Lorenzo Spreafico

Misurale le parole. Analisi lessicale quantitativa di un apprendente di italiano L2[1]

Not much is known about the vocabulary acquisition of learners of Italian as a second language. Aim of this paper is to give a first quantitative sketch of the lexical development of a nonnative speaker of Italian through the complete lemmatization of a group of productions included in the so called Banca Dati di Pavia and based on the seminal work done by Broeder *et alii* (1988). The first part of the article (§2) deals with the presentation of the informant, an *ab initio* learner of Italian and native speaker of Tigrinya, the Semitic language of Eritrea. The second one (§3) describes how the lemmatized data base was organized and which problems were to be faced in building it. The last section of the article (§4) introduces the results drawn from the data base and shows how it is possible to relate them with some of the hypotheses advanced to explain the acquisition of words in a second language.

1. *Introduzione*

Dopo anni di trascuratezza, e a partire dalla denuncia di Paul Meara del 1980, gli studi sull'acquisizione del lessico in lingua straniera cominciano ad essere sempre più frequenti, come dimostrano le ricche bibliografie poste a conclusione di due recenti lavori in materia, Singleton (1999) e Nation (2001), oppure contenute nei siti internet di importanti gruppi di ricerca, quali quello della Prifysgol Cymru Abertawe (University of Wales Swansea) <www.swan.ac.uk/cals/calsres/index/index.htm>, e quello della Victoria University of Wellington, Nuova Zelanda <www.vuw.ac.nz/lals/staff/paul_nation/vocrefs.htm>, guidato da Paul Nation. Più di recente poi, il lessico è stato oggetto di attenzione da parte di studiosi di italiano L2, come testimoniano gli articoli di Bozzone Co-

[1] Questo lavoro presenta alcuni risultati delle ricerche in corso per il progetto cofinanziato MIUR/Università degli Studi di Bergamo "Strategie di costruzione del lessico e fattori di organizzazione testuale nelle dinamiche di apprendimento e insegnamento di L2" (biennio 2002-2003, codice 2003105251).

sta (2002), riferito agli errori lessicali in produzioni scritte di apprendenti formali; di Bernini (2003b, in stampa), dedicati rispettivamente all'apprendimento di alcuni verbi pronominali e ad osservazioni sui processi di acquisizione delle proprietà delle parole; di Valentini (2003), incentrato sul ruolo delle parole composte nella compensazione delle lacune lessicali di apprendenti sinofoni di italiano L2.

A differenza di questi ultimi lavori, nel presente contributo si riportano invece i risultati di uno studio pilota volto a testare la fattibilità di una ricerca quantitativa su ampia scala dello sviluppo lessicale di apprendenti italiano L2. Grazie ad essa si vorrebbero fornire sia una prima descrizione dell'andamento del numero di parole conosciute ed utilizzate da un parlante non nativo nei diversi stadi del suo sviluppo linguistico; sia una verifica empirica di alcune delle ipotesi avanzate dai ricercatori circa l'acquisizione del lessico in lingua straniera ed i fattori capaci di influenzare l'apprendimento delle parole.

L'intervento sarà strutturato come segue. Dapprima verrà offerta una breve presentazione dell'apprendente sottoposto ad indagine. In seguito, poiché si è consapevoli che in un progetto di stampo qualitativo più che in altri alcune delle scelte adottate nell'ordinamento dei dati possono influire sugli esiti della ricerca, si renderà conto dei criteri di lavoro seguiti nella stesura dei lemmari che hanno costituito la base di partenza per le analisi. Infine si illustrerà quanto è possibile dedurre dalla lemmatizzazione, ovvero si procederà ad una descrizione macroscopica dell'andamento dello sviluppo lessicale del parlante mettendo in luce le relazioni con alcuni parametri considerati rilevanti nell'acquisizione del vocabolario in lingua straniera.

2. *L'apprendente*

Il soggetto da cui sono stati ricavati i dati analizzati nel progetto pilota è un ventenne eritreo soprannominato Markos (MK)[2]. I dati sono parte della cosiddetta Banca Dati del Progetto di Pavia (BDPV) sull'acquisi-

[2] L'informante è già stato oggetto di studio in precedenza. Cfr. per esempio Bernini (1987) oppure Giacalone Ramat (1999). Per una rassegna completa si rimanda ad Andorno (2001).

zione dell'italiano come lingua seconda[3]. L'informante, immigrato in Italia nel 1986 a seguito dei disordini politici nella sua patria, ha competenze linguistiche di arabo ed inglese, oltre che ovviamente di tigrino di cui è parlante nativo. Il primo è stato appreso durante una breve permanenza a Khartum, nel Sudan. Il secondo, invece, durante la frequenza dei corsi di materie scientifiche presso la scuola media superiore del suo paese d'origine. MK risiede a Milano con la madre, una domestica che parla una varietà elementare di italiano. Al momento della prima rilevazione, effettuata a circa trenta giorni dall'arrivo in Italia, egli è disoccupato. Dopo cinque mesi però comincia a frequentare una scuola professionale che gli dà la possibilità di visitare alcuni cantieri di lavoro. Nonostante l'informante segua un corso di italiano per stranieri per circa tre ore al giorno, il suo contesto d'apprendimento prevalente è quello spontaneo. L'acquisizione dell'italiano da parte di MK è stata monitorata longitudinalmente per cinque mesi, periodo durante il quale egli è stato sottoposto a dodici interviste orali per un totale di 391 minuti di conversazione. Argomento delle registrazioni sono state sia alcune sporadiche attività guidate come il racconto di film, o il commento di vignette fornitegli al momento; sia, più spesso, conversazioni libere, relative soprattutto alle esperienze personali in Eritrea e a quelle quotidiane in Italia. I dati relativi a MK contenuti nella BDPV ammontano a 12 interviste, raccolte nel periodo compreso tra l'ottobre del 1986 ed il giugno del 1987[4].

3. *Presentazione del lemmario*

L'analisi quantitativa dello sviluppo lessicale dell'apprendente si è basata sulla completa lemmatizzazione della trascrizione delle interviste di MK comprese nella BDPV. Tale lemmatizzazione, plasmata sul mo-

[3] Il Progetto di Pavia, così detto dal nome della sede universitaria coordinatrice, nasce nella seconda metà degli anni Ottanta al fine di studiare l'apprendimento spontaneo di italiano L2. I risultati del progetto, che vede coinvolte, oltre a quella di Pavia, le università di Bergamo, Milano Bicocca, Siena (Università per Stranieri), Torino, Trento, Vercelli, Verona, sono stati da poco raccolti in un volume a cura di Anna Giacalone Ramat (2003). La banca dati è invece consultabile sia on line al sito www.unipv.it/wwwling, sia ricorrendo ad un CD-rom curato da Cecilia Andorno (2001).

[4] Di queste però soltanto le prime undici sono state da noi considerate. L'ultima infatti, siglata LTIIMK12, contiene un test grammaticale sulle forme verbali la cui inclusione nelle rilevazioni avrebbe portato ad una rappresentazione distorta della competenza lessicale.

dello in uso presso il progetto *Second Language Acquisition by Adult Immigrants* (Perdue 1993a, 1993b) della European Science Foundation per lo studio dello sviluppo della ricchezza del vocabolario degli apprendenti (Broeder *et alii* 1988, Broeder *et alii* 1993), e mirata ad organizzare i dati linguistici – e nello specifico lessicali – così da semplificare le operazioni di calcolo necessarie a valutare l'andamento della competenza, è stata operata sfruttando un foglio di calcolo elettronico (Microsoft Excel 2000). Grazie ad esso è stato possibile costruire, per ognuna delle interviste considerate, un *file* costituito da una serie di *records*, ovvero di righe contenenti informazioni tra loro correlate, ciascuno dei quali composto da nove campi (colonne) contenenti un elemento informativo, rappresentati da C1, C2, ..., C9 nella figura 1.

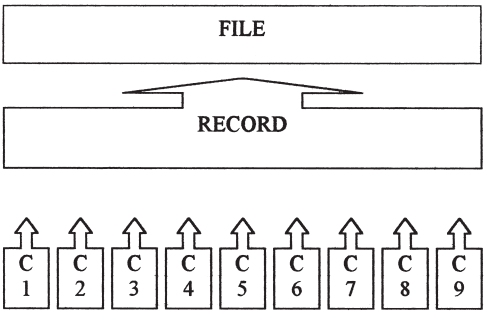


Fig. 1

FO	TO	LE	PO	SI	CL	RI	OR	FR

FO = FONTE TO = TOKEN LE = LEMMA

PO = POLIREMATICA SI = SIGNIFICATO CL = CLASSE

RI = RIPETIZIONE OR = ORIGINE FR = FREQUENZA

Fig. 2

I campi, contrassegnati per comodità da una sigla riproducente le prime due lettere del nome del campo stesso (fig. 2), sono stati compilati così come segue.

Entro il primo, nominato FONTE (FO), è stato riportato il codice alfanumerico sfruttato nella BDPV per identificare ognuna delle trascrizioni dell'apprendente (per esempio: LTIIMK01; LTIIMK02; ...)[5]. In tal modo si è reso possibile sia distinguere tra loro i *records* identici provenienti da registrazioni differenti, sia rintracciare la dislocazione di ciascuna entrata del lemmario.

