INDICE

0 Premessa.....5
0.1Obiettivi della ricerca.....5
0.2 Il lessico nell'insegnamento della L2.....6
0.2.1 Quanto lessico si deve insegnare.....8
0.2.3Quale lessico si deve insegnare.....9
1Dizionari di frequenza e vocabolari di base.....11
1.1Il dizionario di frequenza.....11
1.1.1Definizione e tipologia.....11
1.1.2Precedenti storici.....13
1.1.3Aree d'interesse.....14
1.1.3.1I requisiti del corpus.....14
1.1.3.1.1Rappresentatività.....16
1.1.3.1.2Estensione.....17
1.1.3.1.3Riferimento standard.....18
1.1.3.2La parola come occorrenza e come unità lessicale.....19
1.1.3.2.1La parola come unità elementare dei testi.....19
1.1.3.2.2La parola come unità lessicale elementare.....20
1.1.3.3Frequenza lessicale.....21
1.1.3.3.1Dizionari di frequenza "grezza".....22
1.1.3.3.2Frequenza modificata da criteri distribuzionali "oggettivi".....22
1.1.4La struttura statistica del lessico.....23
1.2Il vocabolario di base.....26
1.2.1Definizione.....26
1.2.2Storia della lessicografia dei dizionari fondamentali.....27
1.2.2.1Le origini: dizionari basati sulla frequenza.....27
1.2.2.2Frequenza e criteri aggiuntivi.....30
1.2.2.3L'approccio comunicativo-pragmatico.....32
1.2.2.4Un approccio frequenziale-comunicativo.....34
1.2.3Tipi di lessici.....35
1.2.4Problemi di applicazione.....37
1.3Quali e quante sono le parole fondamentali?.....38
2I dizionari fondamentali e di frequenza della lingua italiana (1927-2006).....40
2.1Autori, finalità e criteri di selezione.....40
2.2Cronologia dei lessici fondamentali e di frequenza.....41
2.2.1Thompson M. E. (1927) A Study in Italian Vocabulary Frequency, University of Iowa Master's Thesis.....42
2.2.2Knease T. M. (1931) An Italian Word List from Literary Sources, The State University of Iowa, Toronto, Toronto U. P.....42
2.2.3Skinner L. H. (1935) A Comparative Study of the Vocabulary of Forty Five Italian Textbooks. In: Modern Language Journal, XX, November, 2, pp. 67-84.....43
2.2.4Migliorini B. (1943) Der grundlegende Wortschatz der Italienischen. Die 1500 wesentlichsten Wörter, Marburg, Elwert.....43

2.2.5Russo G. A. (1947) A combined Italian Word List. In: Modern Language Journal, XXXI, April, pp. 218-240.....44
2.2.6Russo J. A. (1962) An Italian Idiom List. In: Modern language Journal, XLVI, January, pp. 14-21.....44
2.2.7Giovanelli P. (1966) Grund- und Aufbauwortschatz. Italienisch, Stuttgart, E. Klett.....44
2.2.8Bortolini U., Tagliavini C., Zampolli A. (1971) Lessico di frequenza della lingua italiana contemporanea, Milano, Garzanti-IBM. (LIF).....45
2.2.9Juilland A., Traversa V. (1973) Frequency Dictionary of Italian Words, The Hague, Mouton. (FDI).....46
2.2.10Alinei M. (a cura di) (1973) Spogli elettronici dell'italiano letterario contemporaneo, Bologna, Il Mulino. (SEILC).....47
2.2.11Reiske H. (1974) VHS-Italienisch Unterricht und VHS-Zertifikat Italienisch, Frankfurt, Pädagogische Arbeitsstelle-Deutscher Volkhochschulverband. 47
2.2.12Baldelli I., Mazzetti A. (1974) Vocabolario minimo della lingua italiana per stranieri, Firenze, Le Monnier.....48
2.2.13Sciarone A. G. (1977) Vocabolario fondamentale della lingua italiana, Bergamo, Minerva Italica.....48
2.2.14Dizionario illustrato per l'uso essenziale della lingua (1978) Torino, SEI. (D.I.L.).....49
2.2.15De Mauro T. et alii (1980), Il vocabolario di base della lingua italiana. In: De Mauro T., Guida all'uso delle parole, Roma, Editori Riuniti, pp. 146-172. (VdB).....50
2.2.16Gabrielli A. (a cura di) (1981) Il mio primo dizionario illustrato, Milano, Mondadori.....52
2.2.17Galli de' Paratesi N. (1981) Livello soglia per l'insegnamento dell'italiano come lingua straniera. Strasburgo, Consiglio d'Europa.....52
2.2.18Batinti A., Trenta W. (1982) Ricerche sul lessico di base dell'italiano contemporaneo, Perugia, Guerra.....54
2.2.19Dizionario illustrato della lingua italiana per la scuola dell'obbligo (1986) Milano, Editrice Piccoli/Mondadori.....55
2.2.20Vocabolario elettronico della lingua italiana. Il vocabolario del 2000 (1989) Milano, IBM Italia. (VELI).....55
2.2.21Cannella M. (a cura di) (1991) Il primo Zanichelli. Vocabolario elementare di italiano, Bologna, Zanichelli.....57
2.2.22Katerinov K., Boriosi Katerinov M. C., Sciarone A. G. (1991) Calcolo di rendimento di liste base: italiano parlato, italiano scritto, livello soglia, Perugia, Guerra. (KBS).....57
2.2.23De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993) Lessico di frequenza dell'italiano parlato, Milano, Etaslibri. (LIP).....58
2.2.24Marconi L., Ott M., Pesenti E., Ratti D., Tavella M. (1994) Lessico elementare. Dati statistici sull'italiano letto e scritto dai bambini delle elementari, Bologna, Zanichelli. (LE).....59
2.2.25Bertinetto P. M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., Thornton A. (1995) Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS).....61
2.2.26Savino E. (a cura di) (1995) Il mio primo vocabolario, Milano, Mursia.....62
2.2.27De Mauro T., Moroni G. (1996) Dizionario di base della lingua italiana, Torino, Paravia. (DIB).....62

2.2.28De Mauro T., Moroni G. (1997b) Prime parole. Dizionario illustrato di base della lingua italiana, Torino, Paravia....63
2.2.29De Mauro T., Moroni G. (1997a) DAIC. Dizionario avanzato dell'italiano corrente, Torino, Paravia....63
2.2.30Putzer O. (a cura di) (1998a) Bilinguismo. Zweisprachigkeit. Lessico di base. Grundwortschatz, Bolzano, Arkadia. Putzer O. (a cura di) (1998b) Lessico approfondito. Aufbauwortschatz, Bolzano, Arkadia....64
2.2.31Cresti E., Moneglia M. (a cura di) (2004) Campionamento Corpus Stammerjohann....64
2.2.32Rinaldi P., Barca L., Burani C. (2004) Caratteristiche semantiche, grammaticali e di frequenza delle parole del "Primo Vocabolario del Bambino." In: Psicologia clinica dello sviluppo, 8, pp. 119-143....65
2.2.33Baldelli I., Mazzetti A., Falcinelli M., Servadio B. (2005) Dizionario fondamentale della lingua italiana per stranieri, Firenze, Le Monnier....66
2.3Tabella riassuntiva e comparativa....66
2.4La parola come segno linguistico nei lessici di frequenza e di base....68
2.5Riassunto, prospettive....70
3I dizionari fondamentali e di frequenza della lingua ungherese (1933-2006)....74
3.1Autori, finalità e criteri di selezione....74
3.2Cronologia dei lessici fondamentali e di frequenza dell'ungherese....75
3.2.1Nemes Z. (1933) A magyar parlamenti nyelv leggyakoribb szavai. Az Egységes Magyar Gyorsírás Könyvtára 66., Szeged....75
3.2.2Cser J. (1939) A magyar gyermek szókincse. Gyakorisági és korszótár, Budapest, Magyar Pedagógiai Társaság....76
3.2.3Nemes Z. (1941) Szóstatisztika egymillió szótagot felölelő újságszövegek alapján. Az Egységes Magyar Gyorsírás Könyvtára 190., Szeged....77
3.2.4Csirikné Czachesz E., Csirik J. (1986) Újságnyelvi gyakorisági szótár. Szeged, Juhász Gyula Tanárképző Főiskola....78
3.2.5Füredi M., Kelemen J. (1989) A mai magyar nyelv szépprózai gyakorisági szótára (1965-1977), Budapest, Akadémiai Kiadó....80
3.2.6Aradi A., Erdős J., Sturcz Z. (2000) Küszöbszint, Magyar mint idegen nyelv. Strasbourg, Európa Tanács....81
3.2.7Csirikné Czachesz E., Csirik J. (2002) 10-16 éves tanulók írásbeli szókincsének gyakorisági szótára, Budapest, BIP....83
3.2.8Bácsi J., Kerekes J. (2003) Az első osztályos olvasókönyvek szóanyagából készült gyakorisági szótár: "Van szó". In: Módszertani közlemények, 43, 2, pp. 53-58....84
3.2.9Dizionario di frequenza WordSword (2004)....84
3.2Tabella riassuntiva e comparativa....86
3.3Riassunto, prospettive....86
4 Alcune applicazioni dei lessici di frequenza e di base....88
4.1Lessicografia....88
4.1.1La selezione dei lemmi....89
4.1.2La definizione lessicografica....92
4.1.3Dati di frequenza e marche d'uso....96
4.1.4Accezioni, collocazioni ed espressioni polirematiche....102
4.2Stilometria e scrittura di testi di difficoltà controllata....106
4.2.2Il mensile due parole....108
4.2.3Codice di stile....110
4.3Applicazioni informatiche....111

4.3.2Trattamento automatico del linguaggio....112
4.3.3Tecnologie della lingua parlata....116
5Il lessico nei sistemi di certificazione....118
5.1Gli apporti del Consiglio d'Europa....118
5.1.1Livello soglia....118
5.1.2Il Quadro comune europeo per le lingue....121
5.1.2.1La competenza lessicale nel Quadro comune....124
5.2Certificati nazionali di italiano L2....126
5.2.1Certificazione di Italiano come Lingua Straniera (CILS)....126
5.2.2Certificato di Conoscenza della Lingua Italiana (CELI)....128
5.2.3Certificato di Competenza della Lingua (PLIDA)....131
6Il lessico del manuale Progetto italiano 1....133
6.1Il corpus di riferimento....134
6.2La realizzazione....134
6.2.1Reperimento dei dati....135
6.2.2Lemmatizzazione ed organizzazione dei dati....137
6.3I risultati....139
6.3.1Caratteri quantitativi del lessico....140
6.3.2Caratteri qualitativi del lessico....145
6.3.2.1Sostantivi....146
6.3.2.2Verbi....148
6.3.2.3Aggettivi....149
6.3.2.4Avverbi....151
6.3.2.5Esotismi....151
6.3.2.6Nomi propri....152
6.3.2.7Interiezioni e fonosimboli....153
6.3.2.8Polirematiche....154
6.3.2.9Conclusioni....155
6.4Comparazione del glossario con il VdB....156
6.5Grado di copertura del lessico....161
6.6Apporti di un dizionario di base ungherese....164
6.7Suggerimenti per gli autori dei libri di testo....168
7Conclusioni....171
Bibliografia....173
Appendice....184
Lista A184
Lista B230
Lista C230

0 Premessa

L'insegnamento dell'italiano come lingua straniera in Ungheria ha lunghe tradizioni. Attualmente s'insegna italiano in più di 250 scuole primarie (elementari e medie secondo l'ordinamento italiano) e secondarie, a più di 15.000 studenti.[1] Esistono due licei bilingui, il Kodály Zoltán di Pécs e il Szent László di Budapest in cui si può sostenere l'esame di maturità italiano. È inoltre in funzione una sezione privata elementare-media italiana presso una scuola di Budapest. Concentrandoci primariamente sull'istruzione pubblica, non si enumerano le varie scuole di lingua, sparse su tutto il territorio nazionale, in cui, fra l'altro, si insegna la lingua italiana. L'attività didattica dell'italiano come lingua straniera, quindi, è e continua ad essere molto intensa.
Con la liberalizzazione dell'istruzione pubblica nel 1989, è cambiata la prassi di adottare in tutte le scuole primarie e secondarie gli stessi libri didattici, previsti dal Ministero dell'Istruzione Pubblica. D'allora in poi, spetta agli insegnanti di scegliere dall'offerta degli editori ungheresi ed internazionali il libro didattico da usare. È una scelta che andrebbe operata in vista di una serie di criteri: da un lato occorre tener presente il livello di conoscenza, le capacità, gli obiettivi dei discenti; dall'altro bisogna avere indicazioni precise sulle competenze linguistiche che gli allievi alla fine del periodo di apprendimento dovrebbero padroneggiare. Mentre il primo gruppo di criteri richiede in ogni caso una valutazione ad hoc da parte del docente, la fissazione degli obiettivi didattici nei vari livelli di conoscenza è una questione che va affrontata in modo sistematico. La mancanza di richieste standardizzate, infatti, causa diversi problemi, non solo nelle scuole elementari e medie, ma soprattutto a livello universitario, giacché il grado di conoscenza degli studenti che hanno imparato l'italiano in diverse scuole superiori o licei del Paese, può essere, e spesso è molto differente. La standardizzazione dei requisiti è anche al centro delle varie certificazioni di conoscenza delle lingue straniere, tuttavia a livello nazionale non è ancora avvenuta né l'armonizzazione dei requisiti dei vari tipi d'esame di lingua, né la normalizzazione degli obiettivi didattici dell'insegnamento delle lingue straniere nelle scuole primarie e secondarie.

0.1 Obiettivi della ricerca

Questa tesi si propone di offrire un aiuto teorico e pratico agli insegnanti di lingua, ma anche a coloro che si occupano di valutazione e di certificazione di conoscenza di lingue straniere. Ambedue i gruppi si trovano in una situazione difficile, poiché non esiste un repertorio (né lessicale, né grammaticale) comunemente accettato che definisca con sufficiente chiarezza, di quali elementi si può pretendere la conoscenza ad un determinato livello. La tesi mira a soddisfare esigenze sussistenti, fornendo delle proposte concrete per stabilire un lessico fondamentale della lingua italiana, che si pone come livello lessicale minimo a livello elementare, ossia come livello di "entrata" per ottenere un certificato di conoscenza di livello base (A2) della lingua italiana. Si chiariscono i criteri e i metodi della selezione del materiale lessicale, e tramite l'analisi

[1] Dati ricavati dal sito dell'Istituto Italiano di Cultura di Budapest: www.iicbudapest.esteri.it.

del lessico di uno dei manuali attualmente più diffusi nell'insegnamento dell'italiano, si forniscono proposte concrete in merito al materiale da inserire nella progettazione e/o valutazione. Ci occuperemo esclusivamente del problema della selezione del materiale linguistico, che precede tutte le altre attività didattiche, e che ha la funzione di identificare quel nucleo di conoscenze che funge da punto di partenza e di riferimento costante durante il processo d'apprendimento e/o di valutazione.
La tesi si articola nel modo seguente. Questa parte introduttiva ha lo scopo di definire gli obiettivi e i contenuti della ricerca. Nel capitolo 1 si passa in rassegna l'evoluzione dei dizionari fondamentali e di base, due strumenti indispensabili che tutt'oggi influenzano in modo determinante la selezione del vocabolario fondamentale. Scopo di questo primo capitolo è di situare la ricerca nel contesto bibliografico internazionale e di discutere alcune questioni fondamentali. Seguono i capitoli 2 e 3, dedicati alla descrizione dei dizionari fondamentali e di frequenza di cui disponiamo per l'italiano e per l'ungherese. Si confrontano le loro caratteristiche principali (destinatari, corpus e metodi di selezione, carico degli esponenti ecc.), al fine di reperire quei lavori che possono fungere da punti di riferimento per la costruzione di un vocabolario fondamentale per l'italiano L2.[2] Nel capitolo 4 si fa un piccolo excursus rispetto all'obiettivo originario del lavoro: si presentano alcune applicazioni linguistiche e informatiche dei lessici di frequenza e di base, con particolare riguardo ai risvolti lessicografici. Nel capitolo 5 si descrivono le iniziative prese a livello europeo in merito alla definizione dei livelli di conoscenza delle lingue straniere; la seconda parte è una presentazione dei requisiti lessicali previsti da alcuni sistemi di certificazione di competenza in italiano L2. Nel capitolo 6 si discutono i dati empirici emersi dall'analisi del glossario di uno dei libri didattici più diffusi nell'insegnamento dell'italiano L2 in Ungheria. Il *Progetto italiano 1* di T. Marin e S. Magnelli (2006) è uno dei pochi libri didattici che dichiaratamente si orientano al *Quadro comune europeo di riferimento per le lingue* e alle tipologie delle certificazioni italiane (cfr. sez. 5.2). Si esaminano le caratteristiche quantitative e qualitative del lessico insegnato, successivamente si confronta il materiale lessicale proposto con due liste di riferimento per l'italiano: il *Vocabolario di base della lingua italiana* di De Mauro *et al.* (1980) e il *Lessico di frequenza dell'italiano parlato* (De Mauro *et al.* 1993) e il vocabolario di base della lingua ungherese (Aradi *et al.* 2000). Si verifica il grado di copertura del lessico proposto rispetto ad alcuni esercizi della prova di lettura dell'esame di certificazione CELI. Il capitolo termina con delle proposte indirizzate agli autori dei manuali didattici, in merito al lessico da includere e da escludere dai libri di testo. L'ultimo capitolo riassume le conclusioni relative ai criteri di selezione di un vocabolario fondamentale, pensato per l'insegnamento dell'italiano L2. Il glossario allegato contiene il lessico incluso nel manuale *Progetto italiano 1*, integrato con delle indicazioni emerse dal confronto con le suddette opere.

0.2 Il lessico nell'insegnamento della L2

Nello sviluppo della competenza comunicativa in una L2 quale importanza deve avere il fattore lessicale? Si può intuire che il lessico ha un ruolo non trascurabile nell'acquisizione di una L2. Basta pensare al fatto che chi va in un Paese di cui non conosce la lingua, per tentare di sopravvivere linguisticamente, si procura un vocabolario di quella lingua, non una grammatica.

[2] Per L2 (lingua due) si intende una lingua che non è la lingua madre (Ciliberti 1994: 214).

Giacalone Ramat (1993: 347) in uno studio sull'acquisizione spontanea dell'italiano L2 individua nel processo d'apprendimento tre fasi. Le varietà basiche comprendono i primi tentativi dei parlanti di farsi capire e consistono prevalentemente di elementi lessicali e di pochi elementi funzionali, e non sono presenti opposizioni morfologiche di nessun tipo. La seconda fase è caratterizzata dall'aumento degli elementi lessicali, mentre le variazioni morfologiche restano fenomeni casuali. L'acquisizione delle strutture della lingua d'arrivo avviene nelle fasi successive, in cui l'apprendente acquisisce le classi di parole insieme alle loro proprietà morfologiche. Il fattore lessicale ha un ruolo decisivo nelle fasi iniziali dell'apprendimento, giacché la comunicazione in questo primo periodo è affidata esclusivamente alla carica semantica degli elementi lessicali. In realtà, anche se il rapporto tra i due tipi di apprendimento (quello spontaneo e quello guidato) non è stato ancora chiarito, le ricerche hanno dimostrato che "anche gli apprendenti guidati, quando devono usare la lingua in conversazioni naturali e non in esercizi scolastici, si comportano in maniera molto simile agli apprendenti stranieri" (ivi, p. 342). Inoltre, Bettoni (2001: 61) osserva che perfino nei livelli più avanzati, gli errori lessicali sono, rispetto a quelli grammaticali, notevolmente più comuni, sono quelli che i parlanti nativi notano di più, infine sono quelli che recano più danno ai fini comunicativi. Pertanto il lessico è della massima importanza, in tutti gli stadi dell'apprendimento.

Nel dibattito su metodi e approcci di insegnamento delle lingue straniere esistono tendenze che mirano ad accentuare la portata del lessico. Il c.d. "approccio lessicale" è emerso agli inizi degli anni Novanta. Esso parte dall'idea che si possa arrivare alla competenza nella lingua straniera attraverso un insegnamento basato sul lessico, e in particolare che la stessa grammatica vada appresa contemporaneamente e insieme al lessico, poiché essa è integrata nel lessico.[3] In una visione tradizionale l'apprendimento del lessico riguarda l'acquisizione delle parole, cioè del "vocabolario". Vocabolario e grammatica sono due entità separate: da una parte si studiano le regole della grammatica, dall'altra si studiano i vocaboli. La lingua si identifica prevalentemente con la struttura grammaticale piuttosto che con il lessico, considerato come un insieme di "pacchetti di significato". L'approccio lessicale contesta questa concezione, in quanto lo studio della grammatica va integrato in quello del lessico, poiché il lessico è esso stesso una componente portatrice di informazione grammaticale. Le parole non si combinano liberamente fra loro, ma seguono alcune preferenze di selezione. Le collocazioni sono specifiche di ogni vocabolo e permettono di prevedere il significato in base alla compatibilità fra due parole. Certe parole si presentano con frequenze stabili accanto ad altre, consentendo di individuare anche nel lessico aspetti sistematici che non sono regole vere e proprie, ma almeno tendenze.

Un contributo pratico lo ha fornito il progetto COBUILD, basato sull'analisi di frequenza e di contesto di un corpus dell'inglese di ca. 20 milioni di occorrenze. Il progetto ha dato luogo a un dizionario, una grammatica e un corso di lingua. La ricerca ha fornito non poche sorprese anche dal punto di vista della verifica delle regole grammaticali generalmente diffuse nei materiali didattici. Sembra, infatti, che ci sia una correlazione tra la frequenza delle parole e la frequenza delle strutture grammaticali in cui esse occorrono (Willis 1990: VI), pertanto concentrandosi sulle parole più frequenti, si possono illustrare tutti gli aspetti principali della grammatica di base. Lessico e grammatica non vanno appresi separatamente, visto che l'identificazione corretta delle unità lessicali comporta anche un'elaborazione dei nessi sintattici ed altri aspetti trattati tradizionalmente dalla grammatica.

L'importanza del lessico è riconosciuta oggi sempre di più, eppure gli studi sul lessico sono relativamente pochi. Ciò si deve principalmente a un motivo fondamentale: il

[3] Sull'approccio lessicale si vedano Serra Borneto (1998) e Bettoni (2001).

lessico, assai più che la grammatica, è un sistema aperto, difficile da ricondurre a regole, pieno di potenzialità creative.[4] Per questo a prima vista appare più facile esaminare la grammatica, che per sua natura tende ad essere più sistematizzabile. Non sorprende quindi, che mentre per la fonologia e per la sintassi la didattica delle lingue straniere ha ormai addottato i risultati degli studi in materia, l'insegnamento del lessico continua a seguire più il buon senso che le statistiche derivanti da corpora. Tuttavia, nonostante l'evidente difficoltà di ricondurre il lessico a generalizzazioni, si stanno facendo passi avanti nella direzione di una maggiore consapevolezza del suo funzionamento, basta pensare ai lavori di analisi di vasti corpora in tutte le principali lingue del mondo. È stato finalmente riconosciuto che proprio perché il lessico ha un'organizzazione interna meno sistematica, e perciò nell'apprendimento presenta minori regolarità di sviluppo, le considerazioni esterne (i criteri formali, semantici e di contrastività tra L1 e L2) assumono nel caso del lessico un'importanza maggiore che non nel caso della grammatica.[5]

0.2.1 Quanto lessico si deve insegnare

Quando possiamo dire di conoscere il vocabolario di una lingua? Il vocabolario di ogni lingua comprende un numero considerevole di parole. Ecco alcuni esempi, per limitarci ai testi più noti.

Dizionario della lingua italiana (Devoto-Oli 2004)	100.000
Dizionario della lingua italiana per il terzo millennio (De Mauro 2000)	160.000
Grande dizionario della lingua italiana (Battaglia 1961)	210.000
Grande dizionario italiano dell'uso (De Mauro 1999)	250.000 + 100.000 espressioni polirematiche
Dizionario italiano Sabatini Colletti (Sabatini-Coletti 1997)	153.500
Vocabolario della lingua italiana (Zingarelli 2006)	136.000
Zingarelli minore (2001)	56.000

Quante parole ci sono nei vocabolari?

Numeri del genere sono però solo prime approssimazioni alla numerosità del lessico: la quantità di parole che ciascun parlante riconosce e sa usare varia da individuo a individuo, secondo il livello d'istruzione, le conoscenze, le esperienze, gli ambiti professionali praticati, e perfino le conoscenze lessicali individuali cambiano nel tempo. Valutazioni intuitive sostengono che il lessico di una lingua di cultura comprenda ca. 200.000 parole e insieme ai termini dei linguaggi speciali si arriva facilmente a 500.000, di cui il parlante medio conosce ca. il 10% (Merzagora 1983: 34). A numeri simili giungono Corda e Marello (1994: 29): dei circa 100.000 lemmi registrati dai dizionari d'uso un italiano conoscerebbe circa il 60-80%. Rispetto a quello ricettivo, il vocabolario produttivo è di dimensioni molto inferiori. Le parole conosciute ricettivamente e produttivamente dalla maggior parte dei parlanti, sarebbero, secondo De Mauro, tra le 30.000 e le 50.000. Queste voci appartengono al c.d. *vocabolario comune* che

[4] Halliday (1964: 207) parla di sistemi chiusi (*closed systems*) a livello grammaticale e di serie aperte (*open sets*) a livello di lessico. Martinet (1966: 17-19) a proposito delle unità significative della lingua, ha distinto i morfemi limitati di numero dai lessemi che sono invece di numero illimitato.

[5] Notiamo, tuttavia, che criteri esterni, come la frequenza, sono stati introdotti anche nell'insegnamento della fonetica e per la sintassi (De Mauro 1979: 30-33, Katerinov 1991: 6).

"parecchie persone di parecchie categorie abbastanza diverse tra loro possono capire e perfino usare in qualunque discorso, con un interlocutore di qualunque categoria professionale o regione" (De Mauro 1980: 85).

In realtà, per esprimersi in modo chiaro, non è necessario un grandissimo numero di parole. Secondo Verlée (1954) le persone "semplici" non usano, oltre una quantità variabile di termini tecnici, propri del loro ambiente, più di 2.000 parole diverse, mentre gli intellettuali usano correntemente 4.000-5.000 parole.[6]
Dato che è impossibile dominare una lingua in qualche anno di studio in poche ore la settimana, gli insegnanti si sono resi conto della possibilità di cominciare l'insegnamento del lessico di una lingua da un nucleo ristretto di parole, che permette al discente un livello minimo di comprensione di qualsiasi testo. Tuttavia, data l'assenza di cifre affidabili sull'estensione del lessico produttivo e ricettivo, è difficile stabilire quante parole dovrebbe conoscere un parlante straniero per capire un testo scritto o parlato. Gli studi in materia segnalano solo che l'acquisizione procede molto lentamente: gli allievi che studiano una lingua alle scuole superiori, hanno alla fine un vocabolario ricettivo di 3.000 parole e un vocabolario produttivo di circa 1.500 (Marello 1994: 29).
Quante parole bisogna imparare per comunicare in un livello elementare? La gran parte degli autori (ad es. Kühn 1979, Sciarone 1977, Bettoni 2001) parla di ca. 2-3.000 vocaboli della L2, che l'apprendente deve conoscere e saper usare, insieme alle loro caratteristiche formali, sintattiche e semantiche.

0.2.3 Quale lessico si deve insegnare

Le parole del vocabolario ricettivo e produttivo del parlante straniero dovrebbero rientrare il più possibile nel "vocabolario di base" dell'italiano, a quel nucleo di parole cioè, che sono comprensibili per la maggioranza della popolazione. Tuttavia, per stabilire quali parole deve apprendere uno studente straniero, occorre chiarire a quale scopo sta imparando l'italiano. L'obiettivo che ci si prefigge determina la scelta tra lingua scritta o parlata, il livello di specializzazione nei linguaggi settoriali, ecc. Nella maggioranza dei casi, lo scopo per cui si segue un corso di lingua, è acquisire una competenza globale della lingua straniera, per essere in grado di capire testi scritti e parlati, comunicare per iscritto e oralmente nella lingua straniera.
Se accettiamo che devono essere posti dei limiti all'apprendimento del lessico, sarà opportuno scegliere i vocaboli più utili per il discente, quelli cioè che potranno essere usate in un gran numero di circostanze. Per un ristretto numero di parole – fra 500 e 1.000 – è effettivamente possibile dare un giudizio d'importanza, il problema della selezione di un vocabolario di base affiora invece per le prime migliaia di parole, per le quali non possiamo esprimere un giudizio d'importanza a prima vista.[7] Oltre le prime 500 parole più frequenti si possono avere delle sorprese, appunto per questo basarsi soltanto sulla propria impressione di parlante nativo o di parlante straniero esperto non basta. Vale la pena avvalersi anche dei calcoli di frequenza o delle indicazioni contenute in alcuni dizionari.

