattabili in base al principio della massima entropia, l'identificazione di soggetto o oggetto si esprime come probabilità che un elemento lessicale assuma una delle due funzioni nel suo contesto, espresso come un insieme di tratti. L'approssimazione della probabilità si ottiene contando quante volte, in un *corpus* annotato, una parola assume la funzione ricercata (soggetto o oggetto) nella totalità dei contesti in cui occorre. Il contesto è rappresentato da un insieme di tratti tra i quali vengono selezionati quelli morfosintattici, l'ordine delle parole, e, secondo una gerarchia di marcatezza, l'animatezza e la definitezza. La verifica del modello si attua utilizzando un sistema istruito, secondo lo schema sopra descritto, per classificare i sintagmi di un *test corpus* in cèco e in italiano. Il risultato è dato dalla proporzione di casi di corretta classificazione e di errore. Al di là delle osservazioni di dettaglio, il punto essenziale è che soggetto e oggetto nelle due lingue si riferiscono allo stesso ordine di salienza informativa pur in due contesti idiosincratici diversi. Le particolarità, quindi, sembrano dipendere da una diversa interazione tra restrizioni, che restano però le stesse in ambedue. Risulterebbe così avvalorata un'interpretazione universalista dell'ordine delle restrizioni; da una serie di osservazioni locali, l'articolo riesce a pervenire a una conclusione rilevante per la discussione sulle proprietà universali del linguaggio.

Un articolato procedimento di prova è utilizzato anche da Quaglia, che presenta un saggio di taglio cognitivo intitolato *Is construction-driven argument superimposition realistic?*, in cui propone uno studio sull'espansione argomentale in italiano, in particolare sulla costruzione del beneficiario esteso. Per beneficiario esteso si intendono quei casi di espressione del beneficiario associati a verbi che non reggono un dativo obbligatorio, come *X cucina Y a Z*. Si osserva che il beneficiario può apparire in numerosi tipi di frasi, espresso come sintagma preposizionale dominato da *a*, come clitico, o come sintagma prepo-

sizionale retto da *per*. Quest'ultima costruzione sembra però esclusa quando il beneficiario è caratterizzato dal tratto [- *animato*] e quando manca l'oggetto. Il paradigma teorico scelto per trattare l'argomento è quello della *Construction Grammar* (Fillmore e Kay, 1993), in particolare mediante il procedimento di *Argument superimposition* sviluppato da Goldberg (1995). L'ipotesi di Quaglia è che anche l'italiano abbia una costruzione dativa che subisce, come l'inglese, estensioni per polisemia o per metafora. I dati per questa ricerca sono forniti dal *corpus* giornalistico *La Repubblica*; sono stati selezionati cento verbi transitivi di cui si controllano le costruzioni con *a*, con *per* e con il clitico. Si nota che i verbi che supportano un beneficiario espresso con *a* sono un numero piuttosto limitato e non tutti preferiscono la realizzazione clitica, ma molti accettano il beneficiario espresso per mezzo di *per*. L'esito più rilevante, però, è l'emergere di un'altra classe di verbi in cui il sintagma preposizionale retto da *a* identifica un possessore non argomentale, esemplificato nella frase *hanno rotto il naso al pugile*. Questa categoria suggerisce che l'elemento discriminante sia il possesso, ossia che l'argomento retto da *a* indichi o il futuro possessore o il possessore corrente. Secondo l'autore questa dicotomia spiegherebbe l'ambiguità di interpretazione di alcuni clitici dativi.

Un esame più diretto del dato empirico, non mediato da un processo di dimostrazione, è descritto negli articoli di Strik Lievers e Formentelli.

Anche Strik Lievers (*Italian perception verbs: a corpus-based study*) rivolge il proprio interesse alla categoria dei verbi, in particolare a quelli di percezione. La struttura gerarchica del campo semantico della percezione (vista/udito > tatto/gusto/odorato) assunta dall'autrice è quella di Viberg (1984). Al fine di classificare la realizzazione sintattica e l'agentività dei verbi, Strik Lievers spoglia il *corpus* PAROLE costituito da 21.000.000 di occorrenze, da cui seleziona i verbi di percezione più frequenti nel *Lessico Italiano di Frequenza* (LIF)[8], pur adottando alcuni correttivi. Per bilanciare le modalità percettive, sono inseriti verbi a bassa frequenza, mentre all'interno della stessa modalità sono selezionati verbi semanticamente distanti tra loro. L'analisi delle distribuzioni fornisce indicazioni varie sui singoli verbi percettivi, mentre sul piano generale si conferma la gerarchia di modalità proposta da Viberg. L'autrice propone, però, una differente organizzazione del campo semantico: i verbi classificati come *percept-based* ne sono posti al di fuori perché non descrivono una percezione, ma un evento che presuppone una percezione, come *odorare*; sono anche categorizzati i verbi che causano una percezione, come *toccare*.