Entro il secondo campo, definito TOKEN (TO), si è inteso registrare ognuna delle forme di parola presenti nella trascrizione. L'operazione, solo apparentemente semplice, ha presentato due ordini di problemi: uno dovuto alla nozione di parola che si è deciso di adottare; l'altro conseguente alla scelta di lemmatizzare le produzioni dell'apprendente muovendo da trascrizioni piuttosto che da registrazioni audio originali delle stesse.

Per quanto concerne la questione del che cosa sia una parola va ricordato che, sebbene la discussione impegni i linguisti ormai da lungo tempo, tuttavia ad oggi manca ancora una soluzione del problema che possa dirsi generalmente condivisa. Gli ostacoli maggiori all'affermarsi di una visione unitaria sono rappresentati sia dalla difficoltà di rintracciare una definizione di parola valida per tutte le lingue del mondo[6], sia dalla complessità di elaborarne una che coinvolga tutti i livelli in cui si articola un sistema linguistico[7] come proposto da Klein/Perdue (1997: 313). Considerata la natura dei dati su cui si è operato in questo progetto (trascrizioni in caratteri latini) si è deciso di adottare una definizione di parola su chiave ortografica (Singleton 1999: 11), pertanto è stata ritenuta tale ogni sequenza di lettere confinata tra due spazi bianchi[8]. Tale definizione peraltro si presta generalmente bene ad essere usata con trascrizioni operate in accordo con le norme ortografiche della lingua italiana. Gli unici ostacoli sono rappresentati dal trattamento dei pronomi enclitici, delle preposizioni articolate, delle polirematiche.

[5] Il codice in questione contiene già le informazioni necessarie ad identificare l'apprendente, per esempio la madrelingua (T = tigrino) e la natura del rilevamento (L = longitudinale) (Andorno 2001; Andorno/Bernini 2003).

[6] A ciò tenta di rispondere la soluzione elaborata da Ramat (1990: 7) che essendo di natura prototipica offre la possibilità di un'applicazione graduale.

[7] Cfr. per esempio la rassegna svolta da Singleton (1999: 10-14) che distingue proposte su base fonetica, fonologica, morfologica, semantica, collocazionale.

[8] Questa soluzione, usuale nei lavori di linguistica computazionale, sembra peraltro essere la più diffusa in linguistica (Beccaria 1994).

Nel primo di questi tre ostacoli, la registrazione degli enclitici, le difficoltà sono dovute allo *status* di morfemi semiliberi di cui godono i clitici, *status* che impedisce di separarli, nella scrittura così come nella pronuncia, dal verbo cui si reggono. Pertanto, non essendo compresi tra due spazi bianchi, e data la definizione di parola presentata sopra, la sequenza di grafemi che li costituisce non avrebbe avuto diritto alla registrazione entro un campo TOKEN autonomo del *data base* lessicale (fig. 3b). Tuttavia, poiché i pronomi ed i verbi costituiscono due classi di parola chiaramente distinte, e al fine di scongiurare una rappresentazione falsata della ricchezza lessicale dell'apprendente, si è deciso di riportare gli enclitici, oltre che legati al verbo cui si appoggiano (fig. 3a) anche entro *records* distinti da quelli dei verbi cui risultano legati. Ad ogni modo si è avuta l'accortezza di farli precedere nel campo TOKEN da un trattino sottoscritto che ne segnalasse la natura (fig. 3b).

a	parlalo
b	_lo

Fig. 3

Altrettanto è stato fatto nel trattare le preposizioni articolate. Infatti, affinché la competenza lessicale di MK non risultasse sottostimata, gli articoli determinativi contenuti nelle preposizioni articolate sono stati ricondotti a *records* indipendenti così come avviene anche nel Lessico di Frequenza dell'Italiano Parlato (LIP) di De Mauro *et alii* (1993: 63), e resi riconoscibili perché di nuovo inseriti nel campo TOKEN dopo un trattino sottoscritto (Fig. 4).

dei
_i

Fig. 4

Di diversa natura è invece il problema delle poliremiatiche. In tal caso infatti le difficoltà di trattamento sono da ascriversi integralmente alla scelta di adottare un approccio ortografico al concetto di parola, approccio che porta a trascurare la semantica durante la compilazione dei campi, piuttosto che alle regole di scrittura della lingua italiana. Ciò si mani-

festa soprattutto quando ci si imbatte in polirematiche, ovvero locuzioni complesse costituite da diverse parole grafiche il cui significato non è dato dalla somma "algebrica" di quello dei suoi costituenti. Infatti, proprio perché costituite da più sequenze di lettere racchiuse tra spazi bianchi, applicando rigidamente il criterio di riempimento del campo TOKEN esposto in precedenza si giungerebbe ad una perdita di informazioni. Per ovviare a questo problema e compilare con maggiore fedeltà il *data base* si è stabilito di adottare un campo apposito, chiamato POLIREMATICA (PO), grazie al quale segnalare che un dato *token* appartiene appunto ad una locuzione complessa ma dal significato unitario, così da recuperare la dimensione semantica altrimenti integralmente trascurata.

Come anticipato, il secondo ordine di problemi rinvenuto nel corso della registrazione delle forme di parola nel campo *token* è connesso non alla definizione di parola accolta, come per i casi visti sopra, bensì alla scelta di procedere alla lemmatizzazione partendo da trascrizioni piuttosto che da registrazioni audio originali degli incontri tra l'apprendente e l'intervistatore[9].

A tal proposito è opportuna qualche riflessione. Il linguaggio parlato è un flusso continuo di suoni che l'autore della trascrizione provvede a segmentare. Nella BDPV ciò avviene in accordo alle norme ortografiche della lingua italiana (o, eventualmente, della lingua straniera usata dall'apprendente) integrata, quando necessario, da diacritici quali gli accenti gravi o acuti che intervengono a segnalare i casi di rese fonetiche devianti dallo standard della lingua *target* (Andorno 2001, Andorno/Bernini 2003). Se applicata ad un *corpus* di parlanti nativi, la trascrizione può beneficiare dell'esistenza di una grafia storica standardizzata e contare sulla consapevolezza degli informanti circa le proprietà delle parole utilizzate. Invece, quando operata su un *corpus* di interlingua, l'operazione di divisione del *continuum* fonico risulta problematica e presta il

[9] Ciò è stato fatto nonostante si avessero ben presenti le riserve avanzate da Orletti/Testa (1991) circa l'uso di dati trascritti (e dunque manipolati) da altri che non sia il ricercatore impegnato nel progetto. In questo caso però, data la natura pilota del lavoro, si è ritenuto legittimo ricorrere alle trascrizioni delle interviste risparmiando quindi tempo utile da investire nella stesura del *data base*. Si fa comunque presente che il ricorso alle registrazioni originali degli incontri tra l'informante ed il ricercatore è stato immediato ogni qualvolta si siano avuti problemi nel trattare con le trascrizioni.

fianco ad alcune critiche. In particolare, non è detto che tutto quanto risulta separato nella trascrizione lo sia allo stesso modo anche nel lessico mentale dell'apprendente. Al contrario, spesso le segmentazioni svolte in accordo alle convenzioni della lingua di arrivo si differenziano da quelle che è ipotizzabile abbiano luogo nella mente del non nativo (Bettoni 2001: 56). A tal riguardo è possibile ritrovare in MK un valido esempio, comune anche alle produzioni di altri apprendenti di italiano L2. Si tratta della terza persona singolare del presente indicativo del verbo pronominale *esserci*, usata in italiano con significato prevalentemente esistenziale e resa nel parlato da un unico gruppo tonale ['tʃɛ], nello scritto dalla combinazione di due parole grafiche[10]: <c'> ed <è>. Se si fosse certi che l'apprendente riconosce la differenza tra il pronome ed il verbo si potrebbe procedere registrando le due forme di parola in *records* distinti. I dati in nostro possesso lasciano però intendere che ciò non avviene e che anzi nelle fasi iniziali dell'apprendimento ['tʃɛ] assume il ruolo di "an individual lexical item" (Bernini 2003a) con valore non solo esistenziale, ma anche possessivo, copulativo e locativo a seconda del quadro argomentale in cui compare (Bernini 2003b). Per questa ragione sorge il problema di una corretta collocazione della forma <c'è> prodotta da MK entro il campo TOKEN del *record*. Due sono le possibili ipotesi di trattamento cui si è pensato di ricorrere: registrare la forma in accordo a quanto appare nella trascrizione, ovvero distinguere i due *token* <c'> ed <è> (fig. 5), oppure riportare la forma entro un singolo campo *token* quale fosse un unico *item* nonostante sia trascritta con due parole grafiche (fig. 6).

c'
è

Fig. 5

c'è

Fig. 6

La scelta definitiva sul trattamento della forma <c'è> è stata presa dopo aver verificato la decisione assunta da Broeder *et alii* (1988: 30) a

[10] Per il concetto di parola grafica cfr. De Mauro (1989: 51).

proposito del trattamento di forme devianti dallo standard della lingua di arrivo. Tali ricercatori hanno preferito evitare di registrare i *token* degli apprendenti in accordo con la funzione che questi assumono nell'interlingua dei parlanti dal momento che ciò implicherebbe "a lot of time, effort and doubtful interpretation […]" (1988: 30). Essi hanno invece assunto quale principio guida della loro analisi il mantenimento delle distinzioni categoriali presenti nella lingua d'arrivo, che, nel nostro caso, coincidono con quelle riportate dalle trascrizioni. Poiché in questo lavoro si è deciso di fare altrettanto, la forma <c'è> è stata registrata distribuendola su due *records* distinti (vale a dire adottando la soluzione rappresentata in fig. 5). Tale decisione, peraltro, ha consentito di risolvere un altro problema spinoso, ovvero il trattamento uniforme nel lemmario di dati che subiscono una evoluzione nel corso dello sviluppo della competenza linguistica. Infatti, poiché col progredire dell'apprendimento <c'è> viene riconosciuta dall'apprendente come costituita da due elementi distinti ed il paradigma di *esserci* si sviluppa conformemente alla norma della LT (Bernini 2003a, 2003b), si sarebbe imposto di trattare la forma come *item* non separabile inizialmente, e successivamente come costituito da due elementi[11].