[6] Guiraud (1954) ha dimostrato che Baudelaire, pur disponendo di un vocabolario di ca. 25.000 parole, ne ha usato solo 4.000 ne *Les fleurs du mal*. Egli è giunto alla conclusione che la gran parte degli scrittori non usa più di 4.000 parole, però ciò non gli impedisce di affrontare argomenti complessi.
[7] Sciarone (1995: 33) nota che mentre le prime 500 parole delle liste basate sulla sola frequenza (ad es. Sciarone 1977) coincidono tendenzialmente con le prime 500 parole delle liste basate sull'esperienza degli autori (ad es. Reiske 1974, Baldelli e Mazzetti 1974), con il diminuire della frequenza tale correlazione diminuisce progressivamente.

Nel prossimo capitolo presentiamo la storia e le problematiche relative alla costituzione e all'utilità dei dizionari di frequenza e di base.

1 Dizionari di frequenza e vocabolari di base

I dizionari di frequenza sono il naturale antecedente, logico e cronologico, dell'elaborazione dei vocabolari fondamentali. I vocabolari fondamentali, infatti, si identificano spesso con la zona di massima frequenza d'uso delle parole di una lingua. In questo capitolo si dà un breve panorama della tipologia, dello sviluppo e delle principali aree d'interesse di questi due tipi di prodotti lessicografici, infine si passano in rassegna i problemi relativi alla loro applicazione.

1.1 Il dizionario di frequenza

1.1.1 Definizione e tipologia

> "I lessici di frequenza sono delle liste lemmatizzate organizzate in ordine di frequenza decrescente. L'utilità dei lessici di frequenza basati su grandi corpora di riferimento risiede nella possibilità di fornire un quadro delle principali fasce d'uso dei lessemi e della loro copertura testuale."

In conformità alla definizione di Chiari (2007: 75), non esamineremo liste di frequenza di fonemi, morfemi e strutture grammaticali anche se tali lavori a volte si possono chiamare vocabolari.
Dal punto di vista prototipico, gli elementi macrostrutturali di un dizionario di frequenza sono le parole (lemmi e/o forme), mentre la sua microstruttura informa sugli aspetti frequenziali delle parole tramite dati di frequenza ed altri valori. Per frequenza si intende il numero di volte che una parola o altro fenomeno linguistico occorre in un testo. La frequenza (F) può essere assoluta e si esprime con il numero di volte che il fenomeno occorre, oppure relativa, ossia riportata al totale delle parole di un testo (N) (Beccaria 2004: 343). La frequenza relativa (F/N) si esprime in decimale o in percentuale (quando il valore della frequenza relativa è moltiplicato per 100).
Un estratto da un tipico dizionario di frequenza potrebbe essere il seguente.

	USAGE		FREQUENCY		DISPERSION	
			FIRST FIVE HUNDRED			
1	di *prep*.	24798.22	di *prep*.	28659	dimenticato *adj*.	100.000
2	il *a*.	21375.36	il *a*.	24333	riconoscibile *adj*.	100.000
3	la *a*.	21100.01	la *a*.	23663	rogo *n*.	100.000
4	e *conj*.	13329.33	e *conj*.	14658	vacanza *n*.	100.000
5	a *prep*.	10316.54	in *prep*.	11425	zelante *adj*.	100.000
6	in *prep*.	8727.88	a *prep*.	10901	quattro *adj*.	95.60
7	uno *a*.	8434.71	uno *a*.	93334	anzi *adv*.	95.30
8	essere *v*.	6623.04	essere *v*.	8194	peso n.	95.24

Estratto da Juilland (1973: 405).

Nella prima colonna è riportato il rango di ogni lemma, ovvero il posto occupato da ciascun elemento nelle tre liste di uso, frequenza e dispersione decrescente, seguono poi

in tre colonne i lemmi stessi con indicazione della categoria grammaticale e dei valori d'uso, frequenza e dispersione, rispettivamente. Accanto ai dati di frequenza assoluta appaiono quindi dati di distribuzione e/o di frequenza derivati/modificati, come il coefficiente d'uso o la dispersione, come vedremo più avanti nella sez. 1.1.3.3.
La seguente lista, da trovare sempre in Juilland, ha una microstruttura più dettagliata.

abile *adj.*		**4.13**		**6**		**68.82**	
		6	*2*	*1*	*2*	*1*	*0*
	abile	3	1	0	2	0	0
	abili	3	1	1	0	1	0
abilità *n.*		**2.64**		**7**		**37.73**	
	abilità	7	3	0	4	0	0
abisso *n.*		**7.80**		**14**		**55.68**	
		1 4	*1*	*2*	*7*	*4*	*0*
	abissi	5	0	0	3	2	0
	abisso	9	1	2	4	2	0

Estratto da Juilland (1973: 4).

Nella tabella si trovano le seguenti informazioni: la prima riga riporta in grassetto il lemma seguito dalla categoria lessicale, mentre le cifre rappresentano il coefficiente d'uso, di frequenza e dispersione, rispettivamente. La prima cifra della seconda riga è il numero totale di occorrenze, segue poi il numero delle occorrenze in ogni genere di testo (Juilland 1973) contiene cinque strati con 100.000 parole ciascuno). Le eventuali varianti morfologiche sono menzionate nelle righe seguenti, corredate delle stesse informazioni della seconda riga.
Le differenze di macrostruttura risultano più di tutto dalla scelta lessicale effettuata e dal grado di esaustività che il dizionario di prefigge: mentre le liste di frequenza generali, come il *Lessico di frequenza della lingua italiana* di Bortolini *et al.* (1971), sono nate per fornire dati statistici sull'insieme del lessico di una lingua; altri dizionari sono dedicati a lingue speciali o settoriali: il lavoro di Marconi *et al.* (1994), ad esempio, analizza il lessico dei bambini delle elementari. Si pensi infine alle concordanze letterarie dedicate ad un singolo autore.[1] A causa della variazione delle lingue naturali, i domini più importanti di tale variazione, quello geografico, sociale, individuale, temporale e di contenuto, si ritrovano come restrizioni nella macrostruttura dei dizionari di frequenza.
Strettamente legato ai suddetti è l'ampiezza del corpus che serve come base empirica per il dizionario (cfr. 1.1.3.1): di norma più largo è il corpus, più elementi si trovano nella macrostruttura, ma anche qui si può effettuare una selezione. Juilland (1973), ad esempio, riporta i lemmi con un coefficiente d'uso superiore a 3, ottenendo così 5.014 entrate.
Altre differenze di macrostruttura riguardano la forma e la disposizione delle entrate. Tendenzialmente si distinguono tre soluzioni: alcuni dizionari riportano solo i lemmi, ma non le rispettive forme, ad es. Sciarone (1977), altri invece, elencano solo le forme trovate nel corpus (*WordSword* 2004). Tutte e due le possibilità hanno vantaggi e svantaggi. Dando solamente i lemmi, si priva il lettore di una serie di informazioni grammaticali: egli non verrà a sapere quali forme (declinate o coniugate) delle parole in oggetto sono più frequenti e come sono distribuiti. Elencando invece la sola frequenza delle forme, il lettore curioso della frequenza di un determinato lemma, dovrà effettuare ogni volta i conteggi necessari. La soluzione più felice è di riportare, sotto ogni lemma anche le rispettive forme come in Bortolini *et al.* (1971), De Mauro *et al.* (1993) e altri.

[1] Una concordanza è la presentazione delle parole di un testo con l'indicazione della frequenza con la quale la parola occorre e il contesto linguistico precedente e successivo (Chiari 2007: 72). Le concordanze permettono di osservare i diversi usi di una parola, ed esaminare i contesti in cui appare.

Quanto alla disposizione delle entrate, esse possono essere sistemate in ordine alfabetico, alfabetico inverso, di frequenza decrescente, per lunghezza decrescente, ecc. In Bortolini *et al.* (1971), ad esempio, i lemmi e le rispettive forme sono ordinati in ordine alfabetico (pp. 97-716), mentre i soli lemmi sono riportati in ordine d'uso decrescente (pp. 717-852). Pur potendo ordinare una lista di frequenza in diversi modi, l'ordinamento più utile resta quello per frequenza/uso decrescente, conveniente soprattutto per individuare la distribuzione delle fasce di frequenza/uso.
La presentazione macrostrutturale può includere non solo varianti morfologiche, ma anche varianti semantiche. Ad esempio West (1953) nel suo *A General Service List of English Words* accanto ai "classici" dati di frequenza dà anche la percentuale d'occorrenza delle singole accezioni delle parole polisemiche.
Gli esempi di sopra dimostrano che le differenze tra i dizionari di frequenza possono essere ridotte a differenze della macro- e microstruttura.

1.1.2 Precedenti storici

La storia dei dizionari di frequenza è strettamente legata alla considerazione quantitativa del linguaggio. L'approccio quantitativo risale agli antichi Greci e Romani. L'opposizione fra voci rare o *hapax* da un lato, e parole di frequenza elevata dall'altro, permetteva ai filologi l'attribuzione dei testi ad un autore o ad un altro, anche se in questo caso la frequenza fu un fattore assai soggettivo, giacché basata sull'esperienza e sul senso linguistico dei singoli studiosi. Le prime concordanze nascono nel medioevo per i testi biblici: per ogni vocabolo sono stati riportati tutti i contesti in cui esso appariva. Successivamente appaiono concordanze di autori classici e moderni, lessici speciali di opere particolari di determinati autori. Il presupposto di fondo di tali lavori è che lo stile di uno scrittore si caratterizzi essenzialmente come deviazione quantificabile in rapporto a una norma linguistica standard e si manifesta nell'uso elevato di alcuni termini chiave, la cui frequenza nel corpus oggetto presenta uno scarto specifico rispetto ai normali contesti d'uso (Beccaria 2004: 469). Più tardi si riconosce che lo scarto rispetto alla frequenza standard è solo una delle proprietà che contraddistinguono il linguaggio letterario.
Mentre all'inizio i dizionari di frequenza erano strumenti pensati primariamente per scopi linguistici e/o letterari, nell'era moderna si compilano liste di frequenza per scopi non linguistici. Il primo approccio scientifico al problema della frequenza lessicale è stato adottato da William Gamble (1830-1886). L'opera di Gamble (*Two lists of selected characters containing all in the Bible and twenty-seven other books,* Shanghai, 1861) è un conteggio di frequenza degli idiogrammi cinesi, concepito come un aiuto per il compositore che intendeva migliorare i metodi di stampa. Il secondo e il più vasto studio di tipo quantitativo (11.000.000 occorrenze) si deve a W. Käding (*Häufigkeitswörtebuch der deutschen Sprache,* Berlin, 1899). Trattandosi di una lista compilata per l'ottimizzazione dei sistemi stenografici, l'autore fece a meno di una procedura di lemmatizzazione: la lista contiene solo le forme riscontrate nel corpus, senza i rispettivi lemmi. Oltre le 5.000 forme più frequenti, nel dizionario troviamo anche i grafemi e le sillabe più diffuse. Il materiale lessicale elencato deriva dallo spoglio di dibattiti parlamentari, testi amministrativi e commerciali, giornali, libri di storia, classici della letteratura tedesca e straniera (traduzioni), documenti militari. Käding dimostrò, che le prime 15 parole più frequenti rappresentavano il 25% delle occorrenze totali, e le prime 66 voci coprono circa il 50% dei testi (Arcaini 1978: 243). Negli anni '70 la sua lista fu rielaborata dall'Istituto Goethe per scopi glottodidattici (cfr. Schumacher 1998).

Pionieristico fu anche il lavoro di J. B. Estoup (*Gammes sténographiques,* Paris, 1907) che stabilì uno dei principi fondamentali della statistica linguistica, verificato più tardi anche dal matematico Zipf (1949), ossia che la relazione fra il numero di occorrenze di un termine e il suo rango nella serie delle parole ordinate per frequenza decrescente è costante, in altre parole rango e frequenza sono inversamente proporzionali (cfr. sez. 1.1.4).
Con il crescente interesse verso l'insegnamento delle lingue straniere, è maturata l'importanza dei dizionari di frequenza. Keniston (1920) è stato il primo ad applicare i risultati delle statistiche lessicali nell'insegnamento di una lingua straniera, in questo caso lo spagnolo, e ben presto il suo esempio è stato seguito da altri autori per le altre lingue.
La vera svolta, dal punto di vista storico, per l'elaborazione dei dizionari di frequenza fu la disponibilità dei calcolatori elettronici per automatizzare una buona parte del processo di spoglio e analisi del materiale testuale. In Italia fra gli anni '50 e '60 pionieristico fu il lavoro di Padre Busa, con gli spogli elettronici dell'intero corpus di scritti di S. Tommaso d'Aquino, per un totale di 10 milioni di occorrenze, presso il Centro per l'Automazione dell'Analisi Linguistica di Gallarate, creato principalmente con finanziamenti dell'IBM Italia.
Con l'avvento dell'informatica e grazie allo sviluppo della linguistica durante gli ultimi decenni, anche i dizionari di frequenza hanno subito dei cambiamenti. I calcolatori hanno facilitato la creazione e l'esplorazione di larghi corpora, e ciò ha profondamente influenzato la base empirica sottostante dei lavori.[2] Non solo la linguistica computazionale, ma quella teorica ha dato nuovi impatti agli studi di frequenza lessicale in generale e ai dizionari di frequenza in particolare, poiché ha incoraggiato gli esperti a trovare soluzioni migliori per i problemi di base (la questione del corpus e lo status della frequenza lessicale). Nei decenni passati è cresciuta l'attenzione dedicata alla costruzione del corpus, sono stati pubblicati più dizionari di frequenza di linguaggi settoriali che mai e sono stati elaborati nuovi approcci alla frequenza lessicale. I dati di frequenza sono stati inclusi nei database lessicali e nei dizionari: tutto ciò può essere considerato come conseguenza (diretta o indiretta) della scena linguistica in fase di trasformazione. Grazie al crescente interesse dei linguisti per il lessico e per il ruolo che esso può avere nel trattamento informatico delle lingue naturali, si prevede che questa tendenza continuerà e nel prossimo futuro porterà altri progressi.

1.1.3 Aree d'interesse

Nel paragrafo seguente si analizzano tre aree che nella costruzione dei dizionari di frequenza si rivelano problematiche: le difficoltà concernenti la selezione e l'ampiezza del corpus; la difficoltà di delimitare le unità lessicali; infine passiamo in rassegna gli approcci principali alla frequenza lessicale.

1.1.3.1 I requisiti del corpus

I dizionari di frequenza di norma devono informare sulle caratteristiche numeriche dei lemmi. In molti casi la frequenza assoluta di tali elementi è considerata la data più importante, quasi obbligatoria. Si aspetta quindi, di trovare accanto al lemma il numero delle occorrenze con cui appare in un determinato corpus. A titolo d'esempio,

[2] Per lo sviluppo della linguistica computazionale rinviamo a Bortolini *et al.* (1971) e Calzolari (1992).

nel *Lessico di frequenza della lingua italiana* (Bortolini *et al.* 1971) *deposito* 15 significa che *deposito* appare 15 volte in un corpus di 500.000 parole. Poiché il corpus menzionato intende essere rappresentativo dell'italiano scritto, a ragione possiamo aspettare che *deposito* appaia circa con la stessa frequenza in altri campioni della stessa popolazione statistica. Che questo non sia evidente, lo vedremo di seguito.
Consultiamo il lemma *corpus* nel *Grande dizionario italiano dell'uso* (GRADIT): "raccolta completa e ordinata di scritti, di uno o più autori, riguardanti una certa materia" e come termine linguistico "*campione* di una lingua preso in esame nella descrizione di una lingua" (De Mauro 1999: 346). Un campione è in ogni caso una sezione, una sottoparte della popolazione. Nessun problema emerge se il campione e la popolazione da indagare coincidono, p. es. quando la popolazione statistica si restringe alla produzione orale o scritta di un unico individuo nell'arco di un determinato periodo, in un'unica situazione comunicativa, oppure quando si esaminano insiemi chiusi, come quello dei fonemi. L'utilizzo dei campioni diventa problematico dal punto che si ha davanti un insieme *infinito* e *non omogeneo*, come il lessico. Emerge il problema di come estrapolare dati da una parte ristretta della popolazione (campione, corpus) in modo che essi siano validi per l'intera popolazione.
Secondo Bortolini (1971: 10-13) vi sono almeno due modi diversi di guardare la lingua come un universo statistico. La prima concezione parte dall'idea che le unità del sistema linguistico sono caratterizzate, oltre che da tratti qualitativi, anche dalle loro probabilità d'uso. Anche se queste probabilità non sono direttamente osservabili, è chiaro che le unità ricorrono nei testi con frequenze relativamente stabili, pertanto le frequenze osservabili nei testi si considerano come approssimazioni delle probabilità non osservabili del sistema. I rappresentanti più illustri di questa teoria erano P. Guiraud (1954) e G. Herdan (1964). Nel suo importante contributo Herdan spiega che ogni unità della *langue* è definita, tra l'altro, dalla sua probabilità d'occorrenza nella *parole*.[3] La rata di ripetizione di un determinato elemento del lessico in un certo campione tendenzialmente equivale al numero delle occorrenze nell'intero universo e in qualsiasi parte di esso, e questo indipendentemente dalla larghezza del corpus. Più tardi si è dimostrato che la frequenza non è un attributo stabile delle parole, ma varia da testo a testo in funzione dello stile, del tema ecc. La frequenza dei primi 1.000-1.500 lemmi sembra abbastanza stabile, ma a un livello di frequenza inferiore i dati sono più legati alla scelta dei campioni. Una diversa campionatura potrebbe portare a spostamenti e addirittura a variazioni di lemmi, pertanto la scelta del corpus campione a sottoporre a spoglio, ha estrema importanza.
Nella seconda concezione, rappresentata da Moreau (1962) non è espressa la probabilità come caratteristica intrinseca delle unità del sistema linguistico; l'universo statistico è definito come l'insieme di tutti i testi prodotti in un certo periodo. Si distinguono due categorie di parole: quelle appartenenti a classi aperte e quelle appartenenti a classi chiuse. Della classe chiusa fanno parte termini più o meno comuni a tutti i soggetti e a tutte le situazioni (parole "atematiche"), il cui impiego non varia sensibilmente nei diversi tipi di testo, quindi la stima della loro frequenza in base a testi campione non pone problemi. Alle classi aperte invece, appartengono parole "tematiche" la cui frequenza oscilla nei diversi testi. Per valutare la loro frequenza si propone di stratificare la lingua, delimitando dei centri d'interesse, all'interno dei quali le parole siano "tematiche". Così potranno essere messe in ordine di frequenza, all'interno del loro

[3] Herdan (1964) fu il primo ad interpretare la distinzione saussuriana di *langue* e *parole* in termini di popolazione statistica e campione: l'opposizione tra corpus e lingua non è altro che un'opposizione tra campione e popolazione statistica. La *langue*, in quanto l'insieme delle abitudini linguistiche della comunità, rappresenta il lessico della lingua in questione, mentre la *parole* è l'uso individuale, rappresentato a sua volta da campioni.

centro d'interesse, le parole tematiche. Tuttavia ci sono notevoli difficoltà relative alla determinazione dei centri d'interesse. Ci si chiede se non si debbano cercare anche per le parole di frequenza bassa il centro d'interesse cui si riferiscono, generando così una straordinaria moltitudine di strati, oppure se si debba definire a priori un numero limitato di centri d'interesse e cercare poi, all'interno di ciascuno, solo la frequenza di alcune parole. Si giunge alla conclusione che per alcune parole è impossibile delimitare alcun centro d'interesse in cui ricorrono con una frequenza stabile; tali parole coincidono per larga misura con le c.d. parole disponibili. Questo approccio è stato utilizzato per l'elaborazione del francese fondamentale (Gougenheim *et al.* 1964).
L'utilità teorica dei procedimenti empirici basati sulla raccolta di corpora linguistici è stata messa in dubbio dalla linguistica generativa. Per Chomsky una linguistica dei corpora è inadeguata perché un corpus è finito (chiuso), è incompleto ed esclude enunciati potenziali (possibili ma non realizzati), è influenzato da fattori accidentali e per la sua incompletezza non permette di fare legittimamente delle predizioni sulla grammaticalità delle frasi, ma ci dà *solo* indicazioni sulla frequenza, con la quale si manifestano determinate caratteristiche linguistiche (Chiari 2007: 30). Anche Lepschy (1978: 56) è scettico nei confronti della rappresentatività dei corpus:

> "Una lingua non è un insieme di testi; essa è piuttosto un sistema finito di regole che consentono di generare un numero infinito di testi. Non c'è raccolta che possa essere pienamente rappresentativa della lingua."

Subito dopo aggiunge però che ciò non vuol dire che i conteggi di frequenza siano inutili, ma dalla frequenza di una parola in determinato campione non si può trarre conseguenze dirette riguardo sua frequenza nella lingua. Da un lato tali critiche sono incontrovertibili, da un altro i dati testuali reali, raccolti e organizzati in maniera articolata, per quanto parziali, ci permettono di estrarre dati preziosi sull'uso che i parlanti fanno della lingua.
È un dato di fatto che i corpora sono sempre campioni: il nostro scopo è di estendere le osservazioni condotte sul corpus all'intera popolazione. Per fare ciò, i corpora devono mostrare quattro caratteristiche: rappresentatività, estensione, riferimento standard e forma elettronica (cfr. Chiari 2007). Prima di entrare nel merito dei primi tre criteri, vediamo cosa si intende con forma elettronica. Per i corpora moderni è importante che siano archiviati in forma digitale e che siano consultabili attraverso software di interrogazione. I corpora elettronici permettono di conservare una quantità sempre maggiore di dati, d'altra parte consentono di effettuare ricerche linguistiche complesse. Ora vediamo più in particolare i criteri relativi alla rappresentatività e all'estensione e al riferimento.

1.1.3.1.1 Rappresentatività

I testi sottoposti allo spoglio, secondo le intenzioni dei compilatori, si configurano come campioni della lingua intesa quale universo statistico. Pertanto da un lato occorre delimitare tale universo, dall'alto assicurarsi della rappresentatività del campione. Chiari (2004) elenca tre criteri volti ad assicurare la rappresentatività del corpus.
Per primo, il corpus deve presentare caratteristiche simili (anche quantitativamente) a quelle della popolazione per quell'aspetto che si intende investigare. In secondo luogo occorre garantire la rappresentatività in senso matematico: ogni possibile fenomeno linguistico deve avere la stessa probabilità di occorrere nel campione e nella popolazione. In terzo luogo bisogna assicurare la varietà tipologica (qualitativa) dei testi. Per far ciò, occorre evitare i due estremi, cioè di analizzare lunghi passaggi di alcuni autori, ritenuti

rappresentativi della lingua in oggetto (con più di 5.000-10.000 occorrenze secondo Füredi 1989: XII), oppure di esaminare brevi brani da più testi selezionati casualmente. La maggior parte dei dizionari di frequenza elabora brani non troppo lunghi, di ca. 2.000-3.000 occorrenze ciascuno. Notiamo tuttavia, che sono rari i corpora che contengono parti selezionate dei testi in modo da ottenere, artificialmente, elementi cofrontabili per es. per lunghezza. I corpora campionari (*sample corpora)* richiedono un intervento troppo forte del linguista, il che rende i dati estratti da tali corpora poco affidabili. Nei veri e propri corpora testuali i testi sono presi nella loro interezza, senza tagli o selezioni.
A questo proposito Juilland (1973: XII) osserva che nella selezione del corpus bisogna trovare un equilibrio tra unità (omogeneità) e diversità (eterogeneità): le fonti devono essere abbastanza omogenei al fine di rappresentare più di una meccanica agglomerazione di dati ricavati da diversi periodi o da diversi stili; al contempo devono essere abbastanza eterogenei per essere rappresentativi della lingua standard, non solo di un particolare tipo di testo o stile.[4] Kornai *et al.* (2006) in uno studio recente distinguono tre tipi di corpora in base alla coerenza interna dei componenti. Un c.d. *small-range* corpus è una raccolta statisticamente omogenea, solitamente composto di opere di un singolo autore e raramente eccede l'un milione di occorrenze. Il *medium-range* corpus contiene pochi tipi di testi di vari autori, infine il *large-range* corpus include la maggiore varietà di testi, generi e stili e rappresenta meglio l'uso linguistico standard, di solito comprende 1-100 milioni di occorrenze.
Per quanto riguarda la dicotomia della lingua scritta e parlata, i dati di Vedovelli (1993: 127) e Katerinov *et al.* (1991: 19) confermano che una lista di frequenza basata su campioni scritti è valida anche per la lingua parlata, salvo alcuni elementi specifici con funzione di connessione discorsiva.
La rappresentatività è una caratteristica che varia secondo l'aspetto linguistico che si intende studiare: un corpus rappresentativo per caratteristiche lessicali potrebbe non esserlo par caratteristiche di tipo stilistico o sintattico. In questo senso la rappresentatività è un concetto dinamico e relativo (Chiari 2007: 44).

1.1.3.1.2 Estensione

L'altro problema da affrontare è l'ampiezza del corpus. Il problema dell'infinità del lessico appare, a prima vista, più di carattere pratico che teorico, siccome la differenza assoluta tra la frequenza di una parola in un campione e la sua probabilità nella popolazione tende a zero con l'allargamento del campione. I compilatori dei dizionari di frequenza generali cercano di risolvere tale problema esaminando ampi corpora.
In realtà la dimensione ideale di un corpus di riferimento è tuttora una questione controversa. Generalmente si è concordi sul fatto che con meno di 15.000 occorrenze il corpus non può essere considerato rappresentativo (cfr. Chiari 2004). Per le lingue settoriali sembra accettabile costituire corpora di sole 100.000 occorrenze, giacché si tratta di uno strato ben delimitato del lessico, in cui il numero dei possibili lemmi è assai ristretto. Molti autori (tra cui De Mauro 1993: 29, Füredi 1989: XIII) affermano che per accertare la probabilità di alcune prime migliaia di lemmi il campione ottimale sia di

[4] Thorndike in *The Theacher's Word Book* (1944) ha usato le fonti più svariate, dalla Bibbia e dalle opere di Shakespeare, fino alla corrispondenza privata e ai libri di scuola, senza spiegare i motivi della scelta. Vander Beke nel suo *French Word Book* (1929) ha spogliato 33 romanzi, 13 drammi, 14 giornali e riviste, 13 opere scientifiche e filosofiche, 16 opere storiche e critiche che vanno da Balzac fino a Proust e Bergson.

500.000 occorrenze, poiché mezzo milione di "parole grafiche" sembra garantire la rappresentatività e la stabilità delle frequenze. Altri (Sciarone 1977: 13, Bortolini *et al.* 1973: VII, Chiari 2007: 45) hanno avanzato delle critiche alla dimensione "classica" di 500.000 occorrenze, in quanto il campione risente di fattori settoriali, in particolare per i lemmi di frequenza inferiore ai primi 1.000-1.500 la stabilità delle frequenze è molto variabile e diversi lemmi possono (s)comparire asseconda dei settori considerati.[5] Per Sgroi (1982: 87) il minimo sufficiente è il computo di almeno un milione di parole.
Quanto al limite superiore, con l'avvento dell'era dell'informatica, esso può essere facilmente spinto verso su. Tuttavia si nota che solo i dizionari pubblicati durante l'ultimo decennio hanno ecceduto il corpus di Käding di 11 milioni di occorrenze, pubblicato tra 1897-1898.[6] Per le lingue maggiormente studiate la media dell'estensione di un corpus linguistico di riferimento si aggira intorno ai 100 milioni di parole (ad es. il *British National Corpus* per l'inglese o il *Corpus di italiano scritto contemporaneo*, altri corpora, tra cui la *Bank of English*, raggiungono quasi 500 milioni di occorrenze). Rare sono peraltro i corpora di estensione non finita che prevedono l'aggiunta periodica di testi (i c.d. *monitor corpus*), come il *Bank of English* (COBUILD corpus), la maggior parte dei corpora è statica.
In realtà, sia l'ampiezza, sia la rappresentatività del corpus sono caratteristiche relative che dipendono in larga misura dall'aspetto linguistico che si intende indagare. Per giungere a risultati statisticamente attendibili, la misura del corpus va fissata in funzione alla quantità di lemmi o forme che si vuole estrarre. Chiari (2007: 44) osserva giustamente che le estensioni di corpora per analisi lessicali sono più ampie di quelle per le analisi fonologiche, mentre i corpora per analisi sintattiche superano per estensione quelle per analisi lessicali.