Uno studio pragmatico basato sui sistemi di allocuzione è quello di Formentelli, *The vocative 'mate' in contemporary English: a corpus-based study*.

[8] Cfr. BORTOLINI *et al.* (1972).

Accogliendo la classificazione di Biber *et al.* (1999) sulla classe dei vezzeggiativi e termini di amicizia, l'autore analizza l'uso del vocativo *mate*. La letteratura ne mette in evidenza il carattere di amicizia, mascolinità, rinforzo della relazione, informalità e reciprocità. La fonte di informazione è il campione di parlato accessibile nel *British National Corpus* (BNC), e le occorrenze analizzate sono trecentoventitre, raccolte tra il 1990 e 1994. Alcune delle conclusioni confermano quanto suggerito dalla letteratura, altre lo contraddicono. Formentelli sottolinea un uso di *mate* in concomitanza con schemi di *politeness* sia positiva che negativa.

6. Infine, l'articolo collettivo di Cerini, Compagnoni, Demontis, Formentelli e Gandini, intitolato *A Gold Standard for the evaluation of automatically compiled lexical resources for Opinion Mining*, tocca un argomento che tra le linee tematiche di questa miscellanea resta isolato. L'obiettivo è descrivere un sistema di *validazione* per una risorsa lessicale utilizzata in compiti di *opinion mining*. L'*opinion mining* è una tecnica di ricerca testuale che mira a classificare i documenti sulla base del tipo di opinione che manifestano nei confronti di un determinato argomento. Nella maggioranza dei casi l'opinione è ricostruibile attraverso l'uso di determinati elementi lessicali. Il metodo preso in considerazione in questo contributo è quello basato sull'uso di *WordNet*, un *database* lessicale in cui le parole sono raggruppate in insiemi di sinonimi (*synset*) e connesse fra loro (cfr. Fellbaum, 1998). Per utilizzare *WordNet*, gli autori hanno sviluppato *SentiWordNet*, attribuendo ai *synset* selezionati le categorie di positivo, negativo e oggettivo definite dal lessico *General Inquirer* (GI)[9]. Quest'ultimo strumento costituisce un *benchmark*[10] per la valutazione di *WordNet*. Inoltre, un'interfaccia creata *ad hoc* permette ad un gruppo di valutatori di ascrivere le parole alle tre categorie sopra menzionate. Il risultato di questa ricerca ha evidenziato un discreto accordo tra i valutatori. L'articolo non si propone di compiere una valutazione completa, ma si chiude con la discussione di alcuni casi specifici più problematici.

7. In ultima analisi, la raccolta si caratterizza per gradi diversi di complessità e completezza delle ricerche riferite. Anche se alcuni contributi necessitano di ulteriori approfondimenti, la silloge riesce a dare una visione d'insieme dei progetti nell'ambito delle risorse linguistiche, sia per ricchezza che per varietà di temi generali.

[9] È un sistema di analisi testuale, di vecchia data, che svolge analisi del contenuto (STONE *et al.*, 1967).

[10] Indica un *software* specifico per la valutazione e *validazione* di altro *software*.

A parte qualche inesattezza redazionale, il volume ha indubbio rilievo scientifico e sicura collocazione internazionale.

Bibliografia

BIBER, D. *et al.* (1999), *Longman Grammar of spoken and written English*, Longman, New York.

BOD, R. *et al.* (2003 *eds.*), *Probabilistic Linguistics*, MIT Press, Cambridge.

BORTOLINI, U. *et al.* (1972), *Lessico di frequenza della lingua italiana contemporanea*, Garzanti, Milano.

CARLETTA, J.C. (1996), *Assessing agreement on classification tasks: the kappa statistic*, «Computational Linguistics», XXII, 2, pp. 249-254.