Al contrario di quanto verificatosi nella circostanza dell'unione forzata di due elementi distinti quali il pronome e il verbo *essere* nell'*item* <c'è>, la scelta della trascrizione quale punto di partenza per la lemmatizzazione non ha posto problemi nel caso di riduzione di unità inscindibili della lingua bersaglio. Ne è esempio la parola *ufficio*, presente più volte nel *corpus* nelle forme <ficio> o <ficci>. Probabilmente, in alcune circostanze la sillaba iniziale della parola, <uf> è stata analizzata erroneamente come fosse l'articolo indeterminativo <un>, e dunque non è stata pronunciata da MK se già preceduta da un articolo indeterminativo o determinativo (vedi gli ess. 1, 2).

(1) MK 2 \It\ =mh mh ma per esempio dove lavorano?
\Mk\ eh + eh + un mio amico lavora in un - ficio= ficio
\It\ =mh mh
\Mk\ io non so la^_ficio++ ah- lui + parla con me così=
[…] 00:01:09[12]

[11] Del trattamento di <c'è> nel lemmario si tornerà a parlare nuovamente tra breve.

[12] Gli esempi conservano le norme di trascrizione adottate dalla Banca Dati di Pavia. L'intervistatore, l'apprendente ed il mediatore linguistico sono indicati rispettivamente con \It\, \Mk\, \Ki\.

(2) MK 5 \Mk\ in Kartùm anche ci sono_ficci ci sono (poxx) ehm
[…] 00:02:22

In queste occasioni il lemmario ha conservato la peculiarità dell'interlingua, ovvero la riduzione della sillaba iniziale, come testimonia il fatto che la registrazione nel campo *TOKEN* del *record* coincide con la forma di parola presente nella trascrizione (Fig. 7).

Fig. 7

ficio
ficci

Conclusa la discussione di questi due ordini di problemi affrontati nella compilazione del campo TOKEN, passiamo ora a considerare quelli emersi nella stesura del campo LEMMA (LE), vale a dire il campo in cui è stata inserita la forma lemmatizzata di quanto presente nel campo TO del medesimo *record*. Anzitutto si è reso necessario stabilire come procedere nella trasposizione dalle forme di parola alle forme di citazione, ovvero scegliere tra il rispetto delle particolarità dell'interlingua e la proiezione sulla lingua d'arrivo delle produzioni dei parlanti per come sono riprodotte nelle trascrizioni. Ciò è di estrema importanza perché l'assunzione dell'una o dell'altra possibilità comporta sensibili variazioni nel computo del numero di parole presenti nel lessico mentale dell'apprendente.

Ad esempio si consideri nuovamente il caso della forma <c'è> peculiare, come visto sopra, dell'interlingua di apprendenti italiano L2. A tal riguardo si possono ipotizzare quattro diverse modalità di lemmatizzazione, tre delle quali associate alla registrazione del pronome e del verbo quali *token* unitari (Figg. 8, 9, 10); una invece operata a partire dalla collocazione dei due elementi in campi separati (Fig. 11).

I simboli +, ++, +++ indicano pause dalla lunghezza crescente; gli asterischi * * racchiudono parole in lingua straniera; CNV segnala una comunicazione non verbale; / sottolinea autocorrezione; i punti interrogativi ? ? indicano inizio e fine di enunciato interrogativo; % % il volume basso. I numeri tra parentesi alla fine di ogni esempio segnalano anni, mesi e giorni (aa:mm:gg) di permanenza in Italia.

TOKEN	LEMMA
c'è	c'è

Fig. 8

c'è	cessere

Fig. 9

c'è	esserci

Fig. 10

c'	ci
è	essere

Fig. 11

Nel primo caso (Fig. 8) si riporterebbe il *token* nel campo LEMMA senza sottoporlo ad alcun adattamento. Così facendo si riconoscerebbe la particolarità dell'*item* nelle varietà precoci dell'apprendente.

Nella seconda ipotesi invece (Fig. 9) il *token* <c'è> sarebbe associato al lemma **cessere*, ricostruito come l'esito dell'univerbazione del clitico *ci* e del verbo *essere*. Lemmatizzando la forma di parola al modo infinito e mantenendo in posizione iniziale il clitico verrebbe evidenziato il mancato riconoscimento della complessità del *token* da parte dell'apprendente e, contemporaneamente, si ammetterebbe la possibilità che per il non nativo esso possa configurarsi quale verbo autonomo.

Nella terza proposta (Fig. 10) l'*item* <c'è> verrebbe riportato al lemma *esserci*, dove il posizionamento enclitico del pronome starebbe a sottolineare la proiezione della forma presente nell'interlingua sulla lingua *target*, senza però implicare che la forma del lemma faccia necessariamente parte del lessico mentale dell'apprendente.

Nell'ultima variante infine (Fig. 11) il pronome e il verbo verrebbero lemmatizzati autonomamente come fossero forme di parola rispettivamente del pronome *ci* e del verbo *essere*. Proprio questa alternativa, rispettosa della norma della LT così come già lo era stata la compilazione del campo TOKEN, è stata scelta per la redazione del lemmario nel qua-

le, di conseguenza, i lemmi sono sempre stati ricavati in accordo con le convenzioni utilizzate dalle opere lessicografiche della lingua italiana[13].

Oltre a quello della scelta della modalità di formazione dei lemmi, il processo di creazione delle forme di citazione presenta però altre problematiche, comuni sì a tutti i tentativi di lemmatizzazione di *corpora* orali, ma sicuramente accentuate in produzioni di parlanti non nativi.

Innanzitutto talvolta la lemmatizzazione di un *token* è possibile solo grazie al recupero di lemmi presenti nei dizionari della lingua italiana. Si considerino per esempio i *token paurata* e *bugiare* compresi negli esempi (3) e (4) e lemmatizzati rispettivamente come *paurare* e *bugiare* (figg. 12, 13).

(3) MK8 \Mk\ Sì xxxx lei ha **paurato**/? ha **paurata**? 00:04:12

(4) MK11 \Mk\ allora^ se: loro non **bugiàno** così posso^, andare_
se: ho trovato qualcosa non - &non posso tornare 00:06:27

TOKEN	LEMMA
paurata	paurare

Fig. 12

bugiàno	bugiare

Fig. 13

Nonostante siano totalmente assenti dal parlato e dallo scritto quotidiano dei nativi, tali parole vengono registrate dai lessicografi perché inserite in opere letterarie di interesse primario per la storia della lingua italiana. La prima, ad esempio, è riportata nel *Grande Dizionario della Lingua Italiana* (Battaglia 1981) come verbo transitivo dal significato di 'atterrire', 'impaurire', ed è riconducibile al Tommaseo[14]; la seconda invece compare come verbo intransitivo dal significato di 'dire bugie',

[13] Ovvero, i sostantivi, gli articoli e gli aggettivi sono stati riportati al maschile singolare, i verbi al modo infinito, ecc.

[14] *Meditazioni sopra la Passione di Gesù Cristo* [Tommaseo]: "Con grande terrore ei [Gesù Cristo] **paurava** quella povera compagnia. […]".

'mentire' (prima persona singolare indicativo presente: bugìo e bùgio), rintracciabile addirittura nell'opera di Dante Alighieri[15]. Però, proprio perché arcaiche e rare, la presenza di tali occorrenze nell'interlingua del non nativo non può che configurarsi come formazione autonoma dell'apprendente impossibile da ricondurre ad una ripresa di quanto presente nell'*input*.

Ugualmente presente è il problema dell'impossibilità assoluta di riconoscere il significato di alcune forme e dunque dell'incapacità di ricondurle a parole della lingua di arrivo, come per esempio nei casi citati in (5) e (6):

(5) MK1 \Mk\ [...]
*Khartum- Cairo +
ratamilci - Cairo ++
eh: Cairo + Athens + Athens* -un note + un note - hotel
mh + *Athens - Milan* - drito *direct line* 00:01:00

(6) MK4 \Mk\ questo film è c'è la moderno *compiter*
(**quarplessione**) così e c'è tanti sistemi eh anche c'è un uomo non robót come uomo lui andare
[…] 00:02:06

In tali circostanze, poiché è impensabile risalire alla classe grammaticale di appartenenza del *token* si è saturato il campo LEMMA inserendovi la stessa forma di parola rintracciata nella trascrizione (vedi fig. 14).