1.1.3.1.3 Riferimento standard

Il corpus largo non garantisce in sé la rappresentatività e l'affidabilità. Il fatto che la lingua è una popolazione disomogenea, rende difficile, per non dire impossibile, di trarne un campione rappresentativo e casuale. Prendendo l'italiano come esempio, si potrebbe dire che accanto all'italiano standard esistono numerose altre varietà più o meno vicine allo standard, situate lungo un continuum. Di conseguenza, se uno non definisce con chiarezza quale popolazione intende studiare, non è possibile, in termini statistici, trarne un campione rappresentativo. D'altra parte la variazione della lingua non è casuale, ma strutturata. Ogni lingua naturale varia in funzione dell'area geografica in cui è parlata, il gruppo sociale che la usa, la situazione comunicativa e il tempo in cui è utilizzata, il contenuto che esprime. In vista di quanto detto, molti compilatori di dizionari di frequenza hanno deciso, al fine di risolvere il problema dell'infinità e dell'eterogeneità del lessico, di usare un corpus stratificato come base empirica per i conteggi di frequenza. Molti sforzi sono volti a riprodurre nella composizione del corpus campione la complessità di strati e di stili della lingua nel suo insieme. Si veda ad esempio il *Lessico di frequenza dell'italiano parlato* (LIP), in cui De Mauro *et al.* (1993)

[5] Nella linguistica quantitativa si ritiene statisticamente attendibile quella parte delle liste di frequenza decrescente la cui variazione non supererebbe il 30% se si compilasse lo stesso dizionario di frequenza, secondo gli stessi criteri, con un corpus di grandezza uguale, ma composto di altri testi (cfr. Füredi 1989: XIII).
[6] Per l'italiano l'unico dizionario basato su un corpus di tali dimensioni è il VELI (1989) con 26 milioni di occorrenze, seguito dal CoLFIS (1995) tratto da 3 milioni di parole, mentre gli altri lavori come il LIF (1973), FDI (1973) e il LIP (1993) si attestano tutti sulle 500.000 occorrenze.

hanno cercato di creare un campione rappresentativo dei diversi tipi di parlanti in base alla provenienza geografica, titolo di studio, età, sesso e professione.
Il riferimento standard segnala la possibilitá dei grandi corpora di diventare dei punti di riferimento per lo studio delle varietà che rappresenta, attraverso l'esplicitazione delle metodologie di analisi, facilitando la comparazione tra corpora diversi ed evitando la necessità di ripetere i computi per ogni analisi successiva (Chiari 2007: 47).
Martin (1990: 1317) suggerisce che in caso ideale si dovrebbe partire da una stratificazione strutturata, osservando non solo i parametri menzionati (geografici, sociali, comunicativi, temporali e di contenuto), bensì anche quello della modalità (scritto o parlato) e della comprensibilità (varietà generali, letterari e tecnici). All'interno di questa struttura pluridimensionale si possono situare i testi del corpus di riferimento.

1.1.3.2 La parola come occorrenza e come unità lessicale

I lessicografi conoscono da lungo le difficoltà di definire cosa sia la *parola*. Ciononostante, non essendo direttamente interessati nelle questioni di frequenza e non dovendo esaminare dal punto di vista quantitativo l'uso di elementi lessicali, si potevano permettere di lasciare questi interrogativi in sospeso. Per la statistica, invece, affinché sia impeccabile, occorre definire con chiarezza gli oggetti dell'analisi. Tale richiesta può fare sembrare il lavoro con le *parole* un'impresa assai ardua, poiché tutt'oggi non esiste una definizione soddisfacente di questo concetto. Tuttavia, lo statistico lessicale, ad un certo momento del suo lavoro deve arrivare ad una decisione capace di eliminare ogni dubbio a questo riguardo.

1.1.3.2.1 La parola come unità elementare dei testi

Le tradizioni tipografiche ed ortografiche suggeriscono qui una soluzione che è accettabile nella maggioranza dei casi. Se sotto "unità grafica" si intende un gruppo di segni dell'alfabeto, separati da altri segni, da uno spazio o da un segno d'interpunzione, si giunge quasi sempre all'equivalenza 1 unità grafica = 1 parola.[7] Tuttavia, questa regola ha le sue eccezioni nei casi in cui l'unità grafica rappresenta più di una parola, e viceversa, quando una parola sta per più di una unità grafia.
A. *Un'unità grafica rappresenta più di una parola.* Appartengono qui le preposizioni articolate dell'italiano. Uno può decidere di analizzare queste forme nei loro costituenti, riducendo lievemente la mole del vocabolario ed aumentando il numero delle occorrenze, poiché la frequenza di tali parole si potrà perfino raddoppiare. Un altro gruppo di parole, la cui presenza nella stessa sequenza grafica crea dei dubbi, è quello delle parole con apostrofo. Nelle espressioni come *tutt'oggi, Sant'Andrea* la separazione delle due parole significa di spezzare queste espressioni in parti che nella lingua non esistono autonomamente. Sempre un altro gruppo problematico è costituito dai verbi accompagnati da un pronome clitico (*portami, andarci*, *dimmi*), poiché la semplice separazione del verbo dal pronome spesso porta a risultati insoddisfacenti.
B. *Un'unità grafica rappresenta una sola parola.* I motivi possono essere sia di carattere morfologico o sintattico, sia di tipo lessicale. Si pensi ad espressioni del tipo *sono venuto* oppure *tener presente*. Contare questi elementi come un'unica unità, in pratica non è una procedura molto facile. Per quanto riguarda i fenomeni morfo-sintattici, molti

[7] Cfr. Bortolini (1971: 35) "chiamiamo *parole* (o occorrenze) le successive unità grafiche di cui è costituito in testo […] una o più lettere (o caratteri equivalenti) tra spazi o segni d'interpunzione."

compilatori sono inclini ad eliminare completamente dal conteggio le parole grammaticali. Riguardo ai fenomeni lessicali, bisogna distinguere tra gruppi formati dal parlante e gruppi preesistenti nella lingua, ovvero tra combinazioni occasionali e quelli che appartengono al lessico (espressioni polirematiche). I primi vanno contati secondo i loro costituenti, mentre è meglio trattare i secondi come un'unica unità.

1.1.3.2.2 La parola come unità lessicale elementare

Assumendo che il primo problema sia stato risolto e il testo è stato diviso in "parole", il loro numero totale rappresenta il numero di occorrenze. Al fine di conoscere l'ampiezza del vocabolario, dobbiamo riordinare le occorrenze di ogni singola unità lessicale, disperse dappertutto nel testo. Come prima, quando il criterio grafico e ortografico è stato d'aiuto, anche questa volta troviamo che nella grande maggioranza dei casi possiamo seguire la regola 1 unità grafica = 1 parola. Tuttavia, anche in questo caso ci sono eccezioni alla regola, tra cui alcune più serie di quelle menzionate prima.
Prima di tutto occorre rilevare che per i diversi tipi di investigazione vanno effettuati diversi tipi di conteggi. La distinzione più importante a proposito è se le indagini servono a scopi didattici oppure linguistici. Se uno desidera esaminare le proprietà del lessico come sistema di codificazione, allora gli interessa in principio la forma delle parole, indipendentemente dal loro significato. Di contro, per esplorare parti del sistema lessicale per obiettivi didattici, è consigliabile differenziare la frequenza degli elementi lessicali a seconda il significato delle parole.
La lemmatizzazione, ovvero l'operazione di ricondurre le varie forme ad un unico esponente e di separare le forme identiche risalenti a due esponenti diversi, è un lavoro che ogni lessicografo e ogni compilatore di un indice di frequenza deve affrontare.
A. *Una parola è rappresentata da più forme.* L'unità semantica delle diverse forme grammaticali, dovute alla coniugazione o alla declinazione di un'unità lessicale è fuori dubbio, pertanto le varie modificazioni grammaticali di una parola sono considerate come rappresentanti della stessa voce. Questo vale senza restrizione per verbi e aggettivi e anche per le varie forme declinate dei sostantivi. Herdan (1964: 74) avverte che nel caso dei sostantivi denotanti esseri viventi ed aventi diverse forme secondo il sesso (*amico* e *amica*), emerge la questione se trattare queste forme come rappresentanti di due parole diverse. Il femminile di un sostantivo non è una flessione nel senso stretto del termine, perciò la soluzione migliore sarebbe considerare queste forme come rappresentanti di due parole diverse.
B. *Una forma rappresenta più parole.* Per quanto riguarda la separazione delle parole polisemiche da quelle omonimiche, l'etimologia, cioè il criterio seguito dai dizionari non è sempre di grande aiuto. A livello sincronico Herdan propone di seguire il criterio semantico, che nel caso delle parole con più significati prevede una qualche legame tra le accezioni. Tuttavia spesso è veramente difficile stabilire se questo legame esiste oggettivamente, e in questi casi è il dizionario a dare la risposta finale. Si ricorda che, in effetti, è lo scopo del conteggio di frequenza a determinare quale importanza attribuire alla separazione dei significati. Un conteggio con scopi didattici favorirà un alto grado di differenziazione delle forme secondo i rispettivi significati, mentre un conteggio di linguistica quantitativa può permettersi di minimizzare la differenziazione, perfino di farne a meno.
Un problema tipico è quello degli omonimi sintattici: nonostante il numero degli omonimi lessicali sia limitato, quello delle parole con numerose funzioni sintattiche,

dovute al cambio di categoria grammaticale, è infinito (ad es. tutti i participi passati possono essere classificati come forme verbali o aggettivali).
Infine, le parole grammaticali pongono problemi così numerosi e difficili che alcuni compilatori preferiscono eliminarle dai conteggi di frequenza. In questo modo si evita di dover classificare questi elementi, cosa che spesso causa difficoltà quasi insuperabili. Questa semplificazione crea tuttavia dei problemi, poiché non esiste un netto confine tra elementi lessicali e parole grammaticali. Basti pensare all'uso sia pronominale che aggettivale di alcuni elementi. Un altro esempio è quello degli ausiliari che in una certa misura sono elementi lessicali a piena ragione, e solo in misura minore funzionano come parole grammaticali. Herdan (1964: 75) osserva che l'eliminazione delle parole grammaticali dai conteggi potrà portare all'oscuramento di fatti linguistici e stilistici, pertanto in conteggi destinati ad un uso non specifico, l'inclusione degli elementi grammaticali diventa un imperativo.
Concludendo, occorre rilevare che qualsiasi norma si adotti per risolvere i problemi di sopra, essa deve essere sempre esplicita (che così non è vedremo nella sez. 2.4) al fine di permettere la comparabilità dei risultati.

1.1.3.3 Frequenza lessicale

I dizionari di frequenza con il contenuto della loro microstruttura (dati numerici) e/o con l'organizzazione della loro macrostruttura (spesso di frequenza/uso decrescente) creano l'impressione di ordini stabili, gerarchici. Questa convinzione diventa un po' meno ferma se si comparano l'uno con l'altro gli analoghi dizionari di frequenza, poiché si trovano risultati spesso confondenti e contraddittori. Prendiamo il rango di parole piuttosto comuni, ricavate dal *Lessico di frequenza della lingua italiana* Bortolini *et al.* (1971) (LIF), dal *Lessico di frequenza dell'italiano parlato* (De Mauro *et al.* 1993) (LIP), dal *Vocabolario elettronico della lingua italiana* (1989) (VELI) e dalla lista di Katerinov *et al.* (1991) (KBS).

	rango LIP	**rango LIF**	**rango VELI**	**rango KBS**
andare	26	39	66	34
ritornare	398	574	2003	662
veramente	156	482	1411	192
bene	45	63	195	52
papà	1262	735	3942	-
figlia	1248	329	-	310

I dati provengono in parte dal LIP (1993: 129-134).

I dati variano secondo la lista di frequenza, ad esempio la parola *papà* non rientra tra le 2.000 più frequenti secondo il VELI, basato su un corpus scritto di giornali con prevalenza di lingua formale, e non compare nemmeno nella lista KBS, probabilmente perché non ricorre fra gli argomenti del corpus parlato, mentre risulta più frequente nel LIF che nel LIP. Queste differenze vanno ricondotte al modo in cui è stato costruito il corpus.
Confrontando ancora le frequenze d'uso dei primi 100 lemmi del LIF e del FDI, basati entrambi su un campione di 500.000 occorrenze, si costata che i valori sono assai diversi:

	frequenza LIF	**frequenza FDI**
andare	2118	787
cosa	2020	645
fare	4526	2060
no	1127	428

sud	34	949

Osservando i numerali da 1 a 30, troviamo che nel LIF sono presenti tutti, tranne il *ventotto*, mentre nel FDI mancano *uno, undici, sedici, diciotto, diciannove, ventuno, ventidue, ventisei, ventisette, ventotto, ventinove*. Fra i giorni della settimana nel LIF manca il *mercoledì*, il FDI dà solo il *sabato* e la *domenica*.
Guardiamo ora una tavola apparecchiata:

	rango LIF	**rango LIP**	**rango VF**
aceto	-	-	-
coltello	1854	-	2438
cucchiaio	3948,5	4118	-
forchetta	-	-	-
olio	774,5	2279	1296
pepe	-	-	-
sale	1889,5	1551	2098
tovaglia	2664	3770	-
tovagliolo	-	4269	-

Secondo questi risultati, sulla tavola non si troverebbero una forchetta, aceto, pepe e forse neanche un tovagliolo.
Certe differenze tra i lessici di frequenza dipendono dal fatto che i testi spogliati appartengono a periodi diversi. Nel LIF mancano le parole appartenenti alla retorica ufficiale del periodo 1920-1940, età in cui i testi del campione FDI sono stati creati: *duce, dinamico, eroico, eroismo, inesorabile, impero, imperiale*; oppure hanno un numero d'occorrenze minore rispetto al FDI: *fascista* LIF 26, FDI 38; *nazione* LIF 49, FDI 111.
Tutto questo sembra dimostrare che la frequenza lessicale sia una cosa instabile, variabile che spesso dipende dall'intuizione di chi effettua il conteggio. Per affrontare questi problemi i ricercatori hanno assunto diverse attitudini che s'illustreranno di seguito.

1.1.3.3.1 Dizionari di frequenza "grezza"

In questi dizionari la frequenza trovata nel corpus di riferimento è riportata senza qualsiasi modifica, e le parole sono classificate secondo coefficienti uguali o proporzionali al numero delle loro occorrenze nel campione. Questi studi si basano sull'assunzione che il numero delle occorrenze di una parola nel campione rispecchi il suo uso nella lingua. Si creò l'impressione fallace di stabilità, la frequenza fu considerata la vera riflessione di una fissa probabilità d'occorrenza (il che è una contraddizione in termini). In Italia l'esponente più insigne di questo approccio è Sciarone (1977). Più tardi si è scoperto che i risultati basati su quest'idea non corrispondono alle stime intuitive degli specialisti, e spesso non soddisfano i criteri previsti.
Soprattutto i dizionari di frequenza pubblicati prima del 1960 appartengono a questa categoria, anche se alcuni (Vander Beke 1929) oltre alla frequenza pubblicano informazioni sulla dispersione (il numero di fonti diverse in cui la forma occorre), anticipando il prossimo approccio. La pubblicazione di studi fondamentali di linguistica quantitativa, come Guiraud (1954) e Herdan (1964) hanno dato nuovi impulsi ai lavori di lessicologia quantitativa.

1.1.3.3.2 Frequenza modificata da criteri distribuzionali "oggettivi"

Con il miglioramento delle competenze statistiche dei linguisti, comincia a guadagnare terreno l'idea che i dati grezzi di frequenza necessitano di qualche modifica per rispecchiare le probabilità. Siccome le frequenze di una lista valgono per i testi da cui esse sono tratte, se vogliamo servirci della lista per altre situazioni, dobbiamo calcolare in quale misura la frequenza dipende dal corpus in base al quale è stato calcolato, in altre parole, bisogna calcolare la stabilità della frequenza.
Per rappresentare la variabilità della frequenza nei vari tipi di testi è stato introdotto il valore di dispersione (distribuzione), misurato secondo diversi criteri più o meno complessi. La "dispersione" o "distribuzione" o "ripartizione" (o raramente "rango", in inglese "range") indica il numero dei testi o dei sottoinsiemi di un campione in cui un termine è usato (Sgroi 1994: 210). Per misurare la dispersione è necessario suddividere il corpus in parti e calcolare le frequenze relative delle unità per ciascuna parte del corpus. Sciarone (1995: 44) afferma a ragione che calcolare il numero di sottoinsiemi in cui la parola compare e prendere tale numero come valore di distribuzione (come è stato fatto, tra l'altro nel *Français fondamental*), fornisce un risultato rozzo e approssimativo, giacché fondato sul rapporto *presente/non presente,* e perché non si calcola quante volte la parola ricorre nei vari sottoinsiemi. Una soluzione più sofisticata è proposta in Juilland e Chang Rodriguez nel *Frequency Dictionary of Spanish Words* (1964): il coefficiente di dispersione misura la ripartizione delle occorrenze nei singoli sottoinsiemi dell'intero corpus, calcolando il rapporto tra la frequenza media nei 5 sottoinsiemi e lo scarto dalla media delle frequenze realmente trovate. La dispersione varia da 0 (frequenza accumulata in un solo sottoinsieme) a 1 (ripartizione perfettamente uniforme della frequenza del lemma nei *n* sottoinsiemi). Tra due parole di uguale frequenza la più nota avrà probabilmente una dispersione maggiore, perché le parole che compaiono in testi di vari tipi, hanno più probabilità di essere incontrate. Chiari (2007: 79) osserva a proposito che la scelta delle ripartizioni del corpus è il momento più delicato del trattamento, perché da questo deriveranno la misura della dispersione e, come si vedrà, anche l'indicatore dell'uso.
Dalla combinazione dei valori di frequenza e dispersione è nato il coefficiente d'uso, introdotto anch'esso da Juilland e Chang Rodriguez (1964). L'uso non è altro che l'applicazione del coefficiente di variazione, conosciuto dalla matematica, a fatti linguistici. Una parola con la stessa frequenza assoluta di un'altra, ma con una distribuzione più omogenea, avrà un valore d'uso più elevato, poiché l'uso è pari al prodotto della frequenza e distribuzione. L'idea sottostante è che le occorrenze che nei vari sottoinsiemi hanno ottenuto un livello di frequenza consistente (segnalato da un alto coefficiente di dispersione), si manterranno allo stesso livello di frequenza, senza brusche oscillazioni, se aggiungiamo nuovi sottoinsiemi oppure allarghiamo quelli esistenti del corpus. Al contrario si suppone che le occorrenze che oscillano in larga misura nei vari sottoinsiemi, non hanno ancora "trovato" il loro livello naturale di frequenza, e in un corpus ampliato tenderanno a stabilirsi a un livello più basso (Juilland 1973: XLII). La nozione d'uso è centrale nella statistica lessicale, poiché fornisce una rappresentazione quantitativa della presenza effettiva dell'unità lessicale nelle diverse parti del corpus. Questi dati sono usati in una serie di dizionari curati da De Mauro (1996, 1997a, 1997b, 1999, 2000), per indicare le marche d'uso, che corrispondono in molti casi a indicatori statistici.

1.1.4 La struttura statistica del lessico

Gli indici di frequenza si basano sui procedimenti della statistica linguistica che prevede l'applicazione di metodi statistici all'esame dei fatti linguistici. Citiamo le considerazioni esposte da De Mauro nell'*Enciclopedia Italiana* (1961:820):

> "le unità costitutive di una lingua (fonemi, parole, ecc.) soprattutto considerate sotto il profilo della frequenza con cui appaiono nei testi, costituiscono un tipico insieme di fenomeni di massa e sono perciò suscettibili di indagini statistiche per rilevare le frequenze medie del loro distribuirsi nel discorso e, nel tempo, le eventuali trasformazioni di tali frequenze."

Dall'esame delle frequenze delle parole e dei fonemi sono state ricavate alcune leggi fondamentali della statistica linguistica.[8] L'analisi di testi di vario tipo ha portato alla conclusione che:

> "un numero molto ridotto di parole selezionate in modo opportuno copre la maggior parte del testo ed è possibile compilare una lista in cui: le prime 100 parole coprono il 60% di qualsiasi testo; le prime 1000 coprono l'85% di qualsiasi testo; le prime 4000 coprono il 97,5% di qualsiasi testo; il resto copre il 2,5% di qualunque testo" (Guiraud 1957: 10).

Per copertura di un lessico intendiamo la percentuale di occorrenze del testo che sono contenute nel lessico. In sintesi, le prime 1.000 parole coprono più dell'80% del lessico di tutti i testi, i seguenti 1.000 elementi ca. l'8-10%, la terza fascia ne copre ca. 4%, la quarta e la quinta 2% ciascuno. Pertanto le prime 4.000 parole costituiscono in media il 95% del lessico di qualsiasi testo, la seconda fascia di 4.000 parole copre ca. 2-3%, mentre il resto copre meno del 2% (Krohn 1992: 117).
Inizialmente si suppone non solo che questa distribuzione delle frequenze sia comune a tutti i testi, ma anche che le parole di maggiore frequenza siano pressappoco le stesse in ogni testo. Ne deriverebbe dal punto di vista glottodidattico che per identificarle basterebbe spogliare un numero adeguato di testi, e che converrebbe iniziare da questo elenco l'insegnamento del lessico di una lingua. Apprendendo qualche centinaio di parole, teoricamente l'allievo sarebbe in grado di comprendere più dell'80% di qualsiasi testo. Da quest'affermazione è errato trarre la conclusione, coma lo fa Fóris (2002: 29), che chi conosce le 1.000 parole più frequenti di una lingua, capisce l'80% di qualsiasi testo, oppure l'80% di tutti i testi. Conoscendo, infatti, le sole parole grammaticali (articoli, preposizioni, ecc.) di una lingua, si dominerebbe in teoria più del 50% di un testo, senza essere in grado di pronunciare o di capire una sola frase in quella lingua. Le parole grammaticali (dette anche funzionali, sinsemantiche o vuote), infatti, esprimono le relazioni semantico-sintattiche sussistenti tra le parole autosemantiche (piene), ma non trasmettono informazioni cognitive sul contenuto del testo.[9] Al fine di ottenere risultati più attendibili, Krohn (1992: 120) propone di osservare nel calcolo del grado di comprensione il peso dei sostantivi e delle parole autosemantiche in generale. La comprensione delle espressioni idiomatiche ancor meno può essere resa con il numero delle parole coperte, giacché il senso di queste strutture non è identico alla somma dei significati degli elementi, che le compongono. È, infatti, molto difficile prestabilire la reale comprensione dei testi in base alla sola conoscenza delle parole più frequenti. È importante che secondo Guiraud le mille parole più frequenti, che grazie alle loro ripetizioni costituiscono l'85% di qualsiasi testo, forniscono solo il 50% delle informazioni contenute nel testo.

---

[8] Sulla statistica lessicale cfr. Guiraud (1954), De Mauro (1961), Herdan (1964), Heilman (1971) e le importanti introduzioni di Bortolini *et al.* (1971), Juilland (1973), Sciarone (1977, 1995) ai rispettivi dizionari di frequenza.

[9] Questa dicotomia risale ad Aristotele e si è poi ripresentata con diverse denominazioni in molte opere linguistiche. Ullmann p. es. parla di *parole piene* o *autonome*, aventi un significato autonomo dal contesto, e *particelle* o *parole accessorie* che, essendo portatrici di una certa funzione semantico-sintattica, acquistano il loro significato dal contesto in cui si trovano.

Sempre in merito al grado di copertura e comprensione, Sciarone (1977: 93) nota che la percentuale di parole diverse coperte sia una misura di comprensione più attendibile del valore di copertura delle occorrenze, poiché in un testo il numero delle parole diverse aumenta più rapidamente del valore di copertura. Le due percentuali (parole coperte e parole diverse coperte) si congiungono intorno al 90%, ne consegue che per comprendere un qualsiasi testo dobbiamo disporre di una lista di parole che lo copra per circa il 90%. In sintesi si potrebbe dire che nonostante il grado di copertura sia un valore oggettivo, non è stato dimostrato che l'efficienza comunicativa (ricettiva e produttiva), che ne è stata derivata, possa essere misurata altrettanto oggettivamente. È quindi opportuno evitare di identificare il valore di copertura di un lessico con la sua utilità comunicativa.
La linguistica quantitativa ha formulato delle leggi e ipotesi sulla struttura del lessico. Uno dei risultati più interessanti è che, un testo è di solito costituito per larga parte da un ristretto numero di parole di frequenza elevata, e in parte minima da numerosissime parole di bassa frequenza. Heilmann (1971: 87) arriva perfino a dire che il 95,5% circa degli elementi che costituiscono un testo è rappresentato dalle parole e dalle forme più frequenti. Traducendo quest'osservazione sul piano della comunicazione si costata che le parole più frequenti sono quelle più generiche. Il valore informativo di una parola decresce al crescere della frequenza, mentre con il decrescere della probabilità d'occorrenza di una parola cresce il suo valore informativo, ovvero la frequenza e la carica informazionale di una parola sono inversamente proporzionali. Potremmo esprimere questo concetto in una forma molto banale, affermando che una parola rara colpisce più di una voce comune. Le parole grammaticali, che di norma costituiscono il 30-50% percento di un testo, essendo molto frequenti, hanno un valore informativo minore a quello delle parole lessicali legate a contesti specifici.
Zipf (1949) fu uno dei primi a studiare i fenomeni di distribuzione delle parole.[10] La sua legge più nota riguarda un fatto osservato per la prima volta da Estoup (1907), ossia che la relazione fra il numero d'occorrenze di un termine, e il suo rango nella serie delle parole ordinate per frequenza decrescente è costante. Per verificare questa legge, abbiamo estratto alcune delle parole più frequenti del *Lessico elementare* (Marconi *et al.* 1994).

Rango	Lemma	Frequenza assoluta	Rango x frequenza
27	questo	4.962	133.974
30	vedere	4.288	128.640
32	ci	4.094	131.008
35	suo	3.935	137.725
39	volere	3.477	135.603

La legge di Zipf: il prodotto del rango e della frequenza è pressoché costante.

In base alla tabella possiamo rilevare che rango e frequenza sono inversamente proporzionali e il loro prodotto si aggira su 130.000. Più tardi si è dimostrato che ciò non è vero in tutti i casi. Nel corpus del *Lessico elementare,* ad esempio, per la parola più frequente (*lo*) il prodotto del rango e della frequenza dà 48.253, mentre per la parola dal rango 100 (*certo*) questo valore è 114.200.
Zipf spiega le regolarità statistiche delle lingue non come effetto del caso, bensì delle caratteristiche di finitezza psico-biologica dell'essere umano (Chiari 2007: 37). Per Zipf ogni attività è governata dal *principio del minimo sforzo*, che ipotizza un equilibrio tra i bisogni della comunicazione e l'inerzia dell'uomo, la quale porta a usare un numero ristretto di unità di valore più generale e di impiego più frequente (Martinet 1966: 171). In quest'ottica le situazioni comunicative tipiche della lingua parlata spingono il parlante

[10] Le considerazioni di Zipf sono state portate avanti da Guiraud (1954), Martinet (1966) e Heilmann (1971).

a scegliere gli elementi di più vasta applicabilità semantica. Si è anche dimostrato che gli elementi più frequenti sono generalmente quelli più brevi (legge di Zipf e Guiraud), in altri termini, il numero dei fonemi che compongono una parola è direttamente proporzionale al suo rango. D'altra parte, la legge di Zipf e Martinet prevede che più un fonema è frequente, meno netta è la sua articolazione. Si è osservato inoltre che gli elementi più frequenti sono quelli più semplici (di minor costo), più vecchi, si rifanno ad un etimo popolare ed hanno la maggiore estensione semantica, mentre generalmente si tende ad evitare le parole più complesse, lunghe e difficili (Arcaini 1978: 242). Tra frequenza e costo esiste lo stesso rapporto inverso che tra la frequenza e il valore informativo: le parole più frequenti sono quelle di minor costo, mentre gli elementi di maggior costo si presentano con una percentuale minore.

1.2 Il vocabolario di base

Dizionario di base (De Mauro e Moroni 1996), dizionario fondamentale (Baldelli *et al.* 2005), vocabolario minimo (Baldelli e Mazzetti 1974), vocabolario fondamentale (Sciarone 1977), vocabolario di base (De Mauro *et al.* 1980), vocabolario elementare (Cannella 1991), lessico elementare (Marconi *et al.* 1994), lessico di base (Batinti e Trenta 1983) ecc. sono tutte denominazioni del lessico di una lingua, ridotto secondo criteri prestabiliti, volto a facilitare l'apprendimento della lingua in oggetto. L'apparente varietà di nomi prelude, dal punto di vista terminologico, ad un problema fondamentale della lessicografia dei dizionari di base: tutte queste liste di parole, diverse fra loro per il tipo e la ponderazione dei criteri di selezione adottati, pretendono di rappresentare le parole più frequenti, più utili o più importanti per la comunicazione.

1.2.1 Definizione

Secondo Marello (1996: 144), il vocabolario fondamentale di una lingua include:

> "le parole condivise da tutti i parlanti della comunità linguistica ed è formato da parole comuni in accezioni comuni, non dovrebbe, a rigore, includere parole tipiche di una regione, di un sottocodice o di uno strato speciale o di epoche passate."