COMRIE, B. (2005), *Alignment of cases marking of full noun phrases*, in HASPELMATH, M. *et al.* (2005, *eds.*), cap. 98, Oxford University Press, Oxford.

DANZIN, A. (1992), *Strategic planning study group: towards a European language infrastructure*, Doc. No. 5210/92 report for the CEC, 31/03/1992.

HAJIČOVA, E. *et al.* (1998), *Language resources need annotations to make them really reusable*, in RUBIO, A. *et al.*, (1998, *ed.*), pp. 713-718.

FELLBAUM, C. (1998), *WordNet. An electronic lexical database*, MIT Press, Cambridge.

FERRARI, G. (1991), *Introduzione al* Natural Language Processing, Calderini, Bologna.

FILLMORE, Ch. e KAY, P. (1993), *Construction Grammar*, unpublished manuscript, University of California, Berkeley.

GODFREY, J. J. e ZAMPOLLI, A. (1997), *Language Resources: overview*, in «Linguistica computazionale», XII-XIII, VARILE, G.B. e ZAMPOLLI, A. (1997, a cura di), numero speciale *Survey of the state of the art in human language Techonology*, pp. 381-384.

GOLDBERG, A. (1995), *Constructions: a Construction Grammar approach to argument structure*, University of Chicago Press, Chicago.

GROSZ, B. *et al.* (1995), *Centering: a framework for modelling the local coherence of discourse*, «Computational Linguistics», XXI, 2, pp. 202-225.

HASPELMATH, M. *et al.* (2005, *eds.*), *The World Atlas of Language Structures*, Oxford University Press, Oxford.

IGGESEN, O.A. (2005), *Number of cases*, in HASPELMATH, M. *et al.* (2005, *eds.*), cap. 49, Oxford University Press, Oxford.

KEENAN, E.L. e COMRIE, B. (1977), *Noun phrase accessibility and Universal Grammar*, in «Linguistic Inquiry», VIII, pp. 63-99.

KLEIBER, G. (1999), *Associative anaphora and part-whole relationship: the condition of alienation and the principle of ontological congruence*, in «Journal of Pragmatics», XXXI, pp. 339-362.

MAIR, C. e HUNDT, M. (2000, *eds.*) *Corpus Linguistics and Linguistic Theory. Papers from the twentieth international conference on English language research on computerized corpora (ICAME 20) Freiburg im Breisgau 1999*, Rodopi B.V., Amsterdam.

MANNING, C.D. (2003), *Probabilistic syntax*, in BOD, R. *et al.* (2003, *eds.*), MIT Press, Cambridge, pp. 289-341.

MARCUS, M. *et al.* (1993), *Building a large annotated corpus of English: the Penn Treebank*, in «Computational Linguistics», XIX, 2, pp. 313-330.

MOYSE-FAURIE, C. (1983), *Le Drehu, langue de Lifou (Iles Loyauté)*, Selaf, Paris.

POESIO, M. *et al.* (2004), *Centering: a parametric theory and its instantiations*, in «Computational Linguistics», XXX, 3, pp. 309-363.

RATNAPARKHI, A. (1998), *Maximum Entropy models for natural language ambiguity resolution*, PhD Thesis, University of Pennsylvania, Philadelphia.

RUBIO, A. (1998, *ed.*), *First International Conference on Language Resources and Evaluation*, Granada.

SGALL, P. (1996), *What linguists may expect and require from syntactic parsing*, in «TELRI Newsletter», III, pp. 9-11.

STOLZ, T. (2007), *Harry Potter meets Le Petit Prince: on the usefulness of parallel literary corpora in cross-linguistic investigations*, in «Sprachtypologie und Universalienforschung (STUF)», LX, 2, pp. 100-117.

STONE, P.J. *et al.* (1966), *The General Inquirer: a computer approach to Content Analysis*, MIT Press, Cambridge.

TESNIERE, L. (1959), *Eléments de syntaxe structurale*, Klinksieck, Paris.

TOGNINI BONELLI, E. (2001), *Corpus Linguistics at Work*, John Benjamins, Amsterdam-Philadelphia.

VIBERG, A. (1984), *The verbs of perception: a typological study*, in «Linguistics», XXI, 1, pp. 123-162.

WŢLCHLI, B. (2007), *Advantages and disadvantages of using parallel texts in typological investigation*, in «Sprachtypologie und Universalienforschung (STUF)», LX, 2, pp. 118-134.

MONICA MOSCA