TOKEN	LEMMA
ratamilci	ratamilci
quarplessione	quarplessione

Fig. 14

Analogo a quello trattato ora è il problema della difficoltà di attribuire con sicurezza un significato a certe produzioni del parlante. In tali casi la

[15] La Divina Commedia, Purgatorio, XVIII: "O gente in cui fervore aguto adesso/ricompie forse negligenza e indugio/da voi per tepidezza in ben far messo,/questi che vive, e certo i' non vi **bugio**,/vuole andar sù, pur che 'l sol ne riluca/però ne dite ond'è presso il pertugio".

lemmatizzazione viene fatta pur in assenza di certezze integrali sul valore del *token* cercando di aiutarsi inferendo la semantica dal contesto. Rappresentativo a proposito è quanto contenuto nell'esempio (7), in cui la parola *nora*, più volte ripetuta nella narrazione di un aneddoto, non è stata immediatamente riconosciuta nemmeno dall'interlocutore. Però, poiché il racconto è ambientato nel mondo delle costruzioni edili, si è ritenuto legittimo associare il *token nora* al lemma *nòria* (fig. 15), sostantivo femminile che si riferisce alla "macchina per sollevare acqua o materiali minuti [che] consta di una serie di tazze fissate su una catena o su un nastro senza fine che scorre tra due tamburi rotanti, mossi da un motore, da un animale, o anche dall'uomo [Dallo sp. *noria*, der. dell'arabo nā' ūra]" (Devoto/Oli 1990: 1247), riportato da Battaglia (1981: XII, 550) anche nella forma *nòra* equivalente a quella usata da MK.

(7) MK4 \Mk\ c'è – così c'è %**nora**%, %**nora**%
adesso io + v(i)engo, dopo io arrivo el el **nora** + tu parli con me, buono buono buono"= sì?
\It\ = [ride] ma cos'è **nora**=
\Mk\ =eh + c'è
\It\ che non ho capito?=
\Mk\ =sì
il – l'italiano= entro – in – la macchina= dopo ++
\It\ =hm =hm
\Mk\ ero – (**nora**) -=
\It\ =ah e rimane dentro un'ora?
\Mk\ (u)n'ora sì, fuori + nella – terra=
\It\ =mh
\Mk\ dopo, io vado con la macchina nella – nella **nora** dopo io eh ++ **nora** è più, questo cemento eh ++
[…] 00:02:06

TOKEN	LEMMA
nora	nòria

Fig. 15

Altrettanto problematico si è rivelato talvolta il riconoscimento della classe di parola di un *token* quale, per esempio, quello di *lavora* eviden-

ziato in (8), che può essere considerata sia terza persona singolare del verbo *lavorare*, sia, più verosimilmente, realizzazione dalla pronuncia errata del sostantivo *lavoro*.

(8) MK2 \Mk\ eh ++ mh – la madre ++ del figli=
\It\ =mh mh
\Mk\ non hai *lavora*= non hai *lavora*
[…] 00:01:09

In tali circostanze si è risolto il problema riconducendo la parola al significato più probabile dato il contesto e la struttura dell'interlingua a quello stadio (fig. 16).

TOKEN	LEMMA
lavora	lavoro

Fig. 16

Un problema analogo si è ripresentato nel compilare il campo CLASSE (CL), ovvero quello in cui è stato inserito un codice numerico sfruttabile per differenziare tra loro quei lemmi che, come in fig. (17), seppur formalmente identici, tuttavia contengono *token* ascrivibili a classi di parola distinte. Poiché come detto poco sopra è stato deciso di assegnare ogni *token* ad un lemma – e quindi implicitamente di stabilirne la classe di parola di appartenenza anche quando l'operazione presentasse qualche perplessità – si è risolto di abolire la distinzione creata da Broeder *et alii* (1988: 34 e Appendix A, Field 3; Field 4) tra *obligatory word class code* e *optional word class code* proprio per segnalare la presenza di casi ambigui[16]. I codici utilizzati per segnalare le categorie grammaticali di appartenenza cui si è fatto ricorso sono riportati in fig. 18.

TOKEN	LEMMA	CLASSE
questa	questo	4
questa	questo	5

Fig. 17

[16] Nell'eventualità di una prosecuzione del progetto si terrà in considerazione la possibilità di recuperare tale distinzione al fine di una migliore raffigurazione dello sviluppo lessicale del parlante.

1	sostantivo	8	quantificatore
2	verbo	9	avverbio
3	aggettivo	10	interiezione
4	articolo	11	particella
5	pronome	12	formula
6	adposizione	13	toponimo
7	congiunzione	14	antroponimo
00	non specificato	99	non chiaro

Fig. 18

Oltre alle classi di parola tradizionali (sostantivi, verbi, aggettivi ecc.), la tabella contiene alcune categorie che meritano un breve commento. Anzitutto toponimi e antroponimi, scorporati dai sostantivi per via della loro semantica particolare che li rende immediatamente apprendibili. Vi sono poi le interiezioni, costituite soprattutto da fonosimboli, ovvero "sequenze foniche che formano segni non necessariamente doppiamente articolati" (De Mauro *et alii* 1993: 91); e le particelle o onomatopee, "sequenze foniche che riproducono o evocano un suono" (De Mauro *et alii* 1993: 93). Inoltre l'etichetta formule è stata attribuita a tutte quelle espressioni non analizzate dall'apprendente che, come i saluti, sono sfruttate soprattutto per la conduzione del discorso segnalando, ad esempio, l'inizio e la fine dell'interazione verbale. Invece si è riportata la dicitura non specificate per segnalare le parole funzione impossibili da riportare ad alcuna classe perché decontestualizzate. Infine l'etichetta non chiaro è stata riutilizzata per i *token* dal significante incomprensibile o irriducibile ad alcun lemma della lingua italiana.

Il campo SIGNIFICATO (SI) è stato compilato inserendovi il significato presunto con cui l'apprendente utilizza il *token* inserito nel *record* cui appartiene[17]. Tale segnalazione, operata riportando la forma di parola corretta richiesta dalla varietà standard della lingua d'arrivo, serve a riconoscere quando il non nativo fa uso di una parola attribuendole una semantica difforme da quella prevista nello standard della lingua d'arrivo.

[17] Qualora il significato d'uso della parola fosse risultato ambiguo o totalmente oscuro, come negli esempi (5) e (6), figura 14, nel campo è stato introdotto un punto interrogativo (?).

Invece il campo RIPETIZIONE (RI) è stato usato per segnalare se un dato *token* costituisse la ripetizione di una parola già usata da un interlocutore nel turno di conversazione immediatamente precedente a quello del non nativo. Tale segnalazione è stata omessa qualora l'apprendente avesse dimostrato di conoscere già il *token* in questione utilizzandolo in uno dei cinque turni di dialogo antecedenti.

Infine, entro il campo ORIGINE (OR) si è provveduto a inserire un codice che aiutasse a riconoscere la provenienza dei *token* usati da Markos, distinguendo quelli ricavati dalla L1 (marcati col codice 1), da quelli ripresi da un'altra L2 diversa anche dalla lingua target (codice 4).

Terminata la redazione di tutti i campi visti sopra, si sono sfruttate le potenzialità del foglio di calcolo per ordinare alfabeticamente tutti i *records*. Quindi si è "ripulito" il lemmario cancellando i doppioni e compilando il campo FREQUENZA (FR) riportandovi la cifra corrispondente al numero di volte con cui il *token* contenuto in un certo *record* si presenta nella trascrizione considerata nello spoglio.

I nove campi appena presentati testimoniano del *come* si sia proceduto a lemmatizzare le interviste trascritte dell'informante. La presenza di alcuni di essi obbliga però ora a spendere qualche parola sul *cosa* si è lemmatizzato, ovvero a chiarire quanto – e perché – di ciò che è riportato nelle trascrizioni si ritrova anche nel *data base* lemmatizzato. In particolare, la presenza dei campi RIPETIZIONE e ORIGINE testimonia che l'approccio seguito nella compilazione dei lemmari è di tipo conservativo (Broeder *et alii* 1988: 32). Ciò vuole dire che, di tutti i dati disponibili, si sono trascurate solo le false partenze, ovvero le interruzioni del parlante, sia spontanee sia indotte da altri. Invece, come visto, nonostante "the inclusion of word repeats in a lexical data base is debatable" (Broeder *et alii* 1988: 32) si sono mantenute le ripetizioni. In particolare si sono riportati nel lemmario sia i *self-repeats* che gli *other-repeats*. Infatti si è ritenuto che i primi, ovvero le iterazioni di uno stesso vocabolo usato dal parlante, abbiano una funzione comunicativa e che i secondi, ovvero le "imitation[s] of a native speaker" (Broeder *et alii* 1988: 32), rappresentino, oltre che una modalità di compensazione dei vuoti lessicali, anche una valida strategia di apprendimento dei nuovi vocaboli[18].