Per Cardinale (1994: 185) il lessico fondamentale è l'insieme dei lemmi presenti nell'intersezione dei vari idioletti dei parlanti.[11]
Tutti i concetti di lessico fondamentale si basano, almeno intuitivamente, su una nozione della psicologia dell'apprendimento secondo cui all'inizio dell'insegnamento è impossibile insegnare e imparare l'intero lessico della lingua straniera. Pertanto, sembra

[11] Da quanto detto emerge che il *Basic English* di Odgen (1937) e Richards non fa parte di questa categoria. Si consideri che un vocabolario di base non è mai sufficiente a sé stesso, e ha lo scopo di costituire la prima tappa nell'apprendimento di una lingua, mentre il Basic (*British-American-Scientific-International-Commercial*) è una lingua semiartificiale che non potrebbe servire da prima tappa per il possesso per l'inglese, tra l'altro perché 400 delle parole inglesi più frequenti non ne fanno parte. La lista contiene 850 parole (tra cui 18 verbi) più alcune regole grammaticali, scelte con l'intento di esprimere con il minor numero di parole il maggior numero possibile di parole e concetti. Le frasi dell'inglese naturale sono trasposte in modo analitico, p. es. una parola frequente come *enter*, rimane esclusa dalla lista, perché può essere espressa mediante *go into* che ha il vantaggio di essere usato anche in altre occasioni. Il valore di *Basic English* come mezzo versatile di comunicazione internazionale, è tutt'oggi discusso.

logico l'intento di offrire ai discenti un lessico prescelto di utilità e praticità universale. La gamma dei fattori di selezione è molto vasta, dal senso linguistico personale del lessicografo (Migliorini 1943), attraverso criteri comunicativo-pragmatici (Galli de' Paratesi 1981), la neutralità stilistica (Gabrielli 1981), la capacità di formare composti e derivati (Thorndike 1936), l'importanza e l'utilità pratica (Putzer 1998), fino a criteri statistici come il valore d'uso (Marconi *et al.* 1994), la dispersione (Russo J. A. 1962) o la frequenza (Sciarone 1977, Katerinov 1991).

1.2.2 Storia della lessicografia dei dizionari fondamentali

Sulla scia di Krohn (1992) e Kühn (1990) nella storia della lessicografia dei dizionari fondamentali possiamo distinguere quattro grandi fasi. Al criterio della frequenza, dominante nei primi lavori nati agli inizi del XX sec., si associano ben presto altri criteri più o meno oggettivi, volti a migliorare l'attendibilità dei dati di frequenza. A partire degli anni Settanta si fa largo l'approccio comunicativo che ha dato nuovi contributi non trascurabili, mentre nel periodo più recente si cerca di integrare l'approccio statistico e quello pragmatico in un quadro frequenziale-comunicativo.

1.2.2.1 Le origini: dizionari basati sulla frequenza

Le riflessioni teoriche e pratiche sulla definizione del lessico di base hanno le loro origini nella lessicografica dei dizionari di frequenza. Nel primo periodo la compilazione dei dizionari di base è strettamente legata alla statistica linguistica, i primi dizionari fondamentali nascono, infatti, come prodotto di scarto delle ricerche sulla frequenza lessicale. I lessicografi dei dizionari di base hanno assunto senza critica i processi e risultati dei dizionari di frequenza. Con il criterio della frequenza hanno pensato di ridurre il lessico di una lingua alle parole più frequenti, ritenuti le più generali, comuni, usati, fondamentali e utili. Grazie ai calcoli si poteva conferire al lessico fondamentale un carattere oggettivo, poiché la statistica permetteva di definire con metodi quantitativi (scientifici) la frequenza d'occorrenza di una parola. Anche dal punto di vista strutturale si scopre una stretta correlazione e una certa dipendenza dai dizionari di frequenza. Spesso si tratta di pure liste di parole, elencate in ordine di frequenza o alfabetico.

			Uso	**Rango**
160	**allora**	*avv*	722.47	160
161	**divertire**	*v*	720.61	161
162	**correre**	*v*	713.05	162
163	**riuscire**	*v*	711.99	163
164	**sorella**	*sost*	711.57	164
165	**compagno**	*sost*	698.65	165
166	**alcuno**	*agg/pron*	692.96	166
167	**quale**	*agg/pron*	685.63	167
168	**lasciare**	*v*	682.87	168
169	**cui**	*pron*	682.83	169

Estratto dal *Lessico elementare* di Marconi *et al.* (1994: 43).

		Rango
pre	A	8
sos	ABATE	9448
ver	ABBANDONARE	1360

agg	ABBANDONATO	3076
sos	ABBANDONO	3314
sos	ABBASSAMENTO	5600
ver	ABBASSARE	2583
avv	ABBASTANZA	1173

Estratto dal *Vocabolario elettronico della lingua italiana* (1989: 255).

Il primo dizionario fondamentale vero e proprio fu la lista dell'americano Eldridge (1911), redatto secondo il criterio della frequenza, per facilitare l'apprendimento dell'inglese da parte di operai immigrati. Non stupisce che il primo tentativo sia stato compiuto negli Stati Uniti. La forte spinta che viene a questo tipo di ricerca è dovuta alla necessità d'apprendimento dell'inglese da parte di popoli ex-coloniali ed altri, attratti dalla potenza politico-economica del mondo anglosassone.
Sempre negli Stati Uniti, dopo l'apparizione del vocabolario di frequenza di Käding (1898) B. Q. Morgan pubblicò il suo lessico di base (*German Frequency Word Book. Based on Käding's "Häufigkeitswörterbuch der Deutschen Sprache"*, New York, 1928). La lista di frequenza di Käding costituisce il punto di partenza per la costruzione di molti dizionari di base, sia quando i suoi conteggi di frequenza sono stati adottati direttamente (Morgan 1928), sia quando le sue liste sono state modificate e corrette (Bakonyi 1934). Sulle fondamenta dei lavori di Käding (1898) e Morgan (1928) dagli anni Trenta in poi nascono una serie di dizionari di base, soprattutto negli Stati Uniti, in cui il lessico fondamentale è definito con il criterio della frequenza.[12]
Il più importante fra quelli pubblicati per primi, per l'inglese è il lavoro di Faucett, Palmer, Thorndike, West (1936) *Interim report on vocabulary selection*. I collaboratori intendono offrire un materiale suscettibile di essere utilizzato nell'insegnamento: l'obiettivo è di delimitare un vocabolario chiuso, autosufficiente, dal quale consentire espansioni successive. Con queste premesse si arriva ad un elenco di utilità generale, ovvero il *General Service List* di West (1953) che consta di ca. 2.000 vocaboli fondamentali, più 2.000-3.000 termini speciali. Per la compilazione sono stati adottati criteri come la frequenza, il valore strutturale, l'universalità geografica, la specificità semantica, la capacità di formare derivati e composti, il valore stilistico. Oltre che contare la frequenza delle singole parole nel corpus di 5 milioni di occorrenze, ad ogni sotto-significato è associato un numero percentuale che rappresenta la frequenza di quel significato rispetto alle occorrenze della parola-lemma. Un'altra iniziativa interessante è quella di Eaton (1934): si tratta di una lista di frequenza comparativa delle lingue inglese, francese, tedesco e spagnolo. Per ogni lingua sono riportate le 1.000 parole più frequenti, in più il corrispondente di ciascun vocabolo in esperanto.
Ben presto si è notato che per ottenere liste veramente affidabili, la sola frequenza, il cui studio era ancora in fase iniziale, doveva essere integrata con altri criteri. Buchanan già nel 1927 propose di prendere in considerazione la lingua parlata, ma prima di tutti Morgan (1928) ha riconosciuto alcune incongruenze della lista di Käding (dovuti in principio alla mancata lemmatizzazione e distinzione degli omografi nella lista originale). Vander Beke (1929) ha introdotto il criterio della ripartizione,[13] Thorndike

[12] Per l'inglese cfr. Thorndike E. L. (1921) *The Theacher's Word Book*, New York; Faucett L. W., Palmer H. E., Thorndike E. L., West M. P. (1936) *Interim report on vocabulary selection for the teaching of English as a foreign language*, London, King and Son; Thorndike E. L., Lorge I. (1944) *The Theacher's Word Book of 30.000 Words*, New York; per il francese Henmon V. A. C. (1924) *A French Word Book Based on a Count of 40.000 Running Words*, Madison; Vander Beke G. E. (1929) *French Word Book*, New York; per lo spagnolo Buchanan M. A. (1927) *A Graded Spanish Word Book*, Toronto; Keniston H. (1933) *A Basic List of Spanish Words and Idioms*, Chicago. Per ulteriori informazioni si vedano Arcaini (1978: 243-247), Gougenheim *et al.* (1964: 19-58).

(1936) ha sviluppato criteri aggiuntivi, come il rango di applicabilità alle varietà di soggetti, il c.d. *subject range*.
Grazie alla stretta relazione con gli studi di frequenza lessicale, riguardo alla selezione del lessico fondamentale si è diffusa un'idea assai pericolosa, che perfino oggi gode di una certa popolarità, ovvero che le parole più frequenti sono quelle più utili per l'insegnamento delle lingue. Che le parole più utili per la didattica non sono necessariamente quelle più frequenti in un corpus della lingua scritta, è stato spiegato per la prima volta da Bakonyi (1934): egli ha riconosciuto che le espressioni che servono per dare istruzioni, spiegazioni durante l'insegnamento non compaiono nelle liste di frequenza. Tuttavia nella prima fase della lessicografia dei vocabolari di base rimane valido il principio che per ottenere gli elementi fondamentali si conteggia tutto quello che può essere conteggiato.
Riguardo all'utilità dei lessici fondamentali basati sulla frequenza, dagli anni Settanta in poi prevalgono le opinioni negative. La descrizione più completa, insieme alla critica più feroce dei dizionari di base orientati alla frequenza si trova in Kühn (1979). Ricordiamo i principali punti discussi: a livello lessicostatistico egli critica che i dati di frequenza ricavati sono diversi secondo il corpus utilizzato, pertanto sono rappresentativi solamente per il corpus in oggetto, ma non per la lingua standard (1979: 43); i testi di riferimento, a causa della loro età e origine, rappresentano spesso un lessico obsoleto o periferico (1979: 43); giacché le definizioni dell'unità statistica 'parola' sono assai diverse, è molto difficile condurre uno studio comparativo dei dizionari di frequenza (1979: 45). A livello lessicosemantico egli critica che i conteggi statistici considerano il segno linguistico bifacciale solo come significante, pertanto non possono assegnare significati alle singole forme (1979: 49); in alcuni campi lessicali il 'popolamento' lessicale (*Besetzung*) non è consistente, perciò ci sono delle lacune nel lessico fondamentale, ad esempio mancano coppie di antonimia come *caldo – freddo* (1979: 52); non sono considerati sufficientemente i rapporti sintagmatici, ad esempio lessemi apparentati come *mangiare – verdura* non sono rappresentati in modo sistematico (1979: 55).
La maggior parte degli esempi citati da Kühn è ricavata dal dizionario di frequenza di Käding, quindi Krohn (1992: 67) osserva a ragione che si tratta di critiche rivolte ad un certo tipo di dizionario (di base), che si cerca di sostenere con esempi presi da un dizionario di tipo diverso (di frequenza). Ciononostante le critiche di Kühn non sono infondate. A titolo d'esempio, alla scarsa rappresentatività e al carattere obsoleto del corpus sottostante si deve l'inclusione nel *Vocabolario di base* (De Mauro *et al.* 1980), basato a sua volta sul *Lessico di frequenza della lingua italiana* (Bortolini *et al.* 1971), di alcuni termini militari come *aviatore, aviazione, bombardare, brigadiere, caporale, comandante* ecc. Anche nella ricostruzione dei campi lessicali si incontrano delle difficoltà: nel VELI (1989), tratto da un corpus di 26 milioni di occorrenze, nella serie dei mesi dell'anno è assente la parola *ottobre*, mentre nel *Vocabolario fondamentale* di Sciarone (1977), sebbene con ranghi assai diversi (*maggio* 929 – *marzo* 2058), ritroviamo tutti i nomi dei mesi e delle stagioni, ma sono assenti cinque giorni della settimana a causa del loro basso valore di frequenza. Secondo l'autore il sabato e la domenica sono due giorni speciali della settimana, pertanto è naturale che compaiano nella lista, mentre per indicare gli altri giorni spesso usiamo formule del tipo 'ieri', 'domani' ecc. Schumacher (1998: 163), infatti, ha osservato che se in un determinato campo semantico competono diversi elementi appartenenti alla stessa categoria grammaticale, la probabilità che uno di essi appare nella fascia di frequenza più alta, sarà

[13] Vander Beke nel suo *French Word Book* chiama *range* (R) il numero di sottoinsiemi, in cui l'unità lessicale è presente: poiché operava su 88 testi diversi, R variava da 1 a 88. Nel suo vocabolario di base ha eliminato le unità lessicali con un R inferiore a 5.

relativamente scarsa. In altre parole, più parole grammaticalmente imparentate stanno a disposizione, minore è la probabilità d'occorrenza di ognuno di essi. È ovvio che in sede didattica gli elementi mancanti delle serie chiuse vanno aggiunti, così da disporre della serie completa.
Comparando gli elementi lessicali riportati in alcuni vocabolari di frequenza, si giunge alla conclusione che palesemente non è così facile stabilire quali sono le parole più frequenti e quindi più utili. Ad esempio, nel dizionario di Sciarone (1977), nella fascia del vocabolario fondamentale di De Mauro *et al.* (1980) e nel LIP (De Mauro *et al.* 1993) con un rango d'uso superiore a 2.500 sono elencati alla lettera U 47 lemmi in totale, che si suddividono come segue: Sciarone 37, De Mauro 25, LIP 30. Solamente 16 parole compaiono in tutte e 3 le liste (*uccidere, ufficiale, ufficio, uguale, ultimo, unico, unire, università, uno, uomo, urgente, urlare, usare, uscire, uso, utile*), 16 parole sono nominate in due liste e 14 compaiono in una sola lista. Sono esotiche le parole nominate in una sola lista? Perché si tratta in prevalenza di sostantivi? Quale significato è stato considerato?
Riassumendo, la sostanza delle critiche rivolte a questa prima fase della lessicografia dei dizionari di base riguarda la convinzione che i conteggi statistici di frequenza costituiscano il criterio in assoluto per la definizione delle parole più utili.

1.2.2.2 Frequenza e criteri aggiuntivi

Cercando di utilizzare i risultati dei dizionari di frequenza nell'insegnamento del lessico, si è scoperto ben presto che i dizionari di frequenza "grezza" non erano abbastanza informativi. A causa di queste incongruenze e svantaggi si è cercato, in un secondo tempo, da un lato di perfezionare la metodologia statistica, d'altro lato di introdurre altri parametri nella definizione del lessico fondamentale. E' difatti merito del team di Gougenheim di aver modificato dal punto di vista qualitativo i principi della definizione del lessico di base.
Gougenheim, Michéa, Rivenc e Sauvageot ne *L'élaboration du français fondamental* (1964) hanno scelto per primi come base per la definizione del minimo lessicale la lingua parlata, registrando conversazioni di persone poste in situazione tipiche della vita di ogni giorno. Il corpus di ca. 312.000 parole incluse 163 conversazioni informali, con 8.000 voci differenti, raccolte prevalentemente nella regione parigina.[14] Dopo l'identificazione delle 1.063 parole con una frequenza pari o superiore a 20, la lista conteneva soprattutto termini grammaticali, verbi, aggettivi e qualche sostantivo di carattere generale, mentre mancavano molti nomi comuni, legati per lo più a temi e oggetti della vita quotidiana. I sostantivi concreti, infatti, hanno una frequenza molto instabile, poiché in molte situazioni i nomi concreti sono evocati per mezzo di pronomi, e il nome stesso non viene pronunciato. Ma una lista di frequenze non è un vocabolario di base, e la frequenza coincide solo in parte con l'utilità didattica, quindi, in aggiunta alle liste di frequenza e di distribuzione della lingua parlata, sono stati introdotti due concetti: la *disponibilité* e l'*empirisme rationnel.* Con *disponibilité* s'intendeva la disponibilità associativa di parole intorno a determinati *centres d'intérêt*. Il vocabolario disponibile è costituito da parole legate ad esperienze quotidiane eppure caratterizzate da una bassa frequenza e da una

[14] In merito alla scelta di basare il lessico fondamentale su un corpus della lingua parlata, vari autori hanno avanzato le proprie critiche. Sgroi (1982: 86) ricorda che un simile metodo di ricerca nasconde non pochi pericoli, come il numero ridotto di intervistati, l'ampiezza ridotta del corpus, la difficoltà di procedere a una campionatura geografica e sociale degli intervistati e l'impossibilità di conferire alle conversazioni un carattere spontaneo. In più, la differenza tra la lingua scritta e quella parlata non si colloca sul piano del vocabolario, pertanto una lista estratta da testi scritti può senz'altro dare risultati validi per la lingua parlata (cfr. Sgroi 1977, Katerinov 1991).

forte dipendenza dal contesto. Per individuare tali parole, sono stati elaborati 16 centri d'interesse. Di seguito, a 900 allievi di licei francesi è stato chiesto di scrivere le 20 parole che sono venute immediatamente nella loro mente a proposito di un determinato campo.[15] Con l'*empirisme rationnel* si intendevano aggiunte empiriche, cioè intuitive. In questo modo si sono identificate le 1.475 parole assenti dai precedenti spogli. Nell'elaborazione del francese fondamentale sono state dimostrate per la prima volta le insufficienze del puro metodo statistico nella realizzazione di un vocabolario destinato ad uso didattico.

In questa seconda fase della storia della lessicografia dei dizionari fondamentali accanto alla presa in considerazione della lingua parlata e il criterio della disponibilità troviamo ulteriori criteri con cui si è cercato di modificare le semplici liste di frequenza: "familiarity" (Richards 1970), "il fattore pedagogico" (Siliakus, Schlick 1965: 46-52), "abitudini e conoscenze linguistiche" (Zertifikat DaF 1972: 16), "coverage-index" (Mackey, Savard 1967: 3), "efficacia" (Verlée 1954). Si veda in particolare il principio dell'efficacia introdotta da Verlée (1954) e riproposta in una formula di Herdan (1964). L'idea di base è che certe parole si riconoscono e si memorizzano più facilmente di altre, perché sono collegate a una radice che è stata in precedenza assimilata. Ne consegue che in un vocabolario limitato è necessario evidenziare i termini che generano altre parole. Si propone quindi di aggiungere alla frequenza assoluta di un termine produttivo le frequenze dei derivati e composti formati con quella base. Analogamente, secondo Sgroi (1982: 84-94) bisogna seguire scrupolosamente tre principi fondamentali di selezione: frequenza, ripartizione ed efficacia (cioè produttività e disponibilità). Per produttività Sgroi intende la capacità dei singoli lemmi di formare dei derivati e/o composti. Due lemmi di frequenza uguale avranno quindi ranghi diversi se uno di essi compare sistematicamente in derivati e/o composti.

Trattandosi di criteri assai soggettivi, non è facile verificare quale impatto hanno avuto sulla determinazione del lessico fondamentale, poiché sono difficilmente adoperabili nel lavoro lessicografico. Non sorprende che invece di specificare i criteri di selezione adottati, gli autori spesso si accontentano di affermazioni molto generiche del tipo: il dizionario

> "è nato dall'esperienza dell'Università Italiana per Stranieri di Perugia. […] Nel *Vocabolario minimo* si trovano le parole di uso più frequente della lingua italiana" (Baldelli e Mazzetti 1974).

Caratteristica per il secondo periodo è l'adozione dei risultati della lessicostatistica moderna. Grazie al diffondersi delle tecniche di elaborazione automatica dei dati, i dizionari fondamentali potevano beneficiare delle nuove possibilità di spoglio e di analisi statistica dei dati. I grandi dizionari di frequenza di Juilland e i suoi collaboratori (1964, 1965, 1970, 1973) divergono dalle tradizionali liste di frequenza, poiché riportano per ogni unità lessicografica l'indice di dispersione, calcolato alla base della deviazione standard e coefficiente di variazione; inoltre è riportato il coefficiente d'uso, ovvero il prodotto del valore di dispersione e della frequenza. Il rapporto tra frequenza, dispersione ed uso indica l'appartenenza di un elemento al lessico di base e il suo rango all'interno di una lista di parole. Nonostante queste differenziazioni e modifiche della frequenza

[15] Sciarone (1995: 28-32) e Sgroi (1982: 92-94) criticano fortemente il metodo utilizzato da parte degli autori del *Français fondamental* per la selezione delle voci disponibili. Partiti dalla convinzione (errata) che una lista di disponibilità può fornire molti nomi, ma pochi verbi instabili, gli autori hanno chiesto agli studenti coinvolti nell'inchiesta di scrivere, a proposito di numerosi centri d'interesse (ad es. *le parti del corpo, vestiti, mezzi di trasporto* ecc.), soltanto sostantivi. Considerando valide le parole indicate da oltre il 50% degli intervistati, per ogni centro d'interesse troveremo solo una dozzina di parole. Il reperimento dei verbi disponibili, condotto su un centro d'interesse assai vago (*i verbi d'azione usuali*) ha fornito un risultato ancora peggiore con soli 11 verbi indicati da più della metà degli intervistati.

lessicale, la lessicografia dei dizionari di base in questo secondo periodo della sua storia non trova una valida soluzione per i problemi emersi, infatti, tutte le liste di base si basano in principio sul criterio quantitativo della frequenza lessicale; i criteri qualitativi aggiuntivi sono introdotti solo al fine di riparare le lacune dei conteggi di frequenza; le soluzioni metodiche proposte rimangono nell'ambito della lessicostatistica e non danno origine a un nuovo concetto teorico; le liste di base sono sincategorematiche e non hanno il carattere di un dizionario.

1.2.2.3 L'approccio comunicativo-pragmatico

La svolta comunicativo-pragmatica anche nella lessicografia dei dizionari fondamentali ha portato ad un mutamento paradigmatico, grazie al quale i termini come pragmatica, comunicazione e competenza diventano concetti importanti per la selezione lessicale. Nel tentativo di ampliare il lessico di alto uso ad un minimo lessicale atto ad ottenere determinati obiettivi didattici, sono venuti meno i criteri di selezione scientifici, perlomeno se con il termine "scientifico" si intendono i valori rigorosi delle scienze naturali. Per ovviare a questo dilemma, ossia che le esigenze scientifiche non sono compatibili con l'utilità didattica delle liste lessicali, si è cercato di definire con più chiarezza i criteri di selezione "pragmatici" e i processi orientati ai gruppi di discenti. L'obiettivo è di determinare dei minimi linguistici per i bisogni comunicativi di vari gruppi di apprendenti. È stato riconosciuto che sulla base degli obiettivi didattici bisogna definire il minimo lessicale, che a sua volta deve concordare con le strutture grammaticali di base. Un minimo lessicale di questo tipo deve essere non solo chiuso in sé, ma deve rispecchiare adeguatamente le strutture di base della lingua in oggetto, deve soddisfare i bisogni comunicativi dei destinatari, infine deve essere ampliabile in modo sistematico. Per chiarire i bisogni specifici dei discenti, si cercava di catalogare i temi, le situazioni, i testi ed i possibili atti comunicativi che l'apprendente deve essere in grado di compiere. Questi cataloghi (come quello riportato in Ciliberti 1994: 99-110) sono da intendere come elementi per la determinazione degli obiettivi dell'apprendimento, considerando i quali si possono compilare e verificare le liste di parole. Il criterio statistico vi gioca un ruolo modesto.
Nel 1972 è uscita una lista di 2.000 parole, destinato come materiale d'esame per l'ottenimento del certificato di tedesco come lingua straniera (Zertifikat DaF) (Steger: 1972). Fin dall'inizio si cercava di armonizzare la selezione lessicale con gli obiettivi didattici. La selezione, effettuata sulla base di una lista di 4.000 parole estratte da testi didattici, si fondava sulla valutazione di un gruppo di esperti, e il criterio di frequenza ha avuto solo un ruolo indiretto nel giudizio della commissione. Nel lavoro sono state considerate alcune categorie pragmatiche (intenzioni, temi e situazioni comunicativi di base) che permettono al discente di parlare di temi della vita quotidiana. Nelle successive revisioni del 1977 e 1995 è stato sistematicamente verificato in quale misura le situazioni, temi ed intenzioni si lasciano esprimere con l'inventario lessicale previsto. È stata presa in considerazione la morfologia delle parole: sono state scartate le espressioni che si possono derivare tramite regole morfologiche produttive. Accanto a questo perseguimento sistematico dell'approccio pragmatico sono state osservate le relazioni lessicali (sinonimia, iponimia e antonimia) al fine di alleggerire e rendere più consistente il vocabolario. Nelle versioni riviste la frequenza è stata reintrodotta come criterio di lavoro. L'utilità della raccolta si è dimostrata veramente alta: quasi tutti i libri didattici apparsi dagli anni Settanta in poi si basano sulla lista ZDaF.

Un'altra iniziativa che prende spunto dall'approccio pragmatico, è il progetto *Lingue moderne* del Consiglio d'Europa, nell'ambito del quale, a partire dal 1971, si è cercato di definire il lessico fondamentale facendo ricorso al concetto di *Threshold Level*, ovvero *Livello soglia*, applicato finora a ventitré lingue europee.[16] Il lessico del *Threshold Level*, il primo di una serie di lavori finalizzati alla selezione del vocabolario fondamentale, è stato fissato in 1.000 unità lessicali, in più altre 500 per l'uso solamente ricettivo. L'approccio fondamentale del progetto è la teoria degli atti comunicativi, introdotta negli anni Sessanta da Austin e Searle. Il principio di base è che la gente, parlando, compie degli atti linguistici e la struttura di un enunciato dipende dal tipo di atto che ha realizzato. Questo tipo di metodo comunicativo vuole dotare il parlante straniero di mezzi linguistici necessari per raggiungere un certo scopo in una data situazione comunicativa. Per definire il lessico fondamentale bisogna considerare le situazioni in cui si dovrà usare la lingua straniera; i temi che possono emergere durante le discussioni; il tipo di attività linguistica che il discente deve compiere, p. es. leggere, parlare, ecc., le funzioni che l'allievo deve attuare, p. es. comunicare; le espressioni generali di cui avrà bisogno, p. es. le nozioni di luogo come *davanti, dietro*; le nozioni specifiche, legate ad un certo tema, p. es. al tempo meteorologico sono legate parole come *sole, piovere*; le concrete forme linguistiche di cui avrà bisogno; infine il livello di prestazione del discente.
Sull'approccio comunicativo-pragmatico si basa anche la proposta di Kühn (1979: 62). Per selezionare il lessico fondamentale egli identifica una serie di criteri in cui dominano chiaramente i componenti extralinguistici e pragmatici della competenza comunicativa:

> "di quale materiale lessicale (LM) della lingua standard ha bisogno un parlante/uno scrittore (S), per compiere con successo l'intenzione comunicativa (KI) nella situazione (Sit) sul tema (T) nel ruolo (R) con l'aiuto di una determinata modalità di comunicazione (KM)?" (Traduzione nostra.)

Un minimo lessicale comunicativo dovrebbe pertanto contenere le parole che rientrano nel contesto dei temi rilevanti per la comunicazione. I temi rilevanti per la comunicazione sono quelle che "costituiscono la maggioranza delle forme scritte e orali della comunicazione" (1979: 68). Krohn (1992: 72) nota che con questa definizione Kühn ripropone il criterio di frequenza, criticato con tanta veemenza, anche se, invece di un criterio di ricorrenza lessicografica, qui si opera con un criterio di ricorrenza tematica, in cui la selezione si basa sulla frequenza d'occorrenza dei temi nei singoli testi. Altri fattori aggiuntivi sono: nella gerarchia dei termini le parole generiche vanno preferiti a quelle più specifiche; si preferiscono le parole morfologicamente semplici a quelle composte; quelle stilisticamente neutrali e più "utili" per una determinata categoria di apprendenti.
Proposte simili si ritrovano in un articolo di Scharnhorst (1995), in cui l'autore applica alle diverse fasce del lessico i concetti di centro e periferia, introdotti negli anni Sessanta dai linguisti praghesi. Egli giunge alla conclusione che il nocciolo (lo strato centrale) del lessico è composto di parole di origine indigena, monomorfematiche, stilisticamente neutrali, di uso frequente, appartenenti alla lingua standard. È vero, in parte, che in un lessico fondamentale bisogna includere le parole aventi una struttura morfologica abbastanza semplice, che servono come base per la la creazione di altri elementi lessicali, ciononostante è chiaro che non tutti i termini fondamentali hanno una struttura semplice, e non tutte le parole "semplici" sono fondamentali o abbastanza frequenti per far parte del lessico di base. Un altro problema è che al lessico fondamentale appartengono non solo espressioni stilisticamente neutrali, ma anche locuzioni marcati, soprattutto

[16] Per il livello soglia dell'italiano cfr. 2.2.17, per l'ungherese si veda 3.2.6, per il quadro teorico generale del progetto si rinvia a 5.1.1.

nell'ambito della comunicazione orale (cfr. alcune scelte di Galli de' Paratesi in 2.2.17). In molti casi la decisione di assegnare una parola al lessico fondamentale dipende dal peso che si vuole attribuire ai singoli criteri. Quanto all'utilità dei termini, proposta da Kühn, occorre valutare i limiti del concetto "lessico fondamentale", giacché a voler rilevare i termini più utili per ciascun gruppo di apprendenti, ci vorrebbero tanti lessici fondamentali, quanti i gruppi di destinatari stessi. Pertanto l'unica possibilità è di creare un c.d. lessico elementare, il quale, anche se non permette di esprimersi in modo chiaro ed esauriente in tutti i contesti, serve come nocciolo per il lessico fondamentale individuale di ciascun apprendente.
Gli approcci di tipo pragmatico testimoniano, che il punto di riferimento per la determinazione del lessico fondamentale non è più costituito dalle ricerche di frequenza lessicale, la selezione si basa ormai sull'effettivo uso comunicativo del lessico. In accordo con Kühn (1990) e Schumacher (1998) osserviamo però che la definizione dei criteri orientati all'utente e il loro utilizzo comporti numerosi problemi, dovuti in particolare allo stato poco uniforme delle ricerche di pragmatica linguistica. Non è sempre possibile prevedere quali temi e situazioni saranno indispensabili per un determinato gruppo di discenti. Inoltre, più eterogeneo è il gruppo degli apprendenti, più difficile diventa la selezione, se al contempo si punta a un numero basso di elementi lessicali. Krohn (1992: 26) avverte che in molti vocabolari di base le concrete forme lessicali seguono il catalogo degli atti linguistici, di cui consegue che in molte liste abbondano le espressioni stereotipiche, atte ad affrontare determinate situazioni e temi prestabiliti. L'utilità di questi elementi in certe situazioni standard non ci illude sul fatto che la prevedibilità delle situazioni tipiche e delle forme linguistiche appropriate è limitata. La conoscenza di parole e significati generici, non legati a determinati temi e situazioni, è una condizione irrinunciabile per la costruzione e per l'ampliamento di una competenza lessicale.
È difficile valutare se i criteri orientati ai bisogni degli utenti e alla pragmatica abbiano prodotto lessici di base effettivamente migliori, fatto sta, che un vocabolario fondamentale basato su criteri pragmatici costituisce uno strumento di comunicazione più legittimo rispetto alle liste precedenti. In quest'ottica i vocabolari di base stesi utilizzando il solo criterio della frequenza, infatti, sono visti come prodotti "semilavorati."