[18] In realtà poi è difficile distinguere, tra gli *other-repeats*, quelle parole che effettivamente costituiscono una ripresa di quanto detto in precedenza dall'interlocutore da quelle che, seppur ripetute, fanno già parte del vocabolario del parlante.

4. *Presentazione dei dati*

Conclusa la trattazione della metodologia adottata nella redazione del lemmario e di alcuni dei suoi aspetti problematici, si passa ora alla verifica della bontà dello strumento elaborato illustrando come possa essere impiegato sia per una generica descrizione iniziale del lessico di un apprendente, sia per la verifica di alcune ipotesi sull'acquisizione del lessico in lingua straniera. Prima di addentrarsi nella discussione di questi argomenti, è però necessario accennare almeno ad alcune delle riserve avanzate circa la validità dei dati ottenuti e la possibilità di trarne generalizzazioni.

Anzitutto va affrontato il tema della legittimità di ricavare conclusioni a proposito dello sviluppo del vocabolario di apprendenti muovendo da produzioni che riflettono solo parzialmente la complessa articolazione del lessico mentale. Infatti, data l'impossibilità di accedere direttamente ai contenuti della mente del parlante, ci si deve accontentare di tracciare l'andamento dello sviluppo lessicale affidandosi esclusivamente alle sue produzioni. Queste però non sono rappresentative della globalità del lessico da egli *conosciuto*, ovvero di quello che non pronuncia mai ma indubbiamente impiega per la comprensione degli enunciati in lingua straniera (Palmberg 1987: 217), bensì soltanto di quello da lui *controllato*, cioè di quello gestito in tutte le sue componenti[19] e dunque impiegato concretamente nell'esecuzione di un compito comunicativo (Bettoni 2001: 67).

Inoltre, a loro volta le produzioni disponibili non contengono che una frazione del lessico controllato dal parlante, vale a dire solamente quella (forse nemmeno tutta) sollecitata dall'interlocutore nel corso delle interviste. Perciò è necessario sottolineare come modificando o la modalità di elicitazione dei dati (affidandosi per esempio a conversazioni non guidate così come a compiti scritti), o le tematiche dei colloqui (scegliendone per esempio alcune che, diversamente dalla narrazione della fuga dalla guerra in patria, per esempio, implichino un minore coinvolgimento emotivo), oppure entrambe le variabili, è possibile si ottengano risultati (profondamente) differenti[20].

[19] A proposito cfr. Laufer (1997).

[20] Tale aspetto, ovvero la cura nella gestione della raccolta dati, deve essere enfatizzato soprattutto nel caso di lavori trasversali in cui si comparino informazioni provenienti da soggetti diversi, interrogati da svariati ricercatori, sollecitati con metodologie eterogenee.

Il secondo tema riguarda la corretta modalità di valutazione dei dati raccolti. In primo luogo va ridimensionata la fiducia posta nel valore assoluto di *token* di una produzione quale indicatore affidabile della competenza lessicale raggiunta da un parlante non nativo. Infatti è possibile che, come ricordano Broeder *et alii* (1988: 40), pur a parità di durata dell'elicitazione, un apprendente iniziale pronunci più parole di uno avanzato. Non è infatti da escludere che egli possa appoggiarsi ad una strategia del tipo *trial and error*, ovvero che cerchi di avvicinarsi per tentativi alla forma di parola più adatta al contesto. Diversamente, un informante che si trovi in uno stadio di competenza più avanzato potrebbe disporre sia di un maggiore numero di risorse lessicali, che di una migliore padronanza delle proprietà di ciascuna di esse e limitarsi dunque ad impiegare, tra le molte a lui note, solo la parola che si adatta al contesto e risponde al bisogno comunicativo. In questo modo, pur a parità di informazione veicolata, il secondo apprendente farebbe uso di un minor numero di parole.

Inoltre, sempre in riferimento alla significatività del valore assoluto di lemmi e *token* quali indicatori del grado di competenza lessicale raggiunta, va sottolineato quanto questi perdano in affidabilità se ricavati da lavori che, come il presente, non prevedano una stretta regimentazione né della durata delle interviste, né della modalità di conduzione delle stesse[21]. Pertanto, poiché è impossibile risalire ad alcuna correlazione diretta tra la durata delle singole interviste ed il numero di lemmi o *token* in esse contenute (vedi tab. 1), si rende necessario adottare criteri di valutazione che siano insensibili a tale parametro.

Lo studio di Broeder *et alii* (1993: 151)[22] propone di utilizzare in associazione il cosiddetto *indice di Guiraud* ed il *theoretical vocabulary*. Il primo strumento, che prende il nome dal suo ideatore (Guiraud 1954) e che talvolta viene chiamato anche *indice de richesse*, propone di valutare la ricchezza lessicale di campioni di testi per mezzo della formula $[K=V / (\sqrt{N})]$, dove *K* corrisponde al valore calcolato per l'indice; *V* al

[21] A giustificazione di ciò sia detto che, visti gli obiettivi iniziali del progetto che ha indotto a raccogliere delle informazione entro la BDPV e, visto l'approccio metodologico adottato per l'analisi dei dati – di tipo qualitativo –, tali limitazioni non risultavano essere necessarie. Inoltre si ricorda che la trasformazione delle modalità di conduzione delle interviste, cioè l'introduzione della presenza contemporanea di un secondo informante oltre a MK a partire dalla settima rilevazione (LTIIMK07), è stata motivata dalla precisa volontà del ricercatore di alleviare quanto più possibile lo *stress* da intervista tentando di creare un clima familiare in modo tale da favorire la spontaneità delle produzioni.

[22] Supportato anche dai dati di altri ricercatori ivi citati.

numero di *type*[23] di un testo; *N* al numero di *token*. Grazie ad esso dovrebbe essere possibile non solo giungere ad una rappresentazione della ricchezza lessicale indipendente dall'estensione del testo in analisi, ma anche ovviare ad un difetto solitamente imputato ad una formula diffusissima ed ideata allo stesso scopo, quella del TTR[24], ovvero l'elevata sensibilità all'eventuale ridotta lunghezza dei campioni in analisi.

Il secondo strumento proposto dalla statistica lessicale per la misurazione della ricchezza di vocabolario in testi dalla diversa estensione è il *theoretical vocabulary* (Menard 1983: 107-117, Cossette 1994: 111-137). Tale metodo, sviluppato da Muller (1964) a partire dalla nota legge statistica detta binomiale, mira a calcolare il numero di lemmi non uguali presenti in campioni di testi di lunghezza eterogenea dopo che questi sono stati tutti riportati alla dimensione del più breve senza che ne siano state intaccate l'originale ricchezza e varietà di parole. Unico limite del *theoretical vocabulary*, che si traduce nella formula $V' = V - \sum (q^i V_i)$, dove V = numero di lemmi presenti nel testo originale; q^i = probabilità che un *token* del testo originale non torni in quello ridotto; V_i = numero di lemmi con frequenza i presenti nel testo originale, è la perdita di affidabilità quando applicato a produzioni dalla lunghezza molto dissimile tra loro.

Terminate queste brevi considerazioni[25] passiamo finalmente a riportare i dati ricavati dalla lemmatizzazione delle produzioni di MK. Anzitutto vengono illustrati, riportandoli divisi per classe grammaticale di appartenenza, i valori relativi al numero assoluto di *token* contenuti in ognuna delle registrazioni (vedi tab. 2).

Una prima osservazione dei dati contenuti nella riga subtotale[26] della tabella permette di rilevare una tendenza crescente nel numero di *token* usati da MK nel corso delle registrazioni. Tale andamento, pur con le riserve appena espresse sulla validità di questo valore al fine della stima dell'evoluzione della ricchezza lessicale di un parlante, lascerebbe sup-

[23] *Type* è definito come l'unione di forme di parola identiche entro un testo.

[24] *Type*/*Token Ratio*, ovvero rapporto tra numero di *token* e di *type* di un campione.

[25] Si rimanda a Menard (1983) e Cossette (1994) per maggiori informazione sulla misura della ricchezza lessicale.

[26] Poiché interiezioni, particelle, formule, parole non chiare o non specificate mancano di significato lessicale, si è preferito escluderle dal computo (subtotale) utilizzato per le analisi contenute in questo paragrafo.

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
Durata (minuti)	18	23	30	25	30	40	25	55	30	40	45
token	636	798	1024	1004	1678	2081	616	1447	1555	944	1591
lemmi	166	176	234	205	303	316	158	321	300	158	346

Tab. 1

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
sostantivi	177	201	231	203	307	332	97	240	259	130	246
verbi	91	118	145	169	282	434	177	346	388	285	400
aggettivi	37	34	54	52	98	113	45	99	100	45	67
articoli	42	118	97	119	114	166	50	128	126	105	130
pronomi	70	74	142	122	295	235	72	177	205	121	214
adposizioni	34	63	111	107	187	213	44	155	147	77	166
congiunzioni	11	31	38	23	69	129	35	98	90	65	66
quantificatori	16	25	27	26	23	45	3	30	12	7	46
avverbi	80	108	106	142	215	290	85	164	190	107	163
toponimi	70	26	44	25	66	83	4	1	34	0	70
antroponimi	8	0	29	16	22	41	4	9	4	2	23
subtotale	636	798	1024	1004	1678	2081	616	1447	1555	944	1591
interiezioni	185	242	240	181	300	408	123	104	278	167	236
particelle	0	2	0	0	0	0	0	0	0	0	0
formule	0	0	1	0	0	3	0	0	0	0	1
non chiari	4	1	4	15	8	0	7	13	9	7	9
non specificati	1	2	2	0	0	6	0	0	0	2	0
totale	826	1045	1271	1200	1986	2498	746	1564	1842	1120	1837

Token

Tab. 2

porre un progressivo incremento della competenza legato al protrarsi dell'esposizione all'*input*. L'andamento che emerge manca però di linearità. Infatti, la produzione più ricca di forme di parola non è l'ultima, come sarebbe forse lecito aspettarsi, bensì la sesta, prodotta a 3 mesi e 12 giorni dall'arrivo in Italia. Sorprendentemente, poi, la settima rilevazione, sostenuta a soli sette giorni dalla precedente, la più ricca del gruppo, risulta essere la meno estesa delle undici analizzate, più povera persino della prima, quella rilasciata dopo appena un mese di permanenza in Italia. Questa anomalia spinge ad un'osservazione più accurata delle informazioni raccolte. Grazie ad essa si scopre la presenza di un andamento regolare nella progressione del numero di forme di parola impiegate, che è costituita da una sequenza uniforme di quattro cicli, ognuno dei quali caratterizzato al suo interno dalla crescita del numero di *token* utilizzati.

Il manifestarsi di un ordine apparentemente non casuale nel numero di forme di parole pronunciate da MK, porta a verificarne l'eventuale presenza anche nella conta dei lemmi. Grazie all'organizzazione del lemmario, e mantenendo la scelta fatta in Broeder *et alii* (1988: 35) di contare come lemma ciascuno dei *record* in cui compare una diversa combinazione dei campi LEMMA, CLASSE, ORIGINE[27], si perviene ai valori raccolti nella tabella alla pagina seguente.

Anche in questo caso risulta che MK tende a servirsi di un numero sempre maggiore di lemmi per comunicare con l'intervistatore. Qui però la sequenza ciclica notata in precedenza per il valore dei *token* pare limitata soltanto alle due serie iniziali, ovvero alle registrazioni 1-3 e 4-6. Infatti, il numero di lemmi di cui MK fa uso durante la nona intervista interrompe la regolarità dell'andamento ondulatorio, o, meglio, ne riduce il periodo a due sole registrazioni.

Date queste osservazioni si può cercare di indagare se la crescita non lineare del numero di lemmi e forme di parola sia casuale oppure riconducibile a cause specifiche identificabili sempre grazie al lemmario. Pertanto si può verificare se esista qualche rapporto tra le classi di parola presenti nell'interlingua e la crescita del vocabolario. A tal fine si può

[27] Si noti come da questa serie sia stato escluso il campo SIGNIFICATO, e dunque come il computo dei lemmi sia effettuato senza considerare la semantica, talvolta peculiare, delle parole dell'interlingua. Un approccio alternativo a tale posizione può essere rintracciato in Bogaards (2001).

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
sostantivi	76	70	91	71	116	110	39	118	97	44	128
verbi	13	28	31	30	46	42	34	61	58	41	52
aggettivi	13	12	19	23	28	28	19	37	40	15	23
articoli	2	2	2	3	2	2	2	2	2	2	2
pronomi	7	9	14	11	20	16	18	19	16	17	24
adposizioni	8	9	6	12	9	13	8	13	10	8	14
congiunzioni	4	5	5	8	11	11	7	14	10	9	13
quantificatori	11	13	16	10	15	19	3	15	6	1	23
avverbi	11	20	23	22	29	32	22	34	37	20	30
toponimi	18	8	17	9	15	24	3	1	21	0	24
antroponimi	3	0	10	6	12	19	3	7	3	1	13
subtotale	166	176	234	205	303	316	158	321	300	158	346
interiezioni	5	5	6	6	5	10	4	7	14	7	10
particelle	0	2	0	0	0	0	0	0	0	0	0
formule	0	0	1	0	0	1	0	0	0	0	1
non chiari	4	1	4	7	6	6	7	13	9	7	9
non specificati	1	2	2	0	0	0	0	0	0	2	0
totale	176	186	247	218	314	333	169	341	323	174	366

Lemmi Tab. 3

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
funzione	66,02	55,24	52,57	53,33	50,81	54,19	61,35	55,11	56,81	55,56	55,31
contenuto	33,98	44,76	47,43	46,67	49,19	45,81	38,65	44,89	43,19	44,44	44,69

Token Tab. 4

studiare la distribuzione delle parole contenuto rispetto alle parole funzione al procedere delle registrazioni.

Il calcolo dell'incidenza percentuale di ciascun gruppo si fa recuperando le indicazioni di Broeder *et alii* (1988: 60), ovvero riconducendo sostantivi, verbi e aggettivi alla classe aperta di parole contenuto e articoli, pronomi, adposizioni, congiunzioni alla classe chiusa di parole funzione. Si evita invece di conteggiare avverbi e quantificatori: i primi perché ambiguamente collocati a cavallo tra le due classi, i secondi perché costituiti essenzialmente da numerali. Calcolando il peso percentuale di ciascun gruppo sul totale delle parole contenuto e funzione si ottengono i valori riportati nelle tabelle 4 e 5, riferite, rispettivamente, al numero di *token* e a quello di lemmi.

Confrontando i dati con quelli medi ricavati da Broeder *et alii* (1988: 61) per l'insieme degli apprendenti impegnati in attività di conversazione libera, le uniche compatibili con la metodologia delle interviste semiguidate effettuate da MK, si vede (tabella 6) come sia i valori calcolati per i *token*, sia quelli calcolati per i lemmi, risultino molto simili nonostante le notevoli differenze tra le L2 coinvolte nei progetti.

I valori calcolati per Markos indicano una riduzione dell'incidenza percentuale delle parole contenuto rispetto a quella delle parole funzione. L'andamento, in linea con le previsioni di un rafforzamento della classe chiusa di parole nel corso dell'acquisizione di una lingua straniera (Broeder *et alii* 1988: 61), non è però interessato da alcun fenomeno di ricorsività evidente che permetta di collegare l'evoluzione dei due gruppi con i valori emersi nelle tabelle 2 e 3. Infatti, invece che un andamento ciclico si verifica una tendenza alla stabilità dei rapporti tra le classi. Però, mentre per i *token* la differenza percentuale tra parole contenuto e parole funzione, inizialmente molto elevata, si riduce e stabilizza sin dalla seconda rilevazione, per i lemmi si assiste ad un costante distacco tra i due tipi.

Le uniche eccezioni all'assenza di sensibili oscillazioni sono presenti nella settima e nella decima registrazione. Per cercare di fare maggiore chiarezza grazie al lemmario si è potuto analizzare in dettaglio la distribuzione percentuale delle diverse classi di parola sia per quanto riguarda il numero di *token* (tab. 7) che per quanto riguarda il numero di lemmi (tab. 8).

Dalla tabella 7 emerge come in MK7 la classe dei verbi sia molto più

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
funzione	82,93	81,48	83,93	78,48	81,90	81,08	72,44	81,82	83,69	73,53	79,30
contenuto	17,07	18,52	16,07	21,52	18,10	18,92	27,56	18,18	16,31	26,47	20,70

Lemmi Tab. 5

		1° CICLO	2° CICLO	3° CICLO
TOKEN	contenuto	57,10	55,72	54,01
	funzione	42,90	44,28	45,99
LEMMI	contenuto	77,01	73,92	76,95
	funzione	22,99	26,08	23,05

Tab. 6

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
sostantivi	27,83	25,19	22,56	20,22	18,30	15,95	15,75	16,59	16,66	13,77	15,46
verbi	14,31	14,79	14,16	16,83	16,81	20,86	28,73	23,91	24,95	30,19	25,14
aggettivi	5,82	4,26	5,27	5,18	5,84	5,43	7,31	6,84	6,43	4,77	4,21
articoli	6,60	14,79	9,47	11,85	6,79	7,98	8,12	8,85	8,10	11,12	8,17
pronomi	11,01	9,27	13,87	12,15	17,58	11,29	11,69	12,23	13,18	12,82	13,45
adposizioni	5,35	7,89	10,84	10,66	11,14	10,24	7,14	10,71	9,45	8,16	10,43
congiunzioni	1,73	3,88	3,71	2,29	4,11	6,20	5,68	6,77	5,79	6,89	4,15
quantificatori	2,52	3,13	2,64	2,59	1,37	2,16	0,49	2,07	0,77	0,74	2,89
avverbi	12,58	13,53	10,35	14,14	12,81	13,94	13,80	11,33	12,22	11,33	10,25
toponimi	11,01	3,26	4,30	2,49	3,93	3,99	0,65	0,07	2,19	0,00	4,40
antroponimi	1,26	0,00	2,83	1,59	1,31	1,97	0,65	0,62	0,26	0,21	1,45
totale	100	100	100	100	100	100	100	100	100	100	100

Token Tab. 7

ricca delle altre. Anche nella tabella 8 si assiste alla sensibile avanzata dei verbi, che però non giungono a superare la quota dei sostantivi. Inoltre colpisce molto il dato dei pronomi (11,39%) che, sia in MK7 sia in MK10, crescono sino a raddoppiare il valore medio precedente attestato intorno al 5,5%. In nessun caso però emerge una chiara relazione tra l'andamento delle classi di parola e i cicli regolari rintracciati inizialmente. Pare dunque poco probabile che i due fattori siano tra loro correlati.