1.2.2.4 Un approccio frequenziale-comunicativo

La lessicografia del periodo precedente ha dimostrato che i conteggi di frequenza non erano adatti ad offrire al discente un inventario lessicale delimitato rispetto al lessico infinito di una lingua, che gli permetta di valutare la quantità del materiale da imparare, di acquisirlo in un modo sistematico e di farsi capire in situazioni quotidiane. La svolta pragmatica nella definizione del lessico di base ha eliminato le lacune dovute alla selezione orientata alla frequenza lessicale, ponendo in primo piano l'utilità comunicativa degli elementi lessicali, tuttavia non ha definito chi e in quali situazioni adopera un dizionario di base. Ultimamente si cerca di integrare le due teorie opposte in un approccio frequenziale-comunicativo.
È fondamentale da questo punto di vista l'ampia monografia di Krohn (1992), in cui l'autore discute gli aspetti metalessicografici e didattici della struttura dei dizionari fondamentali tedeschi. Egli dimostra che l'approccio frequenziale e quello pragmatico, contrariamente a quanto affermato nel passato, hanno numerose caratteristiche in comune (ivi, pp. 55-71). L'obiettivo è di integrare i due approcci in un unico quadro frequenziale-

comunicativo. Il principio basilare è che per ottenere un lessico fondamentale attendibile dal punto di vista didattico, i dizionari di frequenza vanno integrati con dei temi e situazioni, mentre i dizionari basati sulla pragmatica non possono fare a meno dei dati di frequenza (ivi, p. 75). Sotto queste premesse Krohn fornisce una "lista compatta" che nasce dall'intersezione di sei liste di parole: due liste compilate secondo criteri comunicativo-pragmatici e quattro liste basate sulla sola frequenza. Per ottenere il minimo denominatore comune delle varie liste, sono stati ammessi solo lessemi che figurano in almeno quattro delle sei liste. Tra i 1.450 lemmi troviamo per lo più parole sinsemantiche (articoli, pronomi, preposizioni e congiunzioni, ossia parole grammaticali che esprimono relazioni sintattiche e/o semantiche, ma non hanno funzione referenziale autonoma) e autosemantiche generali (nomi, verbi, aggettivi che non possono essere associate, nel senso del *Frame Theory*, ad un determinato concetto/tema, ma compaiono in tutti i tipi di testi), mentre le parole autosemantiche legate a determinati temi sono in minoranza. Ciò si spiega con il fatto che le divergenze tra i diversi lessici fondamentali sono i minori nella classe chiusa degli elementi sinsemantici, aumentano nell'ambito dei lessemi autosemantici generici, e il più evidentemente si manifestano nella classe aperta degli elementi autosemantici legati a determinati temi (Krohn 1992: 109). In questo senso il lessico "compatto" rappresenta davvero il minimo denominatore comune, giacché comprende parole sinsemantiche e quelle parole autosemantiche che sia nella lingua standard, sia nei linguaggi speciali compaiono. Per ottenere un lessico fondamentale universale, si propone di recuperare le parole sinsemantiche più importanti tramite conteggi di frequenza; per le voci autosemantiche generali, invece, si dovrebbero ricorrere, oltre ai dati di frequenza anche ai criteri dell'approccio comunicativo-pragmatico. Le parole autosemantiche specifiche, infine, vanno selezionate in vista degli obiettivi didattici, facendo uso dei criteri comunicativi e pragmatici.

1.2.3 Tipi di lessici

Il "lessico di base" è un'espressione inflazionata. In primo luogo, sta ad indicare un lessico delimitato in qualche modo, eppure i limiti della selezione sono assai differenti: per l'italiano Katerinov *et al.* (1991) ha scelto 691 parole di base, per Migliorini (1943) esistono 1.500 elementi lessicali fondamentali, Giovanelli (1966) riporta ca. 4.500 termini, di cui 2.000 di base, De Mauro *et al.* (1980) ha scelto ca. 7.000 parole, il Piccoli (1986) e il VELI (1989) comprendono 10.000 vocaboli, il *Dizionario di base* (De Mauro e Moroni 1996) e il *Lessico elementare* (Marconi *et al.* 1994) includono ambedue ca. 15.000 lemmi. Perfino dizionari monolingui relativamente ampi, come il *Dizionario fondamentale della lingua italiana* dell'editore De Agostini (2005) con 40.000 lemmi, oppure il Devoto-Oli compatto (Le Monnier, 2002) con 24.000 voci affermano di contenere il lessico italiano di base.
In secondo luogo, l'espressione "lessico di base" è usata senza differenziazione, sia per dizionari fondamentali destinati a parlanti nativi, sia per quelli concepiti per apprendenti di una lingua straniera. Una distinzione tra lessici di base della lingua madre e dizionari didattici, pensati per l'apprendimento di una L2, è auspicabile perché i due gruppi target esigono dizionari con macro- e microstrutture differenti. Attualmente sono chiamati vocabolari di base tutti i "dizionari" – contenenti liste disposte in ordine alfabetico – destinati adi apprendenti di una lingua straniera. Schumacher (1998: 161) avverte, infatti, che la lessicografia dei dizionari di base in senso stretto indica la codificazione di lessici fondamentali per l'insegnamento di una L2.

La macrostruttura dei vocabolari di base è molto variegata. Kühn (1990: 1359) dà la seguente classificazione dei lessici di base:

- dizionari destinati all'insegnamento della L1
  - dizionari per bambini (De Mauro 1997b)
  - dizionari scolastici (Gabrielli 1981, De Mauro 1996)
- dizionari destinati all'insegnamento della L2
  - di uno dei linguaggi speciali
  - della lingua standard
    - plurilingui (Baldelli e Mazzetti 1974)
    - bilingui (Migliorini 1943)
    - monolingui
      - lemmi disposti in ordine alfabetico, con definizioni ed esempi (Cannella 1991) o senza (Sciarone 1977)
      - lemmi disposti in ordine tematico (Galli de' Paratesi 1981)
      - lemmi riportati in ordine di frequenza (Katerinov 1991)
      - lemmi riportati in ordine d'uso (Marconi *et al.* 1994)
      - lemmi disposti in gruppi di parole sinonimiche (Giovanelli 1966)

Alla classifica di Kühn abbiamo aggiunto altri due tipi: tra i vocabolari fondamentali destinati all'insegnamento della lingua materna (L1) conviene distinguere tra dizionari per bambini e dizionari scolastici. I primi sono destinati a bambini fino all'età di 8 anni, pertanto sono molto selettivi, infatti, il numero dei lemmi varia da 200 a 2.000 (Hausmann 1990: 1365), ed hanno una microstruttura particolare: si rinunciano a definizioni ed informazioni lessicografiche tradizionali, i lemmi di solito sono illustrati da esempi d'uso. Questo modo di definire è più simile al modo di definire del bambino che spesso definisce la parola portando esempi concreti, tratti dalla propria esperienza personale (Marello 1996: 125). I vocabolari scolastici, invece, sono dizionari veri e propri che pur contenendo dei riferimenti didattici, obbediscono alle convenzioni lessicografiche. I confini tra i due tipi non sono netti, e nemmeno i titoli dei dizionari sono indicatori affidabili. *Il mio primo dizionario illustrato* di Gabrielli (1981) per esempio, con i suoi 10.000 lemmi non rispetta i criteri di un dizionario destinato per bambini, infatti, è pensato per gli studenti delle elementari e per quelli della scuola media.

Il lessico riportato nei dizionari di base, pensati per l'insegnamento di una lingua straniera, può rappresentare la lingua standard, ma esistono anche vocabolari fondamentali, monolingui e plurilingui, delle lingue settoriali. I vocabolari possono essere monolingui, bilingui o plurilingui. I dizionari fondamentali plurilingui e bilingui si basano sul concetto rudimentale delle traduzioni equivalenti, e riportano per ogni lemma i rispettivi equivalenti nella lingua d'arrivo; i lemmi possono essere disposti in ordine alfabetico o tematico. I vocabolari monolingui si possono distinguerea seconda se riportano qualche tipo di spiegazione semantica (come in Steger 1972) o meno (Sciarone 1977). Kühn osserva a ragione che i vocabolari di base senza spiegazioni di significato andrebbero chiamati liste lessicali, mentre quelli con delle spiegazioni si dovrebbero chiamare dizionari di base. Nei vocabolari di base monolingui il lessico può essere riportato in gruppi tematici, in base alla frequenza, il valore d'uso o la sinonimia. In Giovanelli (1966) le parole di perfezionamento sono suddivise in 13 campi concettuali, raggruppati in 73 sottoinsiemi. Nei dizionari di base fondati sulla frequenza o valore d'uso, gli elementi lessicali aventi lo stesso indice di frequenza e/o valore d'uso sono riuniti in gruppi ed elencati in ordine alfabetico.

1.2.4 Problemi di applicazione

I compilatori e gli editori pongono l'accento sull'utilità dei vocabolari di base, sostenendo che conoscendo un lessico fondamentale di 2.000 elementi, uno può capire e farsi capire in una lingua straniera. Dal punto di vista statistico può capire ca. l'85% di un testo non specializzato e condurre dialoghi quotidiani, mentre conoscendo solo le 1.000 parole più frequenti, si capisce l'80% dei testi. Gli esperti (Kühn 1990, Schumacher 1998) avvertono che queste affermazioni servano solo a pubblicizzare i prodotti lessicografici, perché dal punto di vista scientifico non sono sostenibili, anzi dichiarazioni del genere conferiscono un carattere poco serio ed affidabile ai tentativi di estrazione dei lessici di base. Le considerazioni di sopra creano l'impressione errata che imparando un lessico ristretto, uno sia in possesso di una competenza comunicativa estesa.

Un altro problema cui allude il titolo, è che la gran parte dei dizionari fondamentali finora usciti non è utilizzabile come opera di consultazione, poiché in contrasto ai dizionari normali sono troppo riduttivi in merito al lessico contenuto e spesso non forniscono definizioni. Il *Livello soglia* (Galli de' Paratesi 1980), oppure il *Lessico elementare* (Marconi *et al.* 1994) sono due liste dedicate agli insegnanti, agli autori dei libri didattici, a chi pianifica i curricula, quindi, non sono adatti all'uso didattico. È impossibile, infatti, utilizzare una lista di parole in modo meccanico, come fonte da cui memorizzare i lemmi. Presentare le parole al di fuori di un testo rende più difficile l'apprendimento, perché le priva del loro contenuto linguistico, dato che il significato di una lingua cambia a seconda del contesto. Questi lavori fungono piuttosto come una specie di catalogo per chi si occupa di didattica linguistica. D'altra parte, con poche eccezioni, tra cui il DIB (1996) di De Mauro, i lessici fondamentali non sono adatti alle esigenze degli apprendenti, giacché il lessico non è ben disposto dal punto di vista didattico, infatti, non sono rappresentate le strutture lessicali semantiche, né i criteri della semantica contrastiva sono presi in considerazione, pertanto non possono assumere il ruolo di una tradizionale opera di consultazione. Infine, nella realtà linguistica non esistono i testi standard sterili e invariabili, ipotizzati dai lessicografi. Visto che i conteggi di frequenza nella grande maggioranza dei casi si basano su testi letterari, non sorprende quanto notato da Kühn (1990: 1360), cioè che negli articoli dei quotidiani ogni ottava parola non figura in nessuna delle liste di frequenza disponibili.

Kühn (1990: 1360) immagina il dizionario fondamentale del futuro come libro di studio e di esercitazione nello stesso tempo, il cui lessico è stato scelto in base a criteri pragmatici che permettono di farsi intendere in situazioni comuni. Per rendere osservabili e analizzabili anche i caratteri qualitativi della lingua, occorre presentare il lessico fondamentale tramite gruppi e/o campi lessicali e paradigmi. Grazie a questa struttura semantica ogni singolo gruppo di parole costituirebbe un insieme da studiare, permettendo di imparare il lessico in modo sistematico, tenendo conto anche delle relazioni semantiche. Analogamente il *Testlexikon* proposto da Krohn (1992: 143) avrebbe lo scopo di sincronizzare in modo sistematico il vocabolario individuale dell'apprendente con la quantità di elementi lessicali da imparare.

Un altro problema è, per l'applicazione didattica, che le liste approntate per la ricerca scientifica non sono immediatamente utilizzabili nell'insegnamento. Dato che le parole più frequenti sono anche le più polisemiche, per essere veramente utili, le statistiche dovrebbero riferirsi ai significati, invece delle forme. I calcoli di frequenza, infatti, per ora non dicono quale dei molti sensi della parola è il più frequente. Per fare qualche esempio, nel *Grande dizionario italiano dell'uso* (De Mauro 1999) in *dare* si distinguono 11 accezioni, in *andare* 14, in *fare* 30. Con le accezioni secondarie si arriva ai massimi di

tenere con 41, *fare* con 43 e *prendere* con 48. Quali di queste accezioni dovrebbero entrare nel bagaglio lessicale dell'apprendente? Esistono ormai software che calcolano la co-occorrenza statistica dei termini, cioè la probabilità che hanno due parole di essere associate in un determinato testo. Tale calcolo consente di trovare l'accezione con cui è più probabile incontrare una parola in un testo. Tuttavia non esiste ancora per l'italiano un dizionario didattico, come il COBUILD inglese, che sulla base dello spoglio di un corpus vastissimo ordina le accezioni di un lemma polisemico secondo le rispettive frequenze (cfr. sez. 4.1.4). Un'altra questione è che al lessico appartengono espressioni idiomatiche e formule fisse non scomponibili, che non sono sempre introdotte organicamente nelle liste di frequenza.
Quanto alla differenza tra vocabolario attivo (produttivo) e passivo (ricettivo), le esperienze rivelano che non c'è un'opposizione netta tra conoscenza ricettiva e produttiva del lessico. Nel corso dell'apprendimento determinate parole da vagamente familiari diventano note, e in stadi successivi si impara a usarle con sempre maggiore proprietà (Corda e Marello 1999: 27). In alcuni vocabolari fondamentali (Galli de' Paratesi 1981) sono segnalate le parole che l'allievo deve conoscere produttivamente; ma esaminando più in dettaglio queste liste, spesso si vede che l'etichetta "conoscenza produttiva" è data semplicemente in base a un criterio di frequenza, mentre non si presta particolare attenzione al consolidamento del lessico da usare in modo produttivo. Nella maggior parte dei manuali d'italiano L2 non c'è una distinzione tra lessico da imparare produttivamente o ricettivamente: gli autori suppongono, infatti, che le parole conosciute solo ricettivamente agli inizi del corso vengano man mano imparate anche produttivamente.
L'individuazione dei lessici fondamentali ha avuto una certa utilità nella pianificazione dei curricoli didattici. La pianificazione sempre più differenziata dei contenuti causa scelte lessicali sempre più specifiche, basate sui bisogni dei vari gruppi di destinatari (soprattutto adulti). Occorre però sempre tenere presente che un vocabolario di base non è concepito per limitare l'insegnamento a quel nucleo lessicale, né significa che occorra prima insegnare tutte le parole di una prima fascia, poi – e soltanto poi – tutte le parole di una seconda fascia. La delimitazione imposta da un vocabolario di base significa soltanto di imparare dapprima il maggior numero delle parole scelte, ricorrendo alle parole di una fascia seguente soltanto quando è necessario per l'espressione spontanea o quando il contenuto di un testo presentato lo richiede.

1.3 Quali e quante sono le parole fondamentali?

In vista dei suddetti, sembra che gli studiosi non siano ancora riusciti a delimitare un metodo di ricerca che dia tutte le garanzie per riprodurre esattamente il fatto linguistico che si vorrebbe circoscrivere. I diversi metodi di ricerca danno risultati a volte sensibilmente diversi, pertanto le mille, duemila parole più usuali sembrano cambiare forma secondo i criteri di selezione adottati. Giulio C. Lepschy (1978: 59) ha espresso una critica severa nei confronti dei lessici fondamentali:

> "Un dizionario di frequenza per principianti che volesse limitarsi alle 5.000 parole più comuni, più usuali dell'italiano moderno, sarebbe diverso se si basasse sul LIF o sul FDI, e in tutti e due i casi sarebbe piuttosto insoddisfacente. Intanto, sembra ragionevole basare la scelta per i materiali didattici sul buon senso, piuttosto che sui calcoli statistici."

I risultati dei conteggi di frequenza possono presentare delle stranezze, basta pensare agli elementi delle serie chiuse, come i giorni della settimana, ecc. È facile correggere queste assenze, però emerge un dubbio: se la frequenza non riflette l'utilità degli elementi in questo caso, perché dovrebbe farlo in altri? Quanto è efficace basare l'apprendimento lessicale su criteri statistici?
Per garantire l'attendibilità della selezione, bisogna tener conto di numerosi criteri extra-statistici: il problema d'interferenza della struttura lessicale della madrelingua, la necessità di includere elementi utili in classe, scegliere parole che si adottano alla metodologia, escludere quelle che presentano particolari difficoltà. C'è inoltre l'esigenza di conciliare l'insegnamento del lessico con quello della grammatica e della pronuncia. Quindi, nello stabilire l'elenco delle parole più utili per chi impara l'italiano, il criterio quantitativo va integrato con quello qualitativo, cioè con l'esperienza dell'insegnante.
Cercando di dare una risposta all'interrogativo che funge da titolo a questo paragrafo, si può dire che sia la quantità, sia la qualità delle parole che costituiscono il lessico fondamentale varia secondo le esigenze e dell'età degli apprendenti, ma le 2.000-3.000 parole risultanti come più frequenti dallo spoglio di corpora bilanciati sono sicuramente quelle fondamentali per comunicare in una comunità, siccome permettono di riconoscere ca. il 90% del lessico usato nei testi. Certo riconoscere il 90% delle parole non significa necessariamente capirle al 90%, ma in ogni caso costituisce un buon punto di partenza. Riguardo alla questione, per quanto tempo occorra attenersi al vocabolario di base, non esiste una risposta valida per tutti i casi. La risposta dipende da numerosi fattori, tra cui il tempo disponibile, la lingua insegnata, la parentela di questa lingua con quella materna degli allievi, l'età degli allievi, l'estensione richiesta del vocabolario e certo le finalità dell'insegnamento. A titolo d'esempio, Sgroi (1982: 107) ammette che un corso di sei anni dovrà comprendere quattro anni dedicati soprattutto (ma non solo) allo studio di un vocabolario di base di ca. 3.000 parole.
Circa la qualità di queste 2.000-3.000 parole siamo concordi con Sgroi (1994: 237) e Sciarone (1977: 101) sul fatto che le liste di frequenza realizzate con il supporto dell'analisi statistica su corpora, anche se criticabili, possono essere utilizzate per un insegnamento razionale se usate in modo non esclusivo, ma arricchite da vari apporti lessicali e semantici suggeriti dalle concrete situazioni e dagli interessi degli allievi.

2 I dizionari fondamentali e di frequenza della lingua italiana (1927-2006)

Per la redazione di questo capitolo ci siamo serviti di due saggi di S. C. Sgroi (1981, 1994) in cui l'autore passa in rassegna i lessici fondamentali e di base dell'italiano, apparsi nel periodo 1927-1980 e 1927-1993 rispettivamente. Per una breve rassegna in lingua ungherese cfr. Fóris (2000: 125-143, 2002: 28-35, 81-99).
Scopo di questo capitolo è di analizzare i criteri qualitativi e quantitativi adottati dai vari autori per la determinazione del lessico di base, e di reperire quelle opere che potrebbero fungere da strumenti di riferimento per la selezione del lessico fondamentale dell'italiano L2.

2.1 Autori, finalità e criteri di selezione

Selezionare le parole più frequenti o quelle che fanno parte di un vocabolario di base, è un'operazione delicata che in Italia è stata affrontata prevalentemente sulla scia di studi simili condotti per l'inglese e il francese. Sgroi (1994: 209) osserva che i primi vocabolari fondamentali e di frequenza dell'italiano sono opere di studiosi stranieri (Thompson 1927, Knease 1931, Skinner 1935, Juilland-Traversa 1973, Reiske 1974, Sciarone 1977) o di italiani operanti all'estero (Migliorini 1943, G. A. Russo 1947, J. A. Russo 1962, Giovanelli 1966) e solo recentemente di italiani in Italia (Bortolini *et al.* 1971, Baldelli-Mazzetti 1974, D.I.L. 1978, De Mauro *et al.* 1980, Gabrielli 1981, Galli de' Paratesi 1981, Batinti-Trenta 1983, Piccoli 1986, VELI 1989, Cannella 1991, Katerinov *et al.* 1991, De Mauro *et al.* 1993 ecc.). Questi testi apparsi tra il 1927 e il 2006 – nell'arco di ottant'anni – sono una trentina, e per gran parte risalgono all'ultimo ventennio.[1]
Molte liste sono state elaborate per esigenze pratiche: per rendere più efficace l'insegnamento dell'italiano come lingua straniera (ad es. Thompson 1927, Knease 1931, Skinner 1935, Migliorini 1943, G. A. Russo 1947, J. A. Russo 1962, Giovanelli 1966, Reiske 1974, Baldelli-Mazzetti 1974, Sciarone 1977, Galli de' Paratesi 1981, Batinti-Trenta 1983, Katerinov *et al.* 1991, Baldelli *et al.* 2005) e solo recentemente per l'italiano come lingua madre (D.I.L. 1978, De Mauro *et al.* 1980, Gabrielli 1981, Piccoli 1986, Cannella 1991, De Mauro *et al.* 1993, Marconi *et al.* 1994, Putzer 1998a, 1998b, Rinaldi *et al.* 2004); più raramente per esigenze non specificatamente glottodidattiche (Bortolini *et al.* 1971, Juilland-Traversa 1973, Alinei 1973, VELI 1989, CoLFIS 1995). Le prime liste sono state compilate in modo manuale (Thompson 1927, Knease 1931, Skinner 1935, Migliorini 1943, Russo 1947), mentre quelle pubblicate dagli anni Settanta in poi (Bortolini *et al.* 1971, Juilland-Traversa 1973, Alinei 1973, Sciarone 1977, De

[1] È interessante che tra la realizzazione del primo vocabolario fondamentale della lingua italiana (Thompson 1927) e la nascita del primo lessico fondamentale compilato da un italiano (Migliorini 1943) sono passati 16 anni, quindi gli studiosi italiani hanno riconosciuto relativamente tardi l'importanza di tali lessici.

Mauro *et al.* 1980, VELI 1989, Katerinov *et al.* 1991, De Mauro *et al.* 1993 ecc.) sono state compilate con l'ausilio di calcolatori elettronici.
I criteri della composizione variano a seconda gli obiettivi delle opere: la scelta degli elementi si basa ora sulla sola esperienza dei compilatori (così Migliorini 1943, Giovanelli 1966, Baldelli-Mazzetti 1974, Reiske 1974, D.I.L. 1978, Gabrielli 1981, Galli de' Paratesi 1981, Piccoli 1986, Cannella 1991, Savino 1995, Putzer 1998a, 1998b), ora sull'esperienza e la dispersione (J. A. Russo 1962, Skinner 1935), ora sulla frequenza e dispersione (Thompson 1927, Knease 1931, Bortolini *et al.* 1971, Juilland-Traversa 1973, VELI 1989, De Mauro *et al.* 1993), ora sulla frequenza, dispersione ed esperienza (G. A. Russo 1947, De Mauro *et al.* 1980, Batinti-Trenta 1983, DIB 1996, De Mauro-Moroni 1997b, DAIC 1997, Baldelli *et al.* 2005) oppure sulla sola frequenza (Sciarone 1977, Katerinov *et al.* 1991, CoLFIS 1995, Cresti-Moneglia 2004, Rinaldi *et al.* 2004) come vedremo di seguito.
Va ricordato ancora che quasi tutti i dizionari riguardano l'italiano scritto, mentre solo cinque (Galli de' Paratesi 1981, Katerinov 1991, De Mauro *et al.* 1993, Cresti-Moneglia 2004, Rinaldi *et al.* 2004) si riferiscono alla lingua parlata.

2.2 Cronologia dei lessici fondamentali e di frequenza

Nella presentazione dei lessici si segue il criterio cronologico (invece di quello dell'estensione, composizione, ampiezza del corpus ecc.), poiché esso rappresenta meglio delle altre soluzioni, come cambiavano nel tempo i vari parametri della compilazione con lo sviluppo dei metodi di ricerca, man mano che gli studiosi riuscivano a fornire descrizioni sempre più fondate e scientifiche del lessico.
Molti sono i dizionari che mirano ad individuare il lessico "fondamentale" dell'italiano, non tutti però sono dizionari di base nel senso rigoroso del termine. Tralasciamo quindi nella nostra analisi le edizioni ridotte dei vocabolari d'uso.[2] Non è il caso di entrare nel merito dei dizionari inversi, spesso basate su liste di frequenza,[3] e non esaminiamo le concordanze e i lessici di frequenza tratti da opere letterarie, ad eccezione del lavoro di

[2] Zingarelli N. (2001) *Lo Zingarelli minore. Vocabolario della lingua italiana*, Bologna, Zanichelli, pp. 1562; *Il nuovo dizionario italiano Garzanti* (1984) Milano, Garzanti, pp. 1087; Palazzi F. (1982) *Il piccolo Palazzi dizionario della lingua italiana*, Milano, Hoepli, pp. vi-986; Melotti F., Sordi A. (a cura di) (1988) *Dizionario italiano. Le 50.000 parole della lingua di oggi,* Milano, Rizzoli, pp. 1193; D*izionario Larus della lingua italiana* (1990) Bergamo, Larus, pp. 1302; Sgarbi V. (1993) *Lo Sgarbino, Dizionario della lingua italiana*, Bergamo, Larus, pp. viii-1412; Cattana A., Nesci M. T. (a cura di) (1995) *Italiano junior. Dizionario della lingua italiana*, Torino, Loescher, pp. 1114; *Dizionario fondamentale della lingua italiana* (2005) Novara, De Agostini, pp. 1100; Devoto G., Oli G. C. (2002) *Dizionario Devoto-Oli compatto. Dizionario fondamentale della lingua italiana,* Firenze, Le Monnier, pp. 1568. Ad un livello notevolmente inferiore troviamo: Miot G. (1980) *Il mio primo dizionario*, Firenze, Giunti, pp. vii-882; Martignon R. (1987) *Dizionario pratico della lingua italiana,* Milano, Mondadori, pp. xviii-998; Grassi C. (1982) *Dizionario della lingua italiana*, Milano, Garzanti, pp. 554; Colombo P. (1989) *Dizionario della lingua italiana*, Milano, Capitol, pp. 540; Valle G. (1979) *Piccolo dizionario della lingua italiana*, Brescia, La Scuola, pp. 544.
[3] Il primo e per lungo tempo l'unico lavoro di questo genere per l'italiano è il *Dizionario inverso italiano* di Alinei (1962). Le 43.506 forme sono state tratte dal *Prontuario etimologico della lingua italiana* di Migliorini e Duro (Torino, Paravia, 1950). A questa monografia pionieristica si affianca il *Dizionario italiano inverso* di Giuliano Merz (disponibile on line sul sito http://culturitalia.uibk.ac.at/wb/WB_it.asp), contenente 129.432 lemmi tratti da *Il dizionario della lingua italiana per il terzo millennio* di De Mauro (2000).

Alinei (1973) che doveva servire come base per la costruzione di un dizionario di frequenza per l'italiano contemporaneo.
Per ogni dizionario si forniscono i dati del corpus sottostante, i criteri di selezione, il numero di lemmi inclusi, i gruppi di destinatari ed altri dati rilevanti al fine di poter dare un panorama delle divergenze e le convergenze delle liste in esame.

2.2.1 Thompson M. E. (1927) *A Study in Italian Vocabulary Frequency*, University of Iowa Master's Thesis.

La lista della Thompson, destinata a studenti principianti d'italiano, è una tesi non pubblicata, ed è circolata fra studiosi che hanno condotto studi simili. Comprende 500 vocaboli, ricavati da un campione presunto di 100.000 occorrenze, formato da 10 manuali di lettura per stranieri, pubblicati negli Stati Uniti. I conteggi sono stati effettuati manualmente ogni dieci pagine di testo, sono stati ignorati gli articoli, pronomi, preposizioni, congiunzioni, aggettivi dimostrativi e possessivi, numerali, ma anche i nomi dei giorni, dei mesi e delle stagioni, i nomi propri e geografici e le interiezioni di meno di quattro lettere. Ai lemmi non si assegna una categoria grammaticale, tranne i c.d. "omonimi sintattici" (*bene* avv. e *bene* n.), ma in alcuni casi manca l'etichetta: *diritto* (sostantivo, aggettivo o avverbio?), *dolce* (sostantivo o aggettivo) ecc. Non è chiaro inoltre, quale significato delle parole polisemiche (*nome, paese, opera*) è incluso, poiché l'opera non riporta i traducenti inglesi.
Sono stati inclusi i lemmi con una frequenza non inferiore a 30. Oltre all'indice di frequenza, l'autrice ricorre anche al valore del c.d. *weighted total*, una sorta di "uso" che si calcola combinando la frequenza con la dispersione.[4] La prima parte dell'opera contiene i 500 lemmi ordinati per frequenza decrescente, con indicazione della categoria grammaticale solo quando è necessario per distinguere tra omonimi sintattici; la seconda parte fornisce una lista alfabetica dei lemmi.

2.2.2 Knease T. M. (1931) *An Italian Word List from Literary Sources*, The State University of Iowa, Toronto, Toronto U. P.

Quattro anni dopo i conteggi pionieristici della Thompson, la Knease ha difeso un lavoro più ambizioso alla stessa università. Mentre l'elenco della Thompson era destinato a studenti principianti d'italiano, la Knease si concentrava sui bisogni dei studenti a livello intermedio, puntando a una lista di "some two thousand words" (p. 13), più precisamente 2.080 secondo G. A. Russo (1947: 218), ma 2.097 secondo Juilland-Traversa (1973: XI). Si basa su un campione di 400.000 occorrenze, la stessa dimensione che Henmon trovò soddisfacente per *A French Word Book* (1924). Il corpus include racconti, commedie, poesie, critiche di 40 autori letterari dal 1850 al 1930, selezionati in base ad una lista pubblicata nella rivista *Italica* nel 1925.[5] I conteggi sono stati effettuati manualmente ogni cinque pagine di testo, e, sulla scia della Thompson, sono state

[4] "The weighted total, an interesting but questionable attempt to combine frequency and dispersion, was established by multiplying the frequency by the square root of the square nearest to the number of sources in which the word occurred. For instance, *guardare* had 289 occurrences and was found in ten sources; 289 was multiplied by 3, the square root of ten to the nearest integer, for a weighted total of 867" (Juilland-Traversa 1973: X).
[5] A Five Foot Shelf on Modern Italian Literature. In: *Italica*, August 1925, Vol. 2, No. 3.

omesse le parole grammaticali, le parole straniere, i numerali e i nomi geografici. La categoria grammaticale è precisata solo per gli omonimi sintattici.
I lemmi sono compresi fra la dispersione e frequenza massima (40 e 9.600 risp.) di *essere*, e dispersione e frequenza minima (9 e 9) di *augurare*. Alla lista in ordine decrescente di dispersione e frequenza (pp. 15-41) segue una lista in ordine alfabetico dei lemmi (pp. 42-70). Ogni lemma è seguito dal coefficiente di dispersione (numero delle fonti in cui la parola occorre) e di frequenza.[6]

2.2.3 Skinner L. H. (1935) A Comparative Study of the Vocabulary of Forty Five Italian Textbooks. In: *Modern Language Journal*, XX, November, 2, pp. 67-84.

La lista di Skinner è tratta da un campione di 19.800 voci di partenza ("root words" p. 69), formato con i glossari di 20 grammatiche, manuali e libri di testo e 25 antologie usate nelle scuole americane per l'insegnamento dell'italiano. Invece di contare le occorrenze di ogni singola parola, l'autore prende in considerazione i vocaboli inclusi nel glossario di ogni libro, scelta che incide negativamente sulla validità dei risultati, siccome in genere gli editori non concordano su quali parole includere nel glossario. Pertanto il verbo *essere*, che ovviamente appare in ciascuno dei 45 fonti, ha il rango 41 (Juilland 1973: XI). A causa del metodo utilizzato, le 2.962 parole fondamentali (le voci grammaticali sono state parzialmente omesse) sono state scelte in base alla dispersione ("range") compresa fra 45 e 10. Notiamo che la dispersione non è un criterio sufficientemente rigoroso, se l'universo linguistico del campione analizzato non si presenta significativamente ampio e vario. Il campione di Skinner però, costituito da grammatiche e antologie presenti nelle scuole americane, difficilmente può essere considerato rappresentativo.
Le traduzioni inglesi e le categorie grammaticali accompagnano i lemmi solo quando richiesto per la distinzione degli omonimi. Manca una lista alfabetica comprensiva di tutti i termini.

2.2.4 Migliorini B. (1943) *Der grundlegende Wortschatz der Italienischen. Die 1500 wesentlichsten Wörter*, Marburg, Elwert.

Migliorini (pp. 2-31) dà una lista di 1.500 voci fondamentali (*die wesentlichsten Wörter*) comprensivi delle parole grammaticali. La lista, pensata per scopi glottodidattici, è stata composta unicamente sulla base del senso linguistico dell'autore. Juilland (1973: XI) tuttavia assume che le scelte dell'autore siano state guidate da conteggi parziali di frequenza. Nell'opera, fra i 500 lemmi più importanti si trovano molte voci legate alla situazione politico-sociale dell'epoca (*balilla, corporativo, aviere, duce*) ed altre che rispecchiano aspetti stilistici e lessicali in decadenza.[7]

---

[6] Da un raffronto sommario della lista di Knease con il LIF risulta che nel primo siano presenti parole di colorito nettamente letterario che non compaiono nel LIF. Bortolini (1971: 8) riporta fra l'altro le seguenti voci, assenti nel LIF: *affogare, agonia, alito, allegramente, ambedue, ardore, arrischiare, assiduo, astrarre, avvedere, balenare, balzare, berretto, borbottare, brontolare, celare* ecc.
[7] Non compaiono nel LIF: *accampamento, affogare, annuncio, ardore, artiglieria, atterrire, automobilista, aviazione, aviere, avvedersi, balilla, codesto* ecc. (Bortolini 1971: 8).

I 500 lemmi più importanti, raccomandati per gli studenti del primo anno, sono stati stampati in grassetto. Le voci sono seguite dalla categoria grammaticale e dal loro corrispondente tedesco con qualche esempio. I derivati si trovano sotto il lemma principale. Accanto a quasi ogni lemma si trova la trascrizione della sua pronuncia con i segni dell'Alfabeto Fonetico Internazionale.

2.2.5 Russo G. A. (1947) A combined Italian Word List. In: *Modern Language Journal*, XXXI, April, pp. 218-240.

La lista comprende 3.173 parole disposte in ordine alfabetico (pp. 220-240). Il corpus di 5.043 parole è costituito dalla somma del lessico di Knease (1931), ricco, secondo Russo di 2.080 termini (p. 219), anziché 2.097, e di Skinner (1935), comprendente, sempre secondo Russo, 2.963 voci, anziché 2.962 (cfr. sopra). Dalla somma delle due liste sono state sottratte le 1.870 parole in comune, così sono stati ottenuti i 3.173 termini.
Ogni termine è seguito da un c.d. "merit number" che indica se esso è presente in tutte e due le liste o in una sola, e dal grado di difficoltà che indica l'appartenenza a uno dei 6 sottogruppi di dispersione decrescente, in cui sono state suddivise le due liste. Dal confronto emerge che 909 voci (dei 1.870 termini in comune) hanno lo stesso "range", mentre 1.303 parole sono esclusive di Knease o di Skinner, in particolare 210 di Knease e 1.093 di Skinner.

2.2.6 Russo J. A. (1962) An Italian Idiom List. In: *Modern language Journal,* XLVI, January, pp. 14-21.

A differenza di tutti gli altri testi quello di J. A. Russo fornisce una lista di locuzioni varie (sintagmi, modi di dire, ecc.) e non già di singole parole (ridotte ad una decina). I 452 "idioms" sono stati tratti da un campione di 1.920 espressioni. Il corpus è formato da testi per l'insegnamento dell'italiano: "novels, plays and short stories" (elencati a p. 21), usati nelle scuole americane, risalenti al 1940. Le locuzioni sono state selezionate in base alla dispersione ("range"), cioè in base alla loro presenza nei 25 testi spogliati, e, siccome nessuna locuzione è presente in tutti i testi, la dispersione varia da 1 a 21. Gli idiomi sono stati ordinati in base alla dispersione (pp. 15-17) e poi alfabeticamente (pp. 17-21), seguiti dall'indice di dispersione.[8]

2.2.7 Giovanelli P. (1966) *Grund- und Aufbauwortschatz. Italienisch,* Stuttgart, E. Klett.

Questo dizionario pubblicato in Germania, include 4.500 unità seguite dal loro corrispondente tedesco e da pochi esempi. I lemmi sono ripartiti in tre gruppi: l'autore ha identificato 74 parole grammaticali (pp. 13-14); 2.000 parole di base (*Grundwortschatz*) con cui si può comprendere l'85% di un testo non tecnico (pp. 15-108), e 2.500 parole di perfezionamento (*Aufbauwortschatz*) con coi si potrà comprendere fino al 95% di un

[8] Sgroi (1994: 209) osserva che G. A. Russo e J. A. Russo (Giuseppe e Joseph, rispettivamente) sembrano essere la stessa persona, dato che nella nota 7 di G. A. Russo (1947: 220) si legge che l'autore intende pubblicare una lista degli idiomi italiani, uscita 15 anni dopo col titolo *An Italian Idiom List* (Russo J. A. 1962).

testo normale (pp. 109-212). Le parole sono state selezionate in base all'esperienza dell'autore.
I vocaboli di base sono disposti in ordine alfabetico, mentre le parole di perfezionamento sono suddivise in 13 campi concettuali, raggruppati in 73 sottoinsiemi (quantità, tempo, universo, uomo ecc.). Per ogni sottoinsieme l'autore riprende anche i relativi termini del lessico di base, omettendo però il corrispondente tedesco.
La distinzione fra lessico fondamentale e lessico di perfezionamento, e soprattutto la scelta di raggruppare i termini in campi concettuali, rendono il dizionario di Giovanelli molto simile ai dizionari pedagogici più moderni.

2.2.8 Bortolini U., Tagliavini C., Zampolli A. (1971) *Lessico di frequenza della lingua italiana contemporanea*, Milano, Garzanti-IBM. (LIF)

È cronologicamente la prima realizzazione lessicografica mediante calcolatore della lingua italiana che vede la luce dopo la pubblicazione di due opere importanti: *Frequency Dictionary of Spanish Words* (Juilland *et al.* 1964) e *Frequency Dictionary of Rumanian Words* (Juilland *et al.* 1965), due opere che hanno posto le basi dei moderni conteggi di frequenza (v. 2.2.9).
Il LIF ha uno scopo duplice: da un lato costituisce uno strumento di lavoro per lo studio della lingua italiana contemporanea, dall'altro ha un interesse pratico per l'insegnamento dell'italiano sia come lingua materna, sia come lingua straniera. La lista registra 5.356 lemmi (comprensivi delle parole grammaticali), risultanti i più frequenti fra i 15.750 lemmi tratti dallo spoglio di 500.000 occorrenze. Il corpus è costituito da testi apparsi fra il 1947 e il 1968, ripartiti in 5 sottoinsiemi di dimensione uguale: copioni teatrali. Romanzi, sceneggiature cinematografiche, giornali (quotidiani e periodici), sussidiari per la scuola elementare.
L'uso (U) è il valore in base al quale i lemmi sono accettati o esclusi dal vocabolario di frequenza. Al fine di garantire la confrontabilità dei dati, il valore d'uso (U) è stato calcolato in base alla formula di Juilland e Chang Rodriguez, combinando la frequenza (F) con la dispersione (D): U = F x D (Bortolini *et alii* 1971: 32). Sia l'uso che il coefficiente di dispersione sono stati proposti per la prima volta da Juilland e Chang Rodriguez nel 1964 (cfr. 1.1.3.3.2). I 5.356 lemmi sono compresi fra uso massimo 45.041,03 dell'articolo *il* ed uso minimo 1,78 di voci come *abbrustolire, abusare*, ecc. In più, i lemmi inclusi sono presenti in almeno 3 dei 5 sottoinsiemi esaminati.
In questo modo i 1.500 lemmi più frequenti garantiscono una "copertura" dei testi pari al 91,13% (p. 57), mentre con i primi 2.500 lemmi la copertura arriva al 95,5%. Gli autori avvertono che, a causa della grandezza del campione, solamente la frequenza dei primi 1.000-1.500 lemmi può essere considerata stabile, poiché gli elementi di più basso rango possono variare secondo la composizione del corpus. Il volume include:

a) Una parte principale in cui i lemmi e le forme relative sono ordinati alfabeticamente (pp. 99-716). Per ogni lemma è specificata la categoria grammaticale, per ogni lemma e forma flessa sono precisati: l'indice di frequenza per sottoinsieme, il valore di frequenza totale nei cinque sottoinsiemi, il coefficiente di dispersione e il rango d'uso.
b) In una seconda lista i soli lemmi (non le forme) sono riportati in ordine d'uso decrescente (pp. 719-852), seguiti dalla categoria grammaticale, l'indice d'uso, il rango, ovvero il numero progressivo corrispondente alla posizione di ogni lemma in

questo ordinamento, e i due ranghi che competono al lemma nell'ordinamento per frequenza e dispersione decrescenti.

Notiamo che il LIF è stato il primo grande progetto di costituzione di una lista di frequenza, non tagliato su un singolo autore o su testi specificatamente letterari, in più, è stato il primo dizionario di frequenza realizzato con l'ausilio di calcolatori elettronici. Il LIF serve come base per la costruzione del *Vocabolario di base* di De Mauro *et al.* (1980).

2.2.9 Juilland A., Traversa V. (1973) *Frequency Dictionary of Italian Words,* The Hague, Mouton. (FDI)

Il volume è stato pubblicato nell'ambito del cosiddetto *Romance Languages Project* dell'Università di Pennsylvania, diretto da A. Juilland, che mirava a fornire una descrizione dei lessici di tutte le lingue romanze. Il dizionario per l'italiano vede la luce dopo la pubblicazione delle liste per lo spagnolo (1964), rumeno (1965) e francese (1970).

La lista per l'italiano contiene 5.014 parole, tra cui anche quelle grammaticali. Il corpus è di 500.000 occorrenze, tratte da testi pubblicati tra il 1920 e il 1940, raggruppati in 5 sezioni di 100.000 parole ciascuno: opere teatrali; romanzi e novelle; saggistica (saggi, memorie, corrispondenze); periodici (giornali e riviste); testi tecnici e scientifici (pp. XII-XVIII).

Le voci sono comprese fra uso massimo 24.798,22 e frequenza massima 28.659 della preposizione *di* ed uso minimo 2,39 di *ascensione, calce*, ecc. e dispersione minima 4,88 di *strato*. Il limite inferiore è quindi la frequenza superiore a 4, l'indice di uso superiore a 3, e l'occorrenza in almeno 3 sottoinsiemi (p. XLVI).

La dispersione è calcolata secondo la regola introdotta da Juilland e Chang Rodriguez nel *Frequency Dictionary of Spanish Words* (1964: XLV-LV), cioè in base al rapporto tra la frequenza media nei 5 sottoinsiemi e lo scarto dalla media delle frequenze realmente trovate. Con 5 sottoinsiemi e una frequenza totale di 10, abbiamo una media di 2 per sottoinsieme. Con una distribuzione di 22222 lo scarto è di zero, la distribuzione è considerata perfetta, avrà il valore 1, mentre nel caso di una distribuzione meno uniforme l'indice tende verso zero.[9]

È merito degli autori di aver introdotto, sempre nell'elaborazione della lista dello spagnolo, l'indice d'uso, ossia un valore che nasce dalla combinazione della frequenza e distribuzione. L'assunzione di base è che le frequenze distribuite in modo uniforme sono "dense", nel senso che non diminuiscono al crescere del corpus, mentre le frequenze mal distribuite sono "gonfie" e in un corpus più rappresentativo tendono a ridursi. Alla correlazione tra frequenza e distrobuzione Sciarone (1977: 59-87) dedica l'intero capitolo quarto del suo libro, giungendo alla conclusione che più una parola è frequente, meglio è distribuita, pertanto il valore d'uso non serve per ordinare le parole secondo un criterio d'importanza. Gli stessi autori riconoscono, infatti, che più numerose sono le occorrenze, più uniforme è la loro dispersione, quindi l'aggiustamento causato dal valore d'uso sarà minore (le prime 100 parole in ordine di frequenza decrescente sono identiche alle prime 100 parole in ordine d'uso decrescente); mentre più bassa è la frequenza e meno uniforme la distribuzione, l'aggiustamento sarà proporzionalmente maggiore (Juilland 1973: XLIV-XLV). Ciò suggerisce che il coefficiente d'uso è rilevante per prevedere

[9] Secondo Sciarone (1977: 61) il punto critico è che la formula non prende in considerazione il numero dei sottoinsiemi in cui la parola è presente, anche se questo fatto sembra importante per la determinazione della distribuzione.

l'occorrenza dei termini di frequenza media e bassa. La fortuna dell'indice è dimostrata dal fatto che lessici prestigiosi come il LIF, il LIP o il LE hanno scelto il valore d'uso in base al quale espungere o immettere i lemmi nell'elenco. Il dizionario consiste di due parti:

a) Elenco alfabetico (pp. 1-402) dei lemmi (stampati in neretto) e le rispettive forme. Per ogni lemma si precisano l'indice d'uso, frequenza e dispersione, il numero totale delle occorrenze e il numero delle occorrenze nei singoli sottoinsiemi. Per ciascun forma si indicano le occorrenze totali e le occorrenze nei sottoinsiemi.
b) La seconda parte comprende tre liste parallele in cui i soli lemmi (non le forme) sono ordinati in tre colonne: in ordine d'uso, frequenza e dispersione decrescente, e sono divisi in gruppi da 500 lemmi (pp. 403-519).

2.2.10 Alinei M. (a cura di) (1973) *Spogli elettronici dell'italiano letterario contemporaneo*, Bologna, Il Mulino. (SEILC)

Il SEILC va segnalato perché il progetto, che continua i 20 volumi di *Spogli elettronici dell'italiano delle origini e del duecento* a cura dello stesso Alinei (1968-1978), mirava a dare "la possibilità di realizzare un dizionario di frequenza dell'italiano letterario contemporaneo, che risulti dalla fusione dei diversi spogli del SEILC" (vol. I, p. VII). Il progetto prevedeva lo spoglio di altri "venti, quaranta romanzi" per "raggiungere un totale di circa due milioni di parole" (vol. I, p. VIII). I tre volumi pubblicati, riguardanti Moravia (*La ciociara*, 1957), Calvino (*Il sentiero dei nidi di ragno,* 1947) e Cassola (*Ferrovia locale,* 1968), sono strutturati in modo identico. Ognuno contiene:

a) il "formario", ossia una concordanza col rinvio alla pagina e alla riga; le diverse forme disposte in ordine alfabetico, comprendendo oltre all'indicazione del luogo del contesto, la frequenza assoluta e relativa se non inferiore a 0,01;
b) l'indice inverso in 4 colonne delle sole forme;
c) tre liste in ordine decrescente di frequenza assoluta, relativa e relativa cumulativa delle terminazioni in base all'ultimo trigramma, l'ultimo digramma e l'ultimo monogramma;
d) la lista in ordine decrescente di frequenza assoluta, relativa e relativa cumulativa delle forme con somma totale delle occorrenze dei tipi lessicali;
e) la tabella di distribuzione delle frequenze, che costituisce la "sintesi numerica della Lista di Frequenza delle Forme" (vol. I, p. XV).

Purtroppo il progetto non è stato portato a termine.

2.2.11 Reiske H. (1974) *VHS-Italienisch Unterricht und VHS-Zertifikat Italienisch*, Frankfurt, Pädagogische Arbeitsstelle-Deutscher Volkhochschulverband.

Il volume, preparato e pubblicato in Germania per l'unione delle scuole popolari per adulti, registra circa 1.580 parole ovvero 1.592 lemmi, precisa Sciarone (1977: 45), tra cui i termini grammaticali (pp. 37-170). Accanto ai lemmi, scelti in base al criterio dell'esperienza, troviamo una ricca esemplificazione e indicazioni grammaticali.[10]

[10] Opere simili sono quelle di Merz (*Langenscheidts Grundwortschatz Italienisch*, 1987) e Deho (*Italienischer Mindestwortschatz,* 1988), destinate entrambe ad apprendenti tedeschi di italiano L2.

2.2.12 Baldelli I., Mazzetti A. (1974) *Vocabolario minimo della lingua italiana per stranieri*, Firenze, Le Monnier.

Nato dall'esperienza dell'Università Italiana per Stranieri di Perugia, gli autori forniscono un repertorio di 1.741 parole e termini grammaticali, "con frasi ed esempi di uso frequente e con traduzione in inglese, francese, tedesco e spagnolo", come precisa il sottotitolo. Pur trattandosi di un'opera compilata sulla base dell'esperienza degli autori, si dichiara che nel vocabolario "si trovano le parole di uso più frequente della lingua italiana" (p. 1). Nella seconda edizione (1978) in appendice si allega un "Dizionarietto italo-somalo" a cura di Bruno Panza, la terza edizione (1982) invece è arricchita da un "Dizionarietto arabo".
La scelta delle parole è basata sull'esperienza degli autori, pertanto in molti casi si trova soltanto la parola più "frequente" di una famiglia lessicale, si trova ad. es. *regalo*, ma non *regalare*, e delle parole polisemiche si riporta solo i significati più "frequenti". L'opera è concepita per studenti stranieri e per i loro bisogni specifici, quindi contiene molte parole inerenti alla grammatica dell'italiano L2. Ad esempio, la presenza di termini come *ausiliare, avverbio, participio* ecc. sulla cui appartenenza del lessico fondamentale c'è almeno da dubitare, è indubbiamente influenzata dall'obiettivo del libro, destinato all'insegnamento, anche grammaticale, dell'italiano, agli stranieri.
I lemmi, disposti in ordine alfabetico, sono seguiti dalla categoria grammaticale, dal loro corrispondente nelle quattro lingue suddette, dalle eventuali forme irregolari e da esempi (frasi, idiomi, proverbi), la cui conoscenza, secondo gli autori, è sufficiente per imparare la lingua d'uso. Il vocabolario è corredato di un inventario grammaticale contenente tutte le forme degli articoli e pronomi, insieme alle principali coniugazioni verbali.

2.2.13 Sciarone A. G. (1977) *Vocabolario fondamentale della lingua italiana*, Bergamo, Minerva Italica.

È una lista di 2.726 termini e parole grammaticali tratti da un corpus di circa 1.500.000 occorrenze. Il corpus nasce dalla combinazione di tre campioni, ciascuno di 500.000 parole. Il corpus del LIF e del FDI è stato integrato con un nuovo campione di 500.000 occorrenze, diviso in 5 sottoinsiemi, non ripartiti in parti uguali come nel LIF e FDI, ma secondo il più plausibile processo di contatto con la lingua da parte di uno straniero. Lo straniero che deve apprendere l'italiano, cercherà nell'ambito della lingua scritta, prima i giornali (da cui sono tratte 126.089 occorrenze del campione) e le riviste (105.120 occorrenze), poi i racconti (105.739 occorrenze) per la loro brevità e quindi i romanzi (74.312 occorrenze) e i lavori teatrali (78.750) (Sciarone 1977: 53). I testi di questo nuovo campione sono stati pubblicati tra il 1952 e il 1974.[11]
La lista è stata composta in base alla sola frequenza, poiché "la frequenza risulta essere il solo criterio valido per fare una selezione di vocabolario a scopo didattico" e tutte le critiche rivolte alla frequenza sono da spiegare con le imperfezioni delle liste elaborate (Sciarone 1977: 51). L'ampiezza del corpus garantisce in sé la stabilità delle frequenze, per questo, utilizzando un corpus di 1.500.000 occorrenze, la frequenza dei primi 2.500

[11] Sciarone (1977: 52-53) fornisce solo le date di pubblicazione dei testi utilizzati per il nuovo campione, senza precisare la data di prima pubblicazione, né quella vera o presunta di composizione. Sgroi (1994: 216) osserva, che Sciarone non ha preso in considerazione, come invece sarebbe stato utile, i testi di Moravia (1957), Calvino (1947) e Cassola (1968), computerizzati da M. Alinei (1973) (cfr. sopra).

lemmi diventa stabile.[12] Il fatto che l'ampiezza del corpus è direttamente proporzionale alla stabilità delle frequenze, è dimostrato dalla legge dei grandi numeri, secondo cui, solo al crescere del numero delle prove, la frequenza di un evento tende a diventare uguale alla sua probabilità di occorrenza. Questa legge "consente di calcolare in misura attendibile entro quali limiti può variare una determinata frequenza" (Sciarone 1977: 75). In quest'ottica si distingue la frequenza reale (effettivamente trovata), da quella "ricalcolata" secondo una formula statistica che permette di eliminare le variazioni di frequenza dovute al caso.
Si giunge alla conclusione che più una parola è frequente, meglio è distribuita, mentre le frequenze basse presentano variazioni sensibili nei singoli contesti, pertanto esiste una forte correlazione tra distribuzione e frequenza. Un lessico di base contenente solo parole molto frequenti sarà utile per la comprensione di qualsiasi testo (Sciarone 1977: 66). In conformità a verifiche empiriche, l'autore costata che per capire un qualsiasi testo occorre conoscere almeno il 90% delle parole che lo costituiscono. Il valore di copertura delle 2.726 parole contenute nella lista giunge fino a 88,73%, pertanto mancherebbero altre 500 parole per coprire quasi tutti i testi possibili. Per recuperare le parole mancanti, Sciarone (1977: 93) suggerisce, stranamente, di affidarsi all'esperienza.
I lemmi sono compresi fra la frequenza massima "ricalcolata" 165.240 (resa attendibile tramite la normalizzazione dello scarto tra le frequenze trovate nei tre corpora) e frequenza massima "reale" 175.895 dell'articolo *il*, e frequenza minima "ricalcolata" 10 e frequenza minima "reale" 40, non però attendibile, di *ala, animo*, ecc.
Sono state elaborate quattro liste:

a) Lista generale in ordine decrescente (pp. 109-163): i lemmi sono ordinati in due liste parallele, secondo frequenza ricalcolata e frequenza reale decrescenti.
b) Lista per classi di parole in ordine decrescente (pp. 167-196): i lemmi sono raggruppati secondo la categoria grammaticale, seguiti dal "rango secondo frequenza ricalcolata".
c) Lista per classi di parole in ordine alfabetico (pp. 198-228): i lemmi raggruppati secondo la categoria grammaticale sono ordinati alfabeticamente, con indicazione del rango "secondo frequenza ricalcolata" e del "rango secondo frequenza reale".
d) Lista generale in ordine alfabetico (pp. 231-285): contiene i lemmi con "rango secondo frequenza ricalcolata" e "rango secondo frequenza reale", categoria grammaticale, "frequenza [totale] ricalcolata", "frequenza [totale] reale", "frequenza FDI", "frequenza LIF", "frequenza VIS" [cioè del nuovo campione approntato da Sciarone].

La seconda edizione aggiornata è uscita nel 1995 (*Vocabolario fondamentale della lingua italiana*, Perugia, Guerra).

2.2.14 *Dizionario illustrato per l'uso essenziale della lingua* (1978) Torino, SEI. (D.I.L.)

Questo dizionario, ricco di circa 13.000 lemmi (privo di termini grammaticali) va incluso fra i lessici fondamentali. Destinato ad un pubblico di ragazzi, ha scopi dichiaratamente didattici. I principi di redazione erano l'"estrema semplicità e chiarezza"

[12] "Oltre la soglia delle 2.500 parole, la frequenza comincia a risultare meno attendibile per l'incidenza del fattore 'caso' e del fattore 'contesto'. Un lessico di 2.500 parole è comunque, a nostro avviso, più che soddisfacente per un corso di italiano di livello fondamentale." Si nota inoltre che i dati del LIF, basati su un corpus di 500.000 occorenze, sono attendibili solo fino ai primi 1.000 lemmi (Freddi in Sciarone 1977: 15).

(p. 3), e si precisa che il dizionario è "opera di un'équipe di linguisti, di studiosi di psicolinguistica e di insegnanti" (p. 2). Gli autori tengono a sottolineare che i lemmi sono stati selezionati con criteri scientifici: "L'oculata e attenta scelta dei vocaboli, che ha avuto come base la loro frequenza nell'uso comune (!) è il risultato di uno studio della lingua parlata e scritta a livello medio (vita sociale, familiare, testi scolastici, narrativi, ecc.)" (p. 3). Purtroppo si specifica secondo quali criteri di frequenza i vocaboli sarebbero stati selezionati. Le illustrazioni "sono soprattutto nelle voci tecniche e scientifiche, di ulteriore chiarimento al testo" (p. 4). L'opera non è priva di preoccupazioni puristiche, infatti, si propone di "correggere gli errori purtroppo diffusi" (p. 4).
Trattandosi di un dizionario ad uso didattico, ogni lemma è esemplificato e definito e/o spiegato. Le informazioni grammaticali, separate nettamente dal resto della voce, seguono gli esempi d'uso, in questa sezione sono riportati anche i sinonimi e i contrari.
Il volume desidera tener conto "del raggruppamento di più vocaboli sotto un solo esponente (famiglia di parole)" (p. 4), ma ciò vale solo per i suffissati più immediati e non per i prefissati e per i composti parasintetici. Sgroi (1994: 227) osserva, infatti, che *giovane, giovanile e gioventù* si trovano tutti a p. 328, mentre *ringiovanire* è rimasto a p. 624.

2.2.15 De Mauro T. *et alii* (1980), *Il vocabolario di base della lingua italiana*. In: De Mauro T., *Guida all'uso delle parole*, Roma, Editori Riuniti, pp. 146-172. (VdB)

Questa lista di 6.690 lemmi, diventati 7.074 nella XI edizione del 1991, comprensivi delle parole grammaticali è stata elaborata sulla base delle 5.356 (Sgroi 1994: 220) unità di "maggiore uso" (De Mauro 1980: 148) del LIF.[13] La selezione dei lemmi di partenza è avvenuta combinando la frequenza con la dispersione, utilizzando lo stesso corpus di 500.000 occorrenze, fornito dal Centro Nazionale Universitario di Calcolo Elettronico dell'Università di Pisa, che costituiva la base del LIF.
La reale comprensibilità delle parole è stata verificata da parte di ragazze e ragazzi di terza media e di adulti con non più che la licenza media (1980: 148).[14] Sono state scartate quelle non comprese dalla metà degli interrogati, così le 5.000 parole sono state ridotte dai collaboratori a 4.937 (4.750 nella X edizione o più esattamente 4.741 secondo Lucisano 1992: 134). Successivamente sono state aggiunte 1.753 parole di "alta disponibilità" (diventate circa 2.300 nella X edizione del 1989), assenti nel LIF, il quale, lo ricordiamo, ha fornito la base per l'elaborazione del VdB.
Il lessico di alta disponibilità è costituito di:

> "parole che può accaderci di non dire né tanto meno di scrivere mai o quasi mai, ma legate a oggetti, fatti, esperienze ben noti a tutte le persone adulte nella vita quotidiana. Sono le parole che diciamo o scriviamo raramente, ma che pensiamo con grande frequenza." (De Mauro 1980: 148)

[13] Il *Vocabolario di base* è uscito nella collana dei Libri di base, avviata dagli Editori Riuniti sotto la direzione di Tullio De Mauro con l'intento di creare una specie di enciclopedia scomponibile, scritta in un linguaggio molto semplice e chiaro. Ogni volume era affidato ad uno specialista, e prima della pubblicazione, il testo prodotto dall'esperto è stato sottoposto ad una riscrittura mirata a raggiungere la più alta leggibilità e comprensibilità.
[14] Nel 1963 con l'avvento della c.d. scuola dell'obbligo è esploso il problema dell'italiano lingua straniera in patria, ovvero "quanta lingua" dovessero sapere gli allievi dopo otto anni di frequenza (Marello 1996: 147).

Nel vocabolario d'alta disponibilità rientrano parole come *abbottonare, adolescenza, alfabeto, astuto, cuffia, pedale, traghetto* ecc. Queste parole rischiano di restare fuori delle liste di frequenza e d'uso, perciò il criterio statistico quantitativo va integrato con criteri qualitativi. Il concetto di disponibilità emerge per la prima volta negli studi di lessicologia francese negli anni Cinquanta, in particolare nell'elaborazione del *francese fondamentale*, che mirava a fornire un vocabolario e una grammatica di base per l'insegnamento del francese a stranieri (v. sez. 1.2.2.2). Il VdB è il primo e tutt'ora l'unico vocabolario di base per l'italiano in cui il concetto di "disponibilità" è stato integrato, facendo un richiamo esplicito al vocabolario passivo, più spesso attivato nella comunicazione orale. Partendo dall'esame dei dizionari dell'italiano comune, si sono isolate le parole di maggiore disponibilità, successivamente sono state controllate tramite l'interrogazione di vari gruppi di parlanti.

I tre livelli del *Vocabolario di base* sono i seguenti:

a) Il *Vocabolario fondamentale* (VF) contiene le parole di maggior uso, 2.000 termini (1.991 secondo Lucisano 1992: 134) stampati in neretto, tratte dalle prime 2.001 parole in ordine d'uso del LIF. Questi termini formano il nucleo più importante all'interno del vocabolario di base. Si calcola, infatti, che qualunque discorso in media utilizza per il 90% della sua estensione repliche di queste 2.000 parole, con un'incidenza variabile secondo gli scopi della comunicazione: si va da un minimo di 80% nei testi più specialistici ad un massimo di 94-95% nei testi divulgativi. Del VF fanno parte soprattutto parole grammaticali, verbi dal significato generale e nomi usuali nella vita quotidiana.
b) Il *Vocabolario d'alto uso* (VAU) include 2.937 termini stampati in grassetto (ridotti a 2.750 nella X edizione del 1989). Queste parole compaiono nel LIF con ragno d'uso tra 2.002 e 5.001. La percentuale del VAU in qualsivoglia testo è intorno a 6-8%.
c) Il *Vocabolario d'alta disponibilità* (VAD) è costituito da 1.753 termini (diventati nella X edizione ca. 2.300, o più esattamente 2.337 secondo Lucisano 1992: 134) stampati in corsivo. Queste parole non appaiono nel LIF, oppure hanno un rango d'uso superiore a 5.001.

L'intero VdB fu sottoposto a una verifica generalizzata attraverso le procedure del *Glottokit* (Gensini, Vedovelli 1983) per due motivi: da un lato c'era la volontà degli insegnanti di applicare a fini didattici la lista del VdB, dall'altro lato gli autori volevano controllarne la reale comprensibilità, perché avevano bisogno di conferme riguardo la puntualità di tale lista. Le rilevazioni sono state effettuate su tutte le 6.690 parole del VdB, in varie scuole italiane, sempre nelle terze medie, per un totale di 152 alunni. Le 2.000 parole più frequenti sono conosciute e usate con sicurezza da circa il 100% degli allievi, a queste dobbiamo aggiungere quel migliaio di termini che nascono dall'intersezione del VAU e del VAD. Notiamo che queste parole costituiscono il "cuore" del lessico italiano. Secondo gli autori usando le parole del VdB possiamo avere buone probabilità di essere capiti da chi ha frequentato almeno la terza media. Usando invece le parole del vocabolario fondamentale, possiamo sperare di essere capiti dal 66% della popolazione italiana, in altre parole da quelle persone che hanno almeno la licenza elementare (1980: 149).[15]

La lista del VdB fu compilata originariamente per definire quel nucleo del lessico italiano, che gli allievi dopo gli otto anni di scuola obbligatoria, devono essere in grado di conoscere. La XII edizione aggiornata del VdB (1997), contenente 2.049 lemmi del VF, 2.576 del VAU e 1.897 voci del VAD per un totale di 6.522 lemmi, sta alla base di

[15] Per avere un'idea della rappresentatività del VdB rispetto all'insieme del lessico italiano, si consideri che i primi 5.000 lemmi del LIF, contenuti quasi interamente nel VdB, coprono il 99,8% delle 500.000 occorrenze del corpus da cui il LIF è stato tratto (cfr. i dati in Bortolini *et al.* 1971: L).

tre dizionari pedagogici (v. sez. 2.2.28-30). L'iniziativa ha una portata che supera la fascia d'età cui si rivolge: il VdB costituisce il principale punto di riferimento anche per la definizione dei livelli di conoscenza nella certificazione d'italiano L2 (v. 5.2). Oltre a scopi didattici, il VdB fu un punto di riferimento per la produzione di testi d'alta leggibilità e comprensibilità. Con l'ausilio del VdB è stato avviato nel 1989 il primo periodico di facile lettura in Italia (v. 4.2.2); e la X edizione del VdB (1989) è allegata al *Codice di Stile delle comunicazioni scritte ad uso delle amministrazioni pubbliche*, pubblicato nel 1993 (v. 4.2.3). A partire dal VdB (nella versione in De Mauro 1991[11]) è stata realizzata la base di dati BDVDB (Thornton, Iacobini, Burani 1997) che contiene i lemmi del VdB corredati di informazioni di carattere ortografico, fonologico, morfologico, sintattico ed etimologico.

2.2.16 Gabrielli A. (a cura di) (1981) *Il mio primo dizionario illustrato*, Milano, Mondadori.

Il volume è la traduzione del *Golden Book Illustrated Dictionary*, pubblicato nel 1962 a New York. Il dizionario, "fatto per gli alunni delle scuole elementari e per quelli delle medie" (p. 9), contiene circa 10.000 parole, tra cui quelle grammaticali. Sarebbe interessante esaminare se sono stati omessi o aggiunti dei lemmi, definizioni, poiché dal carattere dell'opera è lecito pensare ad un adattamento del testo originale (più che ad una semplice traduzione), anche perché accanto ai nomi degli otto curatori inglesi delle illustrazioni troviamo anche due autori italiani.
Come nel caso del D.I.L., ogni lemma è definito e seguito da esempi in cui la "brevità" e la "chiarezza" sono le qualità essenziali (p. 10). Le oltre 3.000 illustrazioni a colori sono "pronte a soccorrere con l'occhio là dove la mente potrebbe in qualche modo tardare" (p. 11). L'opera tende ad assumere atteggiamenti normativi: l'appendice "nutrita e agile" (p. 11) fornisce un riassunto delle regole grammaticali e stilistiche più importanti dell'italiano.
Similmente al D. I. L., la scelta dei lemmi si basa sull'esperienza, infatti, i criteri sono definiti "pratici e logici, naturalmente" (p. 9). Sgroi (1994: 224) osserva che a giudicare dalle parole dell'autore questo non sembrerebbe un lessico "fondamentale" dell'italiano, ma un dizionario delle difficoltà prevalentemente lessicali. Conformemente all'impianto fortemente puristico dell'opera, sono registrate tre diverse classi di voci: "parole di difficile significato"; parole che pur avendo un significato ben noto, hanno altri significati poco conosciuti; infine le parole di significato comune, che però possono presentare delle difficoltà di scrittura, pronuncia, morfologia e sintassi (Gabrielli 1981: 9).[16]

2.2.17 Galli de' Paratesi N. (1981) *Livello soglia per l'insegnamento dell'italiano come lingua straniera*. Strasburgo, Consiglio d'Europa.

[16] Da un mini-sondaggio di Sgroi (1994: 224-227) risulta che circa il 45% dei lemmi, cioè circa 4.500 vocaboli del Gabrielli (1981) sono comuni al D.I.L. (1978). Il lessico di De Mauro (1980) concorda con il D.I.L. (1978) per il 46,6%, in pratica per ca. 3.117 lemmi, e con Gabrielli (1981) nella misura del 44,1%, ovvero in 2.950 voci. Le coincidenze lessicali tra la lista di De Mauro (1980) e quella del D.I.L. (1978) e del Gabrielli (1981) sarebbero del 32,5%, vale a dire circa 2.174 parole.

L'opera è stata pubblicata nell'ambito del progetto "Lingue moderne" del Consiglio per la Cooperazione Culturale del Consiglio d'Europa, con l'obiettivo di fornire un livello di competenza minima, per chi voglia affrontare situazioni d'uso comune nelle principali lingue europee. L'orientamento teorico e l'obiettivo generale di formulare un livello minimo di sopravvivenza linguistica costituiscono il quadro generale, in cui si collocano tutti i lavori del progetto.[17] Sono specifici del *Livello soglia* i gruppi di destinatari, l'analisi dei bisogni comunicativi e la concretizzazione del tutto in un repertorio di forme dell'italiano.
La prima parte dell'opera è dedicata alla pianificazione dei contenuti didattici, in base ai quali è stata elaborata la seconda parte con le concrete forme linguistiche. In appendice il repertorio lessicale è ripresentato in ordine alfabetico; segue un inventario grammaticale che dà un'indicazione schematica delle strutture grammaticali e morfologiche scelte.
Il carico degli esponenti è di ca. 1.500 unità semantiche di cui 400 solo per l'apprendimento passivo. Si nota, che sotto la soglia di 1.500 unità lessicali mancano gli strumenti linguistici necessari per garantire ai destinatari scelti una soddisfacente interazione sociale nelle situazioni che dovranno affrontare (p. 15). A differenza di quanto accade per le liste di frequenza basate su campioni scritti, per l'elaborazione del *Livello soglia* è stata presa in considerazione la lingua parlata.
È destinato a "un pubblico di visitatori temporanei adulti che devono trascorrere periodi non molto lunghi in Italia e che intendono […] intrattenere rapporti di tipo non strettamente professionale con parlanti nativi di italiano" (p. 20). Berretta (1983: 245) osserva che l'espressione "visitatori temporanei" taglia fuori a priori gli immigrati, i lavoratori stranieri e loro famigliari, profughi e simili. I destinatari si suddividono in vari sottogruppi: visitatori temporanei in veste non professionale; studenti stranieri che studiano l'italiano in università all'estero; studenti stranieri (non di italiano) di università italiane; futuri tecnici e lavoratori stranieri; commercianti; studiosi stranieri (pp. 23-27).
Nel quadro teorico fortemente influenzato dalla pragmatica, le quattro attività di base possono essere analizzate ulteriormente in termini pragmatici, secondo i bisogni comunicativi dei discenti. Per procedere a tale analisi, occorre partire dalla situazione in cui si usa la lingua. I componenti di un'interazione linguistica (situazione) sono: l'argomento, su cui si comunica; il ruolo, legato al ruolo psicologico e alle caratteristiche sociali (il grado di conoscenza reciproca e i rapporti gerarchici) dei parlanti; infine l'ambiente, in cui la comunicazione si svolge (pp. 28-29). Le situazioni comunicative dello stesso tipo, aventi alcuni tratti principali in comune, formano i c.d. domini d'interazione. Per i gruppi di destinatari previsti (v. sopra) si precisano quattro domini principali: gruppo dei pari; servizi pubblici; trattative commerciali; istruzione. Si escludono alcuni domini come la famiglia, la vita professionale e i mass media che pur essendo importanti, non rientrano nella competenza comunicativa dei discenti prescelti.[18]
Il risultato è una descrizione puntuale, in termini di interazioni linguistiche, della vita quotidiana ideale dello straniero in Italia: si presenta, cerca casa, viaggia, fa acquisti, ecc. L'autrice stessa ammette che le liste delle concrete forme linguistiche non sono esenti da un certo grado di arbitrarietà, in quanto basate sull'intuizione e sull'esperienza (p. 49).[19]

---

[17] Per il quadro teorico comune dei vari livelli soglia si rinvia alla sez. 5.1.1.
[18] All'interno di ciascun dominio (ad es. istruzione) si precisano gli argomenti (scambi tra i discenti e con gli insegnanti), ruoli sociali (insegnanti e discenti) e psicologici (scambio cognitivo e non di tipo personale), ambienti (locali dell'istituzione educativa), argomenti (temi relativi alla vita dell'istituzione), attività linguistiche (uso attivo e passivo della lingua orale, uso passivo della lingua scritta), atti comunicativi tipici (domandare, rispondere, chiedere spiegazioni ecc.) e le possibili nozioni generali (legate ai processi di deduzione, spiegazione ecc.) e specifiche (legate ai particolari temi del corso) (pp. 37-48).

I bisogni dei discenti sono tradotti in termini di atti linguistici, considerati le unità minime di apprendimento.[20] Le concrete forme linguistiche si suddividono in nozioni specifiche che si riferiscono a concetti particolari in situazioni concrete, e nozioni generali che sono rilevanti nella maggioranza delle situazioni. Quest'ultima lista corrisponde, per gran parte, all'inventario contenuto nel *Threshold Level*, le nozioni generali sono suddivise nei seguenti gruppi: entità, esistenza, spazio, tempo, quantità, qualità, nozioni mentali, rapporti (pp. 105-130). Le nozioni concrete sono raggruppate in base ai tipi di interazione che l'apprendente deve affrontare: identificazione personale; casa; vita privata; istruzione e carriera; tempo libero; viaggi; rapporti con gli altri; salute e benessere; compere; mangiare e bere; servizi; luoghi; lingua straniera; opinioni politiche; tempo (pp. 131-178). Nei due inventari è riportato il concetto che il discente deve saper esprimere, p. es. "causa, conseguenza", insieme alle forme linguistiche che si suppone siano le più utili per esprimere tali concetti. Le nozioni generali e gli atti comunicativi tendono ad essere gli stessi per i vari sottogruppi di destinatari, poiché si applicano in una vasta gamma di situazioni, pertanto essi formano il "nucleo comune" dell'insegnamento (il concetto risale originariamente a Van Ek 1975).
Dal materiale Berretta (1983: 247) ricava che "sembrano preferiti i termini più specifici contro i sovraordinati (*quadrato, rotondo* vs. R *forma* p. 121), i termini di registro meno formale (*avere* vs. R *possedere* p. 128, *vero* vs. R *autentico* p. 123), i termini meno marcati (*odore* vs. R *puzzo* p. 124), le espressioni analitiche (*più grande* vs. R *maggiore* p. 124, *piacere di più* vs. R *preferire* p. 94). Nelle scelte di registro si nota un orientamento a espressioni molto colloquiali (*andare matto per, non poter vedere* p. 140, *fregarsene* p. 94). Si osserva che sarebbe stato meglio limitare questo criterio alla competenza passiva, restando per la competenza attiva fedeli al criterio più usuale di fornire per prime forme di registro medio che meno espongono lo straniero a rischi di errori "sociolinguistici."
Nonostante varie critiche pare che l'opera costituisca tutt'oggi un valido punto di riferimento per chiunque voglia cimentarsi nell'impresa di stilare un sillabo di italiano L2.

2.2.18 Batinti A., Trenta W. (1982) *Ricerche sul lessico di base dell'italiano contemporaneo*, Perugia, Guerra.

L'obiettivo del lavoro è di "offrire agli operatori nel campo della ricerca linguistica materiale per la descrizione dell'italiano, e agli operatori nel campo della didattica dell'italiano L1 e L2 strumenti per la valutazione oggettiva dei libri di testo, e materiale utile per l'elaborazione degli eserciziari" (p. 19). Gli autori hanno messo a confronto tre opere dalle caratteristiche diverse: il *Vocabolario di base* di De Mauro *et alii* (1980), il *Vocabolario fondamentale della lingua italiana* di Sciarone (1977) e il *Vocabolario minimo della lingua italiana per stranieri* di Baldelli e Mazzetti (1974). "È sembrato opportuno integrare la componente socio-linguistica (De Mauro, op. cit.), la proposta della linguistica computazionale (Sciarone, op. cit.) e quella basata sull'esperienza (Baldelli-Mazzetti, op. cit.) per trarne indicazioni utili sul piano

[19] Bisogna ripetere le riserve avanzate subito dopo la pubblicazione da Berretta che in una lunga recensione notava come le liste approntate dal LS (liste di bisogni comunicativi, di situazioni comunicative, ecc.) pur essendo "di molto buon senso e nell'insieme ragionevoli ed equilibrate", risultavano estremamente soggettive "da ogni punto di vista" (Berretta 1983: 245).
[20] Gli atti comunicativi principali sono: convenzioni sociali (iniziare e chiudere un contatto, ringraziare, congratularsi, augurare, fare le condoglianze, scusarsi, brindare); tecniche della comunicazione; stati emotivi, sentimenti; progetti di azione; comunicazioni e opinioni; atti di risposta non specifici. (pp. 81-86.)

descrittivo e applicativo" (Batinti-Trenta 1982: 14). L'integrazione delle tre liste ha portato a una lista di 6.933 unità. La somma delle tre liste dà in realtà 11.147 parole (6.690+2.726+1.741) ma 4.224 sono voci comuni ai tre sottoinsiemi. Per ogni parola è indicata la presenza nei tre vocabolari, per il VdB si precisa anche l'appartenenza ad una delle tre fasce del vocabolario. Dalla combinazione dei tre lavori risultano varie liste: voci presenti in tutte e tre le liste, lemmi inclusi in due delle tre liste, parole presenti in una sola lista. Nelle intenzioni degli autori queste liste possono costituire un punto di riferimento per reperire le 500 parole mancanti in Sciarone,[21] necessarie per coprire quasi tutte le situazioni possibili.
Nella seconda parte della ricerca è stata controllata, sul corpus delle 6.933 parole, la presenza e la distribuzione dei fonemi, in particolare di quelli che in base all'esperienza degli autori costituiscono una difficoltà nell'apprendimento (nessi consonantici, quantità consonantica). Ad esempio, 786 sono le parole sulle 6.933, nelle quali compare il fonema /ε/, ossia 11,337% sul totale (p. 18). Le tabelle offrono materiali per un primo orientamento.

2.2.19 *Dizionario illustrato della lingua italiana per la scuola dell'obbligo* (1986) Milano, Editrice Piccoli/Mondadori.

Il dizionario, destinato alla scuola elementare, registra "oltre 10.000 parole" selezionate in base alla frequenza intesa intuitivamente: "sono solo quelle che facilmente ti può capitare di incontrare e di usare nella prima parte della tua vita di studente" (Piccoli 1986: ix). Ogni termine è illustrato con un esempio, seguito a sua volta da una breve definizione, con indicazioni funzionali di sinonimi e contrari, rinvii a derivati e composti. Non mancano illustrazioni essenziali in bianco e nero.

2.2.20 *Vocabolario elettronico della lingua italiana. Il vocabolario del 2000* (1989) Milano, IBM Italia. (VELI)

Il VELI nelle intenzioni degli autori doveva essere il prototipo di una nuova generazione di vocabolari su supporto elettronico. Piuttosto che un vocabolario, il VELI si può definire un lemmario, cui sono associate informazioni di tipo grammaticale. Il VELI offre varie funzioni: è possibile indicare tutte le possibili forme flesse per ciascuno dei 10.000 lemmi e, inversamente, ricondurre ciascuna forma al lemma o ai lemmi cui fa capo; dividere le parole in sillabe (secondo la scansione ortografica); indicare sinonimi e contrari, ricercare ed elencare parole che cominciano con una data stringa. Il progetto è stato realizzato dalla Direzione Ricerca Scientifica e Tecnologica dell'IBM Italia con la consulenza scientifica di De Mauro, si tratta di un'opera fuori commercio, distribuito esclusivamente fra i linguisti.
I lemmi sono stati ricavati dall'analisi di circa 26.100.000 occorrenze tratte da fonti giornalistiche (*Il Mondo, Europeo, Domenica del Corriere* e materiali dell'agenzia giornalistica ANSA), apparse nel biennio tra il settembre 1985 e il giugno 1987. È stato additato più volte (Sgroi 1994: 229, Vedovelli 1995: 61) che a causa della composizione del corpus risulta il VELI presenta le parole di una lingua più formale ed informativa,

[21] Il valore di copertura delle 2.726 parole contenute nella lista giunge fino all'88,4%, pertanto mancherebbero altre 500 parole per coprire quasi tutti i testi possibili (Sciarone 1977: 94).

ricca di tecnicismi, pertanto non sorprende che certe parole legate a esperienze quotidiane restano fuori del dizionario.
Le parole sono state sottoposte ad una serie di "filtraggi" successivi (p. 62 e seg.) prima di giungere alla lista finale di 10.000 lemmi. Dalle occorrenze iniziali si sono eliminate le parole con frequenza 1 in ciascuno dei quattro sottoinsiemi, ottenendo così un campione di 24.762.136 parole (p. 68). Si sono poi estratte le 80.000 forme più frequenti, ovvero le 20.000 forme più frequenti di ciascuno dei quattro sottoinsiemi. Dopo aver eliminato le intersezioni dei quattro sottoinsiemi, le voci si sono ristrette a 32.876 di cui si sono estratte le 20.000 forme più frequenti che hanno prodotto 18.793 lemmi. Da quest'elenco sono stati eliminati 4.558 nomi propri, sigle, alterazioni, elisioni e troncamenti, parole straniere ed altri 5.235 lemmi (la cui rimozione non è chiaramente giustificata) per arrivare ai 10.000 prefissati dagli autori.
Nella scelta dei lemmi il criterio della frequenza assoluta è stato integrato con quello dell'uso, introdotto in base al LIF (v. 2.2.8).[22] Le parole sono comprese tra "uso" 1.326.167 (con dispersione 0,97) della preposizione *di* e uso 65 (con dispersione 0,32) dell'aggettivo *statuario*. Accanto alle 10.000 parole così selezionate, sono state inserite una trentina di parole che pur presentando un valore d'uso inferiore a quello previsto, sono usate da VELI per i messaggi e le informazioni di Aiuto fornite all'utente. Questa scelta "corrisponde a una istanza di *chiusura in sé* del vocabolario" (VELI 1989: 69), infatti, l'utente deve poter trovare all'interno del sistema le informazioni relative ai vocaboli che il sistema utilizza, anche se sulla base della metodologia statistica essi non sarebbero entrati a far parte del vocabolario stesso.[23] La versione cartacea contiene varie liste.

a) In appendice A i lemmi sono disposti in l'ordine decrescente d'uso (pp. 85-251), con indicazione della categoria grammaticale, della frequenza in ogni sottoinsieme, di dispersione e di uso.
b) In appendice B è riportato lo stesso elenco in ordine alfabetico (pp. 255-331), i lemmi sono seguiti dalla categoria grammaticale e dall'indicazione d'uso.
c) L'appendice C contiene la lista delle 564 differenze che si sarebbero presentate nella lista dei 10.000 lemmi se si fosse utilizzato, come criterio ordinatore, la frequenza complessiva, invece del valore d'uso (pp. 334-339).
d) L'appendice D comprende la lista dei lemmi che non sarebbero entrati sulla base della sola frequenza, e che invece vi entrano avendo utilizzato il valore d'uso (pp. 341-347).
e) In appendice E figurano i 58 lemmi stranieri, di cui 24 grazie all'utilizzo del fattore d'uso rientrano fra i primi 10.000 lemmi (un sottoinsieme dell'appendice D), mentre i restanti 34 sarebbero entrati usando la frequenza complessiva (un sottoinsieme dell'appendice C) (pp. 349-351).
f) L'appendice F è un confronto tra i primi 2.000 lemmi del LIF e del VELI (pp. 353-397).

[22] "Classificare i lemmi in base al numero delle loro occorrenze sommate … è ingannevole …, perché una parola, le cui occorrenze sono tutte o quasi tutte nella stessa fonte, si può presumere sia di uso troppo specialistico. Vorremmo una misura combinata, di quanto bene la parola sia dispersa fra le varie fonti … e di quanto sia frequente in assoluto" (VELI 1989: 66-67).
[23] Marello (1996: 122) avverte che uno dei difetti definitori da cui il lessicografo deve guardarsi è, accanto alla circolarità diretta, l'uso di termini non definiti a loro volta nel dizionario.

2.2.21 Cannella M. (a cura di) (1991) *Il primo Zanichelli. Vocabolario elementare di italiano*, Bologna, Zanichelli.

Il vocabolario, destinato agli alunni delle elementari, registra "oltre 15.000 parole" e secondo l'annuncio pubblicitario della casa editrice "più di 21.000 accezioni" selezionate secondo una frequenza "intuitiva", vale a dire le parole "più frequenti nei libri di studio o di lettura per ragazzi" (Cannella 1991: 3). La quarta edizione del 2004 comprende, secondo le affermazioni dell'editore "oltre 21.000 voci e 27.000 accezioni", pertanto difficilmente potrebbe essere chiamato dizionario di base, tuttavia gli autori non mancano a segnalare le 4.300 parole appartenenti all'italiano fondamentale. In base a quali criteri le parole fondamentali siano stati scelti, non è specificato.
Ogni parola è illustrata come nel Piccoli (1986) con un esempio d'uso, seguito da una definizione. Il volume include anche "5.000 illustrazioni a colori o in bianco e nero" che "insegnano nuove parole, fra le quali anche molte non comprese fra quelle definite nel vocabolario" (ivi).

2.2.22 Katerinov K., Boriosi Katerinov M. C., Sciarone A. G. (1991) *Calcolo di rendimento di liste base: italiano parlato, italiano scritto, livello soglia*, Perugia, Guerra. (KBS)

I 691 lemmi sono stati tratti da corpus parlato di 150.000 occorrenze, raccolto tra il 1987 e il 1990 (p. 8). I criteri adottati nella formazione del corpus sono orientati alla didattica dell'italiano come lingua straniera (domini/argomenti, ruoli socio-psicologici, età, sesso, grado di istruzione, classe sociale, professione, aree dialettali di appartenenza), senza tener conto di una tipologia del parlato come nel LIP. I domini (temi) delle conversazioni registrate e le più frequenti funzioni comunicative (pp. 9-11) coincidono per larga misura con quelli previsti da Galli de' Paratesi (1981). I testi sono stati prodotti da parlanti di ambedue i sessi, dai 3 ai 70 anni di ogni grado di istruzione e di tutte le classi sociali, di tutta l'Italia (p. 12).
La selezione dei lemmi è basata sulla sola frequenza: sono state selezionate e lemmatizzate automaticamente (pp. 14, 52-53) le parole aventi una frequenza da 20 in su (p. 14). Secondo gli stessi autori, il corpus è troppo poco rappresentativo per uno studio attendibile sulla frequenza, per questo la raccolta dei testi dovrebbe proseguire fino al raggiungimento di almeno 500.000 occorrenze, tuttavia "non c'è motivo di ritenere che in seguito si potranno osservare differenze significative per quanto riguarda le prime 700 parole più usate dell'italiano parlato", che rientrano peraltro in un minimo lessicale del *Waystage*[24] (p. 15).
Le 691 parole sono ordinate in due colonne parallele, per frequenza decrescente e alfabeticamente (pp. 23-36). Per ogni lemma è indicata la frequenza (compresa fra 6.969 e 20), ma non la categoria grammaticale di appartenenza.[25]
L'opera si propone di individuare le strutture grammaticali e lessicali tipiche del parlato, e di verificare l'ipotesi di partenza, ovvero che la grammatica del parlato sia diversa da quella della lingua scritta (pp. 6-7). Per questo, la seconda parte contiene un confronto tra i 691 lemmi del parlato e con le 2.726 parole dell'italiano scritto di Sciarone (1977) (pp.

[24] Van Ek J. A., Alexander L. G., Fitzpatrick M. A. (1977) *Waystage, an Intermediary Objective Below Threshold Level,* Strasbourg, Council of Europe; Van Ek J.A., Trim J.L.M. (1980) *Waystage English*, Pergamon Press.
[25] Stranamente si trovano due lemmi *dare* con frequenza 374 e 57 (p. 24 e 28), si trovano inoltre come lemmi, anziché forme: *mia* (fr. 200) e *mio* (fr. 213), *la* (fr. 3.468) e *lo* (fr. 648), *lire* (fe. 32) e non *lira*.

37-51). Nella prima colonna i lemmi sono ordinati per frequenza decrescente (secondo il KBS) con i rispettivi ranghi nel KBS e nella lista di Sciarone, mentre nella seconda colonna i lemmi figurano in ordine alfabetico. Dal confronto emerge che 651 termini sono comuni nelle due liste. In Sciarone mancano 40 termini di KBS: 7 nomi geografici (*America, Capri, Italia, Milano, Napoli, Roma, Sanremo*) e 33 termini comuni (*aids, antipatia, appello, beh, buonasera, buongiorno, cancro, codice, cognata, collegare, cortesia, diario, dibattito, espresso* agg. e s.m., *festival, giornalismo, glielo, imputato, innanzitutto, mago, mah, morto* agg. e s.m., *nucleare* agg. e s.m., *ovviamente, pistola, prevenzione, reato, sanità, soccorso, sentenza, sicuramente, sieropositivo* agg. e s.m., *televisivo*).[26]
Il confronto dei 691 lemmi del parlato con i 691 più frequenti di Sciarone (1977) dà una percentuale di 64,83% di voci in comune (pp. 15-16), in assoluto 454 parole. Questo numero relativamente basso si spiega con il fatto che la variazione della frequenza (e dei ranghi occupati dalle singole parole) è dovuta in parte al caso, pertanto esaminando le parole con un rango più basso nella lista di Sciarone, il numero di parole comuni aumenta in modo considerevole. Quindi, in realtà, la somiglianza fra lessico del parlato e lessico dello scritto "è maggiore di quanto appaia a prima vista" (p. 16).

2.2.23 De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993) *Lessico di frequenza dell'italiano parlato,* Milano, Etaslibri. (LIP)

Il lessico, preparato dalla stessa équipe del VELI, presenta un elenco delle 15.641 parole più frequenti della lingua italiana *parlata*. I lemmi sono stati ricavati da un corpus di riferimento di 496.335 occorrenze, ossia 475.883 forme grafiche (pp. 37, 112). Trattandosi del primo corpus di parlato raccolto secondo criteri espliciti, esso intende essere rappresentativo sia dal punto di vista dei generi di parlato, sia dal punto di vista geolinguistico. Il LIP è un tentativo di rappresentazione statistica dell'italiano comune parlato in tutto il territorio nazionale. Le registrazioni, dalla durata complessiva di ca. 57 ore, sono state reperite tra il novembre 1990 e il luglio 1992 (p. 45) in 4 città italiane: Milano, Roma, Firenze, Napoli (p. 31); 125.000 parole per città, suddivise ulteriormente in pacchetti di 25.000 occorrenze per ognuno dei cinque tipi di discorso (v. più avanti).[27]
I testi sono stati prodotti da 1.653 parlanti di ambedue i sessi (p. 51), dai 6 anni in su, ripartiti per sei fasce d'età (p. 51). È stata presa in considerazione la ripartizione per titolo di studio e il livello di professione (p. 52).
Si è considerato parlato ogni produzione fonica e spontanea. Sono stati identificati cinque tipi di situazioni comunicative su una scala che va da una maggiore ad una minore naturalezza (p. 35): a) scambio bidirezionale faccia a faccia con presa di parola libera (conversazioni); b) scambio bidirezionale non faccia a faccia con presa di parola libera (telefonate); c) scambio bidirezionale faccia a faccia con presa di parola non libera (dibattiti, interviste, interrogazioni ecc.); d) scambio unidirezionale in presenza di destinatari (lezioni, conferenze, relazioni ecc.); e) scambio bi- o unidirezionale a distanza

[26] Sgroi (1994: 231) giustamente osserva che gli autori indicano come assenti in Sciarone (1977) alcuni lemmi che in realtà vi sono presenti: *perché, né*; altri mancano solo in apparenza, perché lemmatizzati sotto altra forma: *oramai* (anziché *ormai*), *mio* (anziché *mia*), *lira* (anziché *lire*), *lo* (anziché *la*).
[27] La suddivisione omogenea dei testi in tutta l'Italia acquista particolare importanza alla luce del fatto che l'italiano è diventato lingua nazionale soltanto negli ultimi decenni, e che i suoi dialetti ancor oggi godono di una grande vitalità. Le varietà regionali sono quelle che effettivamente sono parlate, e dato che si tratta di varietà locali dell'italiano, esse differiscono notevolmente una dall'altra, anche se l'effetto unificatore dei mass media tende a ridurre queste differenze.

(trasmissioni radiofoniche o televisive). Per ciascun tipo di parlato sono stati raccolti testi per un totale di 100.000 occorrenze (25.000 in ciascuna città), e ogni genere di parlato è suddiviso ulteriormente in sottoclassi orientative (pp. 40-41). A titolo d'esempio per il gruppo a) si definiscono le seguenti sottoclassi: conversazioni in casa; conversazioni sul luogo di lavoro; conversazioni nell'ambito scolastico e universitario; conversazioni in luoghi ricreativi e sui mezzi di trasporto.
Dal corpus sono stati estratti i 15.649 lemmi più frequenti. Tra le voci sono incluse oltre le parole grammaticali e anche nomi propri (p. 93), cognomi (p. 90), nomi di ditte (ibid.) e nomi geografici (p. 91). A questi lemmi vanno aggiunti 35 fonosimboli (pp. 531; 92-93).
I lemmi statisticamente più attendibili sono quelli con frequenza superiore a 2 (p. 121), vale a dire i primi 7.213. Tali unità coprono il 97,69% dei testi registrati (p. 116), ciò significa che la copertura del restante 2,31% dei testi richiede la conoscenza degli altri 8.641 lemmi con frequenza inferiore a 3. I 15.649 lemmi sono stati raggruppati in cinque liste:

a) Sono stati ordinati alfabeticamente i 7.213 lemmi con almeno 3 occorrenze e le 20.388 forme relative (pp. 171-400). Per ogni lemma e forma si indicano la categoria grammaticale, la frequenza di ognuno dei cinque generi di discorso, la frequenza totale e l'uso (prodotto della frequenza e distribuzione). Per i soli lemmi, non per le forme, si indica anche il rango d'uso.
b) Segue poi una lista in ordine alfabetico di 8.428 lemmi con meno di 3 occorrenze (pp. 401-435). I lemmi sono stati tratti da 9.044 forme (p. 114). Per ogni lemma si indica la frequenza in ognuno dei cinque generi di discorso, la frequenza totale e il rango d'uso.
c) La terza lista è quella complessiva di tutti i 15.649 lemmi in ordine d'uso e di rango da 1 a 6.505 (pp. 436-530). Per ogni lemma è riportata la categoria grammaticale, la frequenza in ognuna delle quattro città d'inchiesta, la frequenza totale, il valore d'uso nonché il rango d'uso.
d) La quarta è una lista di frequenza (p. 531) dei 35 fonosimboli, ciascuno accompagnato dal numero delle occorrenze nell'intero corpus.
e) L'ultima lista contiene in ordine di frequenza decrescente 1.933 espressioni polirematiche (pp. 532-540), per un totale di 6.874 occorrenze (p. 64, 112). I componenti delle polirematiche sono stati conteggiati sotto i lemmi delle prime due liste (p. 96).[28]

Al volume sono stati allegati due dischetti che permettono la consultazione di tutti i testi delle registrazioni su cui si basa il LIP, consentendo di verificare la provenienza della rilevazione, il contesto d'uso e l'esatta valenza semantica di ogni parola-occorrenza, ovvero la polisemia dei lemmi.

2.2.24 Marconi L., Ott M., Pesenti E., Ratti D., Tavella M. (1994) *Lessico elementare. Dati statistici sull'italiano letto e scritto dai bambini delle elementari*, Bologna, Zanichelli. (LE)

Il *Lessico elementare* è il primo dizionario di frequenza italiano, relativo al lessico dei bambini che frequentano la scuola elementare.[29] Si basa su un corpus di 1.088.725 occorrenze, ricavate da pubblicazioni per allievi della scuola elementare e da

[28] Notiamo che quest'ultima lista, contenente espressioni come *va bene, d'accordo, che cosa* è di particolare utilità per l'insegnamento dell'italiano L2 (cfr. 6.5).
[29] I primi studi svolti nel campo del lessico dell'infanzia si devono a Thorndike (1921, 1944).

composizioni scritte da scolari della stessa fascia. Il periodo di riferimento è tra il 1987 e il 1989, i testi sono stati raccolti su tutto il territorio nazionale. È un lessico di frequenza della lingua scritta che comprende varie liste:

a) *Lessico Elementare in ordine di frequenza* (pp. 41-89): la lista delle 6.095 parole più frequenti nei testi scritti da e per bambini, basato su un corpus di 501.300 occorrenze. Lo spoglio dei testi globalmente ha dato origine a 19.624 lemmi diversi, di cui sono stati eliminati 13.529 con una frequenza inferiore a 3,98. Ogni lemma è corredato dalla categoria grammaticale, indice d'uso e rango.[30]
b) *Lessico di Lettura in ordine di frequenza* (pp. 91-135): la lista delle 5.511 parole che i bambini leggono di più. L'intero corpus di questo settore è stato suddiviso in tre sottosettori: libri di lettura e fiabe, giornalini e fumetti, libri di testo. Per ogni sottosettore sono stati scelti 100 testi, da ciascuno dei quali sono state estratte ca. 1.700 parole, 170.000 parole per sottosettore, per un totale di 587.425 occorrenze. Nella lista in ordine di frequenza decrescente ogni lemma è corredato dalla categoria grammaticale, frequenza d'uso e rango, inoltre s'indica l'eventuale assenza del lemma nel *Lessico Elementare*, nel *Lessico di Scrittura* o nell'intero corpus di scrittura.
c) *Lessico di Scrittura in ordine di frequenza* (pp. 137-164): la lista contiene le 3.543 parole che i bambini scrivono di più, ricavate da 5.000 temi diversi (1.000 per ciascuna delle 5 classi elementari), analizzati in base all'età scolare, sesso, località di provenienza degli alunni. L'argomento dei testi scritti dai bambini è stato lasciato volutamente libero, al fine di rendere il corpus di riferimento il più vario possibile. Nella lista in ordine di frequenza decrescente ogni lemma è corredato dalla categoria grammaticale, frequenza d'uso e rango, inoltre s'indica l'eventuale assenza del lemma nel *Lessico Elementare*, nel *Lessico di Lettura* o nell'intero corpus di lettura.
d) *Lessico Elementare in ordine alfabetico* (pp. 167-393): in questa sezione ogni lemma occupa un'intera riga di due pagine affiancate. La pagina di sinistra contiene le frequenze assolute dell'intero corpus, dei due settori principali (lettura e scrittura) e degli otto sottosettori (libri di testo, libri di lettura, giornalini e fumetti per il lessico di lettura; temi delle cinque classi elementari per il lessico di scrittura). La pagina di destra contiene gli indici statistici utilizzati: rango, uso, dispersione e frequenza normalizzata (modificata dal valore d'uso) per l'intero corpus e per i due sottosettori.
e) *Nomi propri in ordine alfabetico* (pp. 395-416): si elencano i 495 nomi propri più frequenti del *Lessico Elementare,* la presentazione ricalca quella della lista d).
f) *Alcune liste in breve* (pp. 417-437): si riportano separatamente, in ordine di frequenza decrescente, le liste di sostantivi, verbi e aggettivi appartenenti rispettivamente al *Lessico di Scrittura* e al *Lessico di Lettura*, suddivisi in classi di frequenza di 500 lemmi.

Il fatto, che i primi 500 lemmi del *Lessico di Scrittura* coprono quasi l'84,42% di tutte le parole usate nei temi, dà un'idea abbastanza chiara della povertà del lessico attivo degli alunni delle elementari. Il valore di copertura dei primi 500 nel *Lessico di Lettura*, invece, è pari a 74,39% del totale delle occorrenze, il che dimostra una capacità ricettiva più ricca (p. 28). Anche il rapporto fra numero assoluto di lemmi e la percentuale di occorrenze prodotte ci fa capire che il *Lessico di Scrittura* risulta molto più povero rispetto al *Lessico di Lettura*, infatti, il 90% delle occorrenze è coperto per il settore *Scrittura* da 752 lemmi, mentre per il settore *Lettura* da 1.829, cioè ben più del doppio (p. 32). Le prime 1.405 parole del *Lessico Elementare* coprono il 90% del corpus, quindi

[30] Un confronto non sistematico fra i 6.095 lemmi del LE e i 7.050 del VdB mostra che i 2.000 lemmi fondamentali si trovano in entrambi, così pure quelli di alto uso, mentre gran parte dei lemmi di alta disponibilità non compare nel LE (Marello 1996: 149).

sembra che meno di 1.500 lemmi rappresentino la quasi totalità degli strumenti usati per la comunicazione scritta (p. 31).[31]
Grazie alla presentazione separata del lessico di lettura e di scrittura, il LE è uno degli strumenti che meglio permettono all'insegnante di fare previsioni sulla difficoltà di comprensione di una determinata parola.

2.2.25 Bertinetto P. M., Burani C., Laudanna A., Marconi L., Ratti D., Rolando C., Thornton A. (1995) *Corpus e Lessico di Frequenza dell'Italiano Scritto* (CoLFIS)

Quest'opera è stata realizzata da un gruppo di linguisti, col finanziamento del Consiglio Nazionale delle Ricerche. Il corpus di riferimento è composto di 3.150.075 occorrenze. Nel campionare i testi, gli autori hanno ponderato i dati dell'ISTAT, relativi alle letture preferite dagli italiani, per ottenere un corpus che fosse il più possibile rappresentativo di ciò che gli italiani effettivamente leggono. La suddivisione dei 3 milioni di occorrenze ha dato luogo alla seguente ripartizione: 1.500.000 occorrenze tratte da quotidiani del periodo 1992-1994 (La Repubblica, La Stampa, Il Corriere della Sera), 1.000.000 occorrenze da periodici e 500.000 occorrenze da libri, considerando anche i libri letti per motivi scolastici o professionali (per una descrizione più dettagliata del corpus cfr. Laudanna *et al.* 1995).
Il lessico di frequenza include due sottoparti: un formario e un lemmario.

a) Il formario contiene la frequenza di ciascuna delle 140.000 forme presenti nel corpus, senza distinguere fra i vari lemmi cui la forma si può ricondurre. I file del formario contengono le forme estratte dal corpus di italiano scritto, ciascuna corredata dalla frequenza assoluta con cui compare nel corpus e dal numero di caratteri di cui è costituita. I file differiscono fra di loro per il carattere (minuscolo, maiuscolo, numerico) e l'ordine (alfabetico diretto, alfabetico inverso, numerico) in cui le forme sono riportate.
b) Il lemmario contiene invece le frequenze delle forme lemmatizzate. Il lemmario indica inoltre la frequenza delle parole polirematiche. I file del lemmario contengono le forme (ricondotte al lemma originario) e i relativi lemmi (ca. 90.000). Le forme e i lemmi sono corredati da: categoria grammaticale; frequenza assoluta e relativa, dispersione totale e parziale (calcolata sui sottosettori che costituiscono il corpus); rango d'uso; numero di caratteri di cui la stringa è costituita. I diversi file differiscono fra di loro per le informazioni contenute (solo forme, solo lemmi, sia forme che lemmi) e l'ordine (alfabetico diretto, alfabetico inverso) in cui le forme e i lemmi vengono riportati.

Secondo gli autori, i punti di forza di CoLFIS, rispetto ai lessici di frequenza precedentemente disponibili per l'italiano sono due: il bilanciamento delle fonti che conferisce un carattere di non casualità alle rilevazioni numeriche estraibili dall'archivio lessicale; e la sua ampiezza. Non esistono, infatti, molti altri esempi di corpora italiani di queste dimensioni.
I file del lessico di frequenza sono disponibili per il download, l'accesso è gratuito (http://www.istc.cnr.it/material/database/colfis/Pres_gen.shtml). Il corpus (limitatamente

[31] La situazione non è specifica del lessico dei bambini: ciò emerge da un confronto con il LIF (1971) e con il FDI (1973). Mentre i primi 1.000 lemmi del LE coprono l'86,87% delle occorrenze, nel LIF coprono 87,33% e nel FDI 85,59%, ma già le prime 10 parole sono sufficienti a coprire ca. il 30% in tutti e tre i corpora.

alla parte per cui sono state rilasciate le autorizzazioni) e il dizionario sono consultabili presso: www.ge.ilc.cnr.it/strumenti.php.

2.2.26 Savino E. (a cura di) (1995) *Il mio primo vocabolario*, Milano, Mursia.

Questo vocabolario registra ca. 5.000 lemmi, tra cui anche parole grammaticali. Nel vocabolario sono raccolte due categorie essenziali di lemmi: le parole di base (ca. 2.000) consentono di "emettere e recepire messaggi completi (e abbastanza complessi) nella normalità delle varie situazioni quotidiane" (p. III); le parole a più alta frequenza "pur non giungendo ad esprimere contenuti settoriali o specialistici, posseggono però un'area di significati meno generali, meno complessivi delle parole di base" (p. IV). Del gruppo delle parole di base fanno parte i termini grammaticali (*articolo, aggettivo, morfologia* ecc.) che non appartengono al lessico fondamentale, ma devono essere inclusi in un lessico che ha la funzione di completare un testo di grammatica per la scuola.
L'opera, infatti, prende spunto dalla tradizione didattica-linguistica anglosassone, in quanto si basa sul concetto di *elementary dictionary*, ovvero dizionario di base, i cui requisiti sono due: registrare un numero ben definito di lemmi fondamentali; corredarli di spiegazioni del significato e dell'uso semplici e trasparenti (p. III). La lettura del dizionario, inteso come autentico "libro di testo" della lingua, assicura agli studenti mezzi linguistici sufficienti per la corretta comunicazione quotidiana.
Quanto ai singoli lemmi, purtroppo manca un'indicazione sull'appartenenza ad una delle due categorie (parole fondamentali e d'alta frequenza), così come non sono specificati i criteri adottati per la selezione dei lemmi. Ogni lemma è accompagnato dalla categoria grammaticale, eventuali informazioni morfologiche, definizione e fraseologia esplicativa. Contrariamente a quanto promesso sul fronte del libro, non sono riportati i sinonimi e i contrari.

2.2.27 De Mauro T., Moroni G. (1996) *Dizionario di base della lingua italiana,* Torino, Paravia. (DIB)

Cronologicamente questo è il primo nella serie dei dizionari pedagogici redatti con l'ausilio del *Vocabolario di base* (De Mauro *et al.* 1980) (v. anche 2.2.28 e 2.2.29). Questo volume, dedicato a ragazzi da 8 a 11 anni, è per ora il prodotto che più si avvicina a un dizionario didattico per apprendenti stranieri. Si basa sulle stesse categorie del *Vocabolario di base*, ma alle 7.000 parole del VdB sono state aggiunte altre 8.000 che in base all'esperienza didattica degli autori risultano indispensabili in una società moderna e complessa. Esso comprende: 2.000 parole fondamentali (comprese le parole grammaticali) contrassegnate da un cerchietto; 3.000 d'alto uso, indicate con un semicerchio; 2.000 di alta disponibilità segnate con una lunetta e altre 8.000 aggiunte a posteriori.
Per ogni lemma (15.000 in tutto) è indicata l'appartenenza ad una delle categorie suddette. Seguono le informazioni sulla pronuncia, sillabazione, categoria grammaticale ed etimologia. Sono definiti i sensi fondamentali di ciascun vocabolo e le frasi ed espressioni in cui la parola vive. Sono poi indicate le parole da cui essa deriva e quelle che ne derivano, infine segue un elenco dei sinonimi e dei contrari. Il DIB contiene più di 2.000 rinvii al c.d. *Dizionario visuale* che include ca. 2.000 immagini a colori, strutturate intorno a 8 grandi nuclei (lo spazio e la terra, l'uomo e il suo corpo, le costruzioni, il

lavoro, i trasporti, le comunicazioni, l'arte e le attività ricreative). Le illustrazioni sono arricchite da 55 riquadri con 700 verbi di riferimento che suggeriscono le azioni pertinenti a ogni contesto. Segue infine un indice analitico di tutti gli oggetti raffigurati. Il DIB è inoltre accompagnato da una guida didattica (Cattaneo 1996).

2.2.28 De Mauro T., Moroni G. (1997b) *Prime parole. Dizionario illustrato di base della lingua italiana*, Torino, Paravia.

L'opera si inserisce in una serie di dizionari pedagogici, destinati ai ragazzi delle scuole elementari e medie: *Prime parole. Dizionario illustrato di base della lingua italiana* (D'Aniello, De Mauro, Moroni 1997) per i ragazzi da 5 a 7 anni, *Dizionario di base della lingua italiana* (De Mauro, Moroni 1996) per i ragazzi da 8 a 11 anni, *Dizionario avanzato dell'italiano corrente* (De Mauro, Moroni 1997) per i ragazzi da 12 a 15 anni. Si tratta di dizionari progressivamente commisurati alle crescenti capacità ed esigenze dei destinatari. I tre dizionari redatti sulla base del *Vocabolario di base* (De Mauro 1980), mirano diventare strumenti essenziali per chiarire dubbi linguistici dall'infanzia fino all'adolescenza. In mancanza per ora di dizionari pedagogici dell'italiano pensati specificamente per un pubblico straniero, questi dizionari restano le raccolte più vaste di parole frequenti e disponibili corredate da definizioni.
Questo primo volume, destinato per bambini agli inizi della scuola (agli allievi delle prime due classi delle elementari) è un dizionario vivacemente illustrato in 608 pagine, comprensivo delle 7.000 parole del *Vocabolario di base*.

2.2.29 De Mauro T., Moroni G. (1997a) *DAIC. Dizionario avanzato dell'italiano corrente,* Torino, Paravia.

Questo dizionario destinato agli allievi delle scuole medie superiori (ragazzi da 12 a 15 anni), contiene oltre le 7.000 parole fondamentali del VdB, altre 13.000 parole "chiave", tra cui espressioni specialistiche e termini delle principali discipline scolastiche. Si tratta di parole funzionali, che aiutano l'orientamento degli alunni nella vita e negli studi, fornendo conoscenze tecniche e produttive. Compilato per scopi didattici, costituisce un ampliamento del DIB. La struttura del dizionario è identica a quella del DIB.
Sia il DIB che il DAIC sono corredati di cd-rom, la cui interfaccia di consultazione è particolarmente chiara. Nell'ambiente principale del programma si possono visualizzare la definizione di un lemma, la fascia di frequenza a cui appartiene, la sillabazione, i sinonimi e i contrari, le parole derivate, la pronuncia delle parole straniere ed eventuali illustrazioni. Si possono scrivere note personali, abbinate a ciascun lemma e salvarle per successive consultazioni. Si può copiare il testo del dizionario e trasferirlo in un programma di videoscrittura. Si possono cercare parole intere e parti e arrivare ai lemmi partendo dalle forme flesse, e ottenere la flessione di ciascun lemma. Si tratta quindi di opere ridotte per mole, ma con buone funzionalità informatiche. Il DAIC contiene un esile prontuario di grammatica, cui si può accedere anche dai lemmi che sono termini grammaticali, inoltre, come il DIB, è accompagnato da una guida didattica.

2.2.30 Putzer O. (a cura di) (1998a) *Bilinguismo. Zweisprachigkeit. Lessico di base. Grundwortschatz*, Bolzano, Arkadia. Putzer O. (a cura di) (1998b) *Lessico approfondito. Aufbauwortschatz*, Bolzano, Arkadia.

I due volumi sono stati pubblicati su incarico della Provincia Autonoma di Bolzano/Alto Adige per supportare la preparazione all'esame di bilinguismo.[32] Il primo volume presenta il "lessico di base", il secondo è dedicato al "lessico approfondito". Le voci sono state selezionate con l'ausilio di liste di frequenza dell'italiano e del tedesco e di lessici di base esistenti, tuttavia non si specificano le opere utilizzate. Si è tenuti conto delle peculiarità della vita quotidiana in provincia e delle finalità particolari, quindi i glossari non sono esenti da scelte soggettive (p. 8). I due volumi, privi di termini grammaticali, sono interamente bilingui (italiano e tedesco) e hanno la stessa struttura. Sono divise in due sezioni, italiano-tedesco e tedesco-italiano, dedicati rispettivamente agli apprendenti di madrelingua italiana e tedesca. Purtroppo le due raccolte non contengono alcun'indicazione sul numero delle voci incluse, poiché "la peculiarità della raccolta non risiede nel numero delle voci elencate, bensì nel modo in cui è stato trattato il lessico" (p. 8). Secondo i nostri calcoli, il carico degli esponenti si aggira attorno a 1.800 per ciascun volume.
Per ogni voce sono riportate le possibili corrispondenze nell'altra lingua, tuttavia gli autori avvertono che i traducenti vanno considerati semplicemente come supporto agli esempi di frase che accompagnano ogni lemma. Gli esempi sono stati scelti in modo da rappresentare "contesti tipici o costruzioni particolari in cui viene solitamente usata la parola", (p. 4). Alcuni gruppi di parole "dal significato corrente", i cui usi e significati non variano nelle due lingue (i giorni della settimana, dei mesi, delle festività, delle lingue ecc.) non sono stati inseriti nel vocabolario in ordine alfabetico, ma in apposite tabelle nella parte centrale dell'opera, tra le due sezioni italiano-tedesco e tedesco-italiano. Ogni voce del glossario è accompagnata da informazioni grammaticali (articolo e plurale dei sostantivi e degli aggettivi, terza persona singolare del passato prossimo dei verbi). Si indicano inoltre le locuzioni più comuni in cui la parola in oggetto appare.

2.2.31 Cresti E., Moneglia M. (a cura di) (2004) *Campionamento Corpus Stammerjohann*.

Questa lista di frequenza tratta dal corpus realizzato nel 1965 a Firenze da Harro Stammerjohann (Stammerjohann 1970), è consultabile al sito dell'Università di Firenze: http://lablita.dit.unifi.it/corpora/descriptions/stam/. La raccolta rappresenta il primo esempio di corpus di parlato spontaneo italiano. Il corpus era stato trascritto dall'autore solo per piccoli campioni e non era fino ad ora disponibile in nessun formato. Il corpus è stato donato a Emanuela Cresti, perché fosse archiviato e poi diffuso nella comunità scientifica. L'intero corpus è stato trasferito in digitale, ed è stato inserito nel Corpus del parlato spontaneo del Laboratorio Linguistico del Dipartimento di Italianistica

[32] Le direttive concordate per gli esami di conoscenza delle due lingue in Alto Adige prevedono un patrimonio linguistico di base distinto per carriere (*einfache Laufbahn* vs. *mittlere und gehobene Laufbahn*). La Giunta provinciale di Bolzano ha provveduto alla selezione dei vocaboli per mezzo di alcuni esperti guidati da P. K. Egger (1977). I volumi contengono i vocaboli di entrambe le lingue, necessari per la carriera esecutiva e per quella direttiva, tenendo presente in particolare la vita pubblica, i termini usati in loco, pertanto sono stati omesse le parole relative alla vita privata. Questi due volumi appartengono alla stessa serie di manuali di