Per scongiurare ogni possibile errore dovuto al trattamento di dati ricavati da registrazioni aventi diversa lunghezza si fa quindi ricorso al *theoretical vocabulary* ed all'indice di ricchezza introdotti poco sopra, così da scoprire se veramente la ciclicità emersa nella tabella 2 e parzialmente replicata nella tabella 3 sia totalmente occasionale. Ricorrendo alla formula presentata sopra sono stati calcolati due valori possibili per il *theoretical vocabulary* (vedi tab. 9).

Il primo è stato ricavato riconducendo statisticamente tutte le trascrizioni alla lunghezza di 616 *token*, ovvero a quella rinvenuta nella trascrizione più povera del *corpus* (MK7). Le cifre ottenute, pur confermando la tendenza alla crescita della ricchezza lessicale già ipotizzata analizzando il valore assoluto di *token* e lemmi delle tabelle 2 e 3, smentiscono però definitivamente l'esistenza di ogni genere di ciclicità. Inoltre chiariscono come, a differenza di quanto dedotto dalle prime tabelle, e così come preventivabile, a parità di condizioni[28] sia proprio l'ultima registrazione quella più ricca di lemmi dell'intera raccolta. Ciò confermerebbe le attese iniziali di un progressivo incremento della competenza lessicale legata al protrarsi della permanenza in Italia e dell'esposizione all'*input*. Ciò che colpisce della tabella 8 è che la più povera tra le registrazioni sia la penultima (MK10) e non, come forse si poteva attendere, la prima. Comunque, poiché il valore ricavato dall'applicazione della formula del *theoretical vocabulary* per questa trascrizione si pone fortemente al di fuori dell'ordine di grandezza ottenuto per le altre interviste, si ritiene che siano intervenuti fattori occasionali capaci di provocare un peggioramento della prestazione[29].

[28] Fatta salva l'eterogeneità di argomenti discussi durante l'intervista.

[29] Un'ipotesi da verificare è che l'utilizzo di un nuovo tipo di attività nell'intervista (non più incentrata esclusivamente su conversazioni semilibere bensì su attività guidate come la narrazione degli eventi di un film muto) possa aver fatto peggiorare il livello della prestazione il che, peraltro, coinciderebbe con quanto notato da Broeder *et alii* (1988: 46).

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
sostantivi	45,78	39,77	38,89	34,63	38,28	34,81	24,68	36,76	32,33	27,85	36,99
verbi	7,83	15,91	13,25	14,63	15,18	13,29	21,52	19,00	19,33	25,95	15,03
aggettivi	7,83	6,82	8,12	11,22	9,24	8,86	12,03	11,53	13,33	9,49	6,65
articoli	1,20	1,14	0,85	1,46	0,66	0,63	1,27	0,62	0,67	1,27	0,58
pronomi	4,22	5,11	5,98	5,37	6,60	5,06	11,39	5,92	5,33	10,76	6,94
adposizioni	4,82	5,11	2,56	5,85	2,97	4,11	5,06	4,05	3,33	5,06	4,05
congiunzioni	2,41	2,84	2,14	3,90	3,63	3,48	4,43	4,36	3,33	5,70	3,76
quantificatori	6,63	7,39	6,84	4,88	4,95	6,01	1,90	4,67	2,00	0,63	6,65
avverbi	6,63	11,36	9,83	10,73	9,57	10,13	13,92	10,59	12,33	12,66	8,67
toponimi	10,84	4,55	7,26	4,39	4,95	7,59	1,90	0,31	7,00	0,00	6,94
antroponimi	1,81	0,00	4,27	2,93	3,96	6,01	1,90	2,18	1,00	0,63	3,76
totale	100	100	100	100	100	100	100	100	100	100	100

Lemmi Tab. 8

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
616	164,07	159,18	184,94	169,23	190,32	188,77	158,00	166,50	200,49	132,35	217,69
100	59,34	57,08	58,43	59,84	60,97	59,78	32,60	46,30	66,02	52,00	66,54

Tab. 9

	MK1	MK2	MK3	MK4	MK5	MK6	MK7	MK8	MK9	MK10	MK11
indice di Guiraud	8,49	8,64	10,25	9,15	10,89	11,55	9,63	11,80	12,55	8,85	12,89

Tab. 11

valore assoluto	26	10	11	14	26	19	3	2	8	1	13
%	4,09	1,25	1,11	1,36	1,55	0,91	0,49	0,14	0,51	0,1	0,81

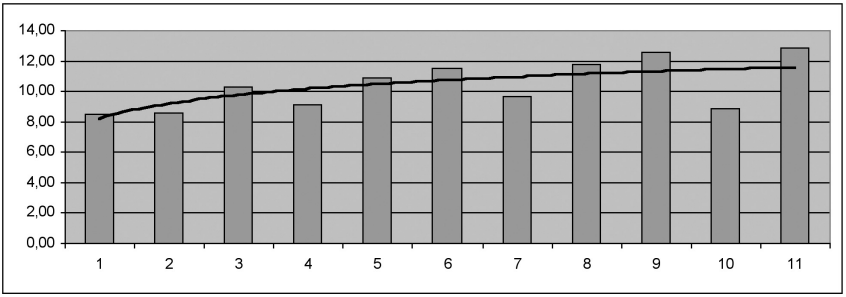
Tab. 12

ciclo	1	2	3
media	59,20	61,87	60,97

Tab. 10

La seconda riga riportata in tabella 9 si riferisce invece all'ammontare del vocabolario teorico computato dopo aver ridotto tutti i testi originali all'ipotetica lunghezza di 100 *token*. La scelta di tale valore è stata attuata per permettere un confronto tra i valori calcolati da Broeder *et alii* (1988: 53) per gli informanti del proprio progetto ed i risultati elicitati da MK. Se si considerano le cifre ricavate dagli autori della ricerca europea per i compiti di conversazione libera, stupisce quanto i valori si avvicinino a quelli presenti nella tabella 9. Infatti il valore medio calcolato per MK (56,26 lemmi) non si discosta significativamente da quello ricavato nel primo ciclo di raccolta dei dati del progetto della European Science Foundation (Broeder *et alii* 1988: 55) (tabella 10) che si riferisce a soggetti impegnati nell'apprendimento della L2 da un lasso di tempo simile o di poco superiore a quello speso in Italia da Markos al termine delle registrazioni.

Come detto, accanto al computo del *theoretical vocabulary* esiste un secondo metodo per ricavare un'idea dell'andamento della ricchezza lessicale di un apprendente indipendentemente dalla lunghezza dei testi da questi prodotti: l'*indice di Guiraud*. Gli esiti dell'applicazione di tale strumento alle informazioni ricavate dalla BDPV per MK figurano nella tabella 11, che tradotta graficamente produce la linea di tendenza visualizzata nel grafico 1.



Graf. 1

Tale rappresentazione si approssima a quella riportata in Broeder *et alii* (1988: 43) e può essere interpretata come indicativa di un contesto di arricchimento lessicale che va progressivamente rallentando. Come visto grazie all'uso combinato di *theoretical vocabulary* ed *indice de richesse* applicati ai dati estratti dal lemmario è possibile operare una prima descrizione generale dello sviluppo lessicale di MK, ovvero affermare che si verifica una tendenza alla crescita lessicale nel tempo, ancorché interessata da un rateo di incremento sempre più ridotto. Però, oltre che per una macroscopica descrizione dello sviluppo lessicale di un parlante, come detto, il lemmario si presta anche alla verifica di alcune ipotesi sull'acquisizione del lessico. Qui si darà ora brevemente conto di una di esse ancora in fase di elaborazione e della quale ci si riserva di dare adeguata documentazione in futuro.

Uno dei fattori proposti in letteratura tra i più influenti sull'apprendimento delle parole è la frequenza delle stesse nell'*input* (Ellis 1997, Bettoni 2001: 68, Ellis 2002: 152). Tramite il lemmario è possibile rilevare se ciò sia confermato anche nel caso dell'apprendente MK. Infatti, facendo ricorso al LIP di De Mauro *et alii* (1993), ed in particolare al sottoinsieme di dati relativi alla città di Milano, ovvero quella di residenza di MK, qualora siano registrate, è possibile determinare il rango d'uso di ciascuna delle forme di parola sfruttate dall'informante. In tal modo è possibile constatare se vi sia una prevalenza di parole dal basso rango d'uso (ovvero frequenti) nelle fasi iniziali del soggiorno in Italia, e contestualmente una loro diminuzione nel tempo. Se così fosse allora sarebbe possibile sostenere legittimamente di aver trovato una conferma dell'ipotesi. Peraltro, qualora ciò si avverasse, sarebbe possibile pensare di adottare il *Lexical Frequency Profile* proposto da Laufer/Nation (1995) anche alle produzioni orali di apprendenti di italiano L2. Il *Lexical Frequency Profile* (LFP) costituisce una nuova misura della ricchezza lessicale di apprendenti basata sul computo della frequenza nelle produzioni di parlanti non nativi di parole dalla decrescente reperibilità nell'*input*. Il presupposto è che, se davvero le parole più disponibili vengono apprese prima delle altre, allora il ritrovare negli elaborati degli informanti parole poco diffuse è indice di sempre maggiore competenza[30].

[30] Una tale supposizione è stata avanzata anche da Blok-Boas (1997) nel trattare dell'ipotetica composizione del lessico mentale di apprendenti avanzati di italiano L2.

Qualora si riuscisse a trasferire il sistema elaborato da Laufer/Nation (1995) anche alle produzioni di apprendenti italiano L2, si potrebbe finalmente disporre di uno strumento oggettivo per diagnosticare il livello di competenza lessicale di parlanti non nativi di italiano.

Oltre che per la verifica di ipotesi sulle variabili che semplificano o complicano l'apprendibilità delle parole, il lemmario può essere applicato inoltre per la verifica della frequenza con cui gli apprendenti ricorrono a strategie interlinguali quali il prestito da lingue diverse da quella di arrivo per la compensazione delle proprie lacune lessicali. Infatti, grazie alla presenza del campo ORIGINE entro cui è segnalata la lingua di appartenenza di ciascun *token* dell'informante, si riesce a scoprire se e quali lingue siano da lui impiegate durante la conversazione con altri individui.

5. *Conclusione*

L'obiettivo di questo articolo, annunciato nell'introduzione del contributo, era la verifica della fattibilità di uno studio quantitativo volto ad indagare lo sviluppo della competenza lessicale di apprendenti italiano L2. I dati raccolti e qui presentati paiono dimostrare che l'operazione, nonostante i lunghi tempi di lavoro necessari per la sua attuazione, è fattibile.

In particolare la lemmatizzazione, soprattutto se svolta in combinazione con strumenti elaborati dalla statistica lessicale quali il *theoretical vocabulary* e l'indice di *Guiraud,* sembra risultare di grande aiuto per procedere ad una descrizione iniziale del repertorio lessicale di un apprendente, eventualmente da approfondirsi poi per mezzo di ricerche di marca qualitativa. Inoltre, il lemmario pare dimostrarsi funzionale alla verifica delle ipotesi qualitative avanzate per giustificare l'acquisizione del vocabolario in lingua straniera. Al contrario invece tale strumento sembra essere poco efficace nel fare emergere nuove ipotesi che rendano conto dell'apprendimento delle parole straniere da parte di un non nativo. Per questi motivi non pare irragionevole pensare di procedere alla lemmatizzazione delle produzioni di un numero maggiore di apprendenti italiano L2, anche al fine di scoprire se le ricorsività apparenti rilevate in precedenza sono confermate anche in altri parlanti, oppure costituiscono esclusivamente un dato peculiare di Markos.

BIBLIOGRAFIA

Andorno, Cecilia (a cura di), 2001, *Banca dati di italiano L2* (CD-Rom), Dipartimento di Linguistica, Università degli Studi di Pavia. (cfr. http://www.unipv.it/wwwling)

Andorno, Cecilia/Bernini, Giuliano, 2003, "Premesse teoriche e metodologiche". In: Giacalone Ramat (a cura di): 27-36.

Battaglia, Salvatore, 1981, *Grande Dizionario della Lingua italiana*, Torino, Unione Tipografica Editrice Torinese.

Beccaria, Gian Luigi (a cura di), 1994, *Dizionario di linguistica: e di filologia, metrica e retorica*, Torino, Einaudi.

Bernini, Giuliano, 1987, "Le preposizioni nell'italiano lingua seconda". *Quaderni del Dipartimento di Linguistica e di letterature comparate* 3: 129-152.

Bernini, Giuliano, 2003a, "The copula in learner Italian: Finiteness and verbal inflection". In: Dimroth, Christine/Starren, Marianne (eds.), *The dynamics of learner varieties*, Berlin-New York, Mouton de Gruyter: 159-185.

Bernini, Giuliano, 2003b, "Verbi pronominali: la costruzione del lessico in italiano L2". Comunicazione tenuta alla Giornata di Studio su *Morfologia, tipologia e acquisizione di lingue seconde* (Milano-Bicocca, 30 maggio 2003).

Bernini, Giuliano, in stampa, "Come si imparano le parole: osservazioni sull'acquisizione del lessico in L2". *ITALS*.

Bettoni, Camilla, 2001, *Imparare un'altra lingua*, Bari, Editori Laterza.

Blok-Boas, Atie, 1997, "Italiano L2: il lessico mentale del discente avanzato". *Rassegna Italiana di Linguistica Applicata* 29/3: 155-170.

Bogaards, Paul, 2001, "Lexical units and the learning of foreign language vocabulary". *Studies in second language acquisition* 23: 321-343.

Bozzone Costa, Rosella, 2002, "Rassegna degli errori lessicali in testi scritti da apprendenti elementari, intermedi ed avanzati di italiano L2 (ed implicazioni didattiche)". *Linguistica e Filologia* 14: 37-67.

Broeder, Peter/Extra, Guus/van Hout, Roeland/Strömqvist, Sven/Voionmaa, Kaarlo (eds.), 1988, *Processes in the developing lexicon*, (= *Final Report to the European Science Foundation*, III), Strasbourg, Tilburg, Göteborg.

Broeder, Peter/Extra, Guus/van Hout, Roeland, 1993, "Richness and variety in the developing lexicon". In: Perdue, Clive (ed.), *Adult language acquisition: cross-linguistic perspectives. Volume I. Field methods*, Cambridge, Cambridge University Press: 145-172.

Cossette, André, 1994, *La Richesse Lexicale et sa Mesure*, Paris, Honoré Champion Éditeur.

De Mauro, Tullio, 1989[10], *Guida all'uso delle parole*, Roma, Editori Riuniti.

De Mauro, Tullio/Mancini, Federico/Vedovelli, Massimo/Voghera, Miriam (a cura di), 1993, *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri.

Devoto, Giacomo/Oli, Gian Carlo, 1990, *Il dizionario della lingua italiana*, Firenze, Le Monnier.

Ellis, C. Nick, 2002, "Frequency effects in language processing". *Studies in second language acquisition* 24: 143-188.

Ellis, Rod, 1997, *Second language acquisition*, Oxford, Oxford University Press.

Giacalone Ramat, Anna, 1999, "Grammaticalization of Modality in Language Acquisition". *Studies in Language* 23/2: 377-407.

Giacalone Ramat, Anna, (a cura di), 2003, *Verso l'italiano*, Roma, Carocci.

Guiraud, Pierre, 1954, *Caractères statistiques du vocabulaire. Essai de méthodologie*, Paris, P.U.F.

Klein, Wolfgang/Perdue, Clive, 1997, "The Basic Variety (or: Couldn't natural languages be much simpler?)". *Second Language Research* 13/4: 301-347.

Laufer, Batia, 1997, "What's in a word that makes it hard or easy: some intralexical factors that affect the learning of words". In: Schmitt, Norbert/McCarthy, Michael (eds.), *Vocabulary: Description, Acquisition and Pedagogy*, Cambridge, Cambridge University Press: 140-155.

Laufer, Batia/Nation, Paul, 1995, "Vocabulary Size and Use: Lexical Richness in L2 Written Production". *Applied Linguistics* 16/3: 307-322.

Meara, Paul, 1980, "Vocabulary Acquisition: A neglected aspect of language learning". *Language Teaching and Linguistics: Abstracts* 15/4: 221-246.

Menard, Nathan, 1983, *Mesure de la Richesse Lexicale*, Genève-Paris, Slatkine – Champion.

Muller, Charles, 1964, "Calcul des probabilités et calcul d'un vocabulaire". *Tra Li Li* 2/1: 235-244.

Nation, Paul, 2001, *Learning vocabulary in another language*, Cambridge, Cambridge University Press.

Orletti, Franca/Testa, Renata, 1991, "La trascrizione di un corpus di interlingua: aspetti teorici e metodologici". *Studi italiani di Linguistica Teorica e Applicata* 20/2: 243-283.

Palmberg, Rolf, 1987, "Patterns of vocabulary development in foreign-language learners". *Studies in Second Language Acquisition* 9: 201-220.

Perdue, Clive (ed.), 1993a, *Adult language acquisition: Cross-linguistic perspectives. Volume I. Field methods*, Cambridge, Cambridge University Press.

Perdue, Clive (ed.), 1993b, *Adult language acquisition: Cross-linguistic perspectives. Volume II. Results*, Cambridge, Cambridge University Press.

Ramat, Paolo, 1990, "Definizione di <parola> e sua tipologia". In: Berretta, Monica/Molinelli, Piera/Valentini, Ada (a cura di), *Parallela 4. Morfologia/Morphologie*, Tübingen, Narr: 3-15.

Singleton, David, 1999, *Exploring the Second Language Mental Lexicon*, Cambridge, Cambridge University Press.

Valentini, Ada, 2003, "Da *giardino vacanza* a *campeggio*: il ruolo delle parole composte". Comunicazione tenuta alla Giornata di Studio su *Morfologia, tipologia e acquisizione di lingue seconde* (Milano-Bicocca, 30 maggio 2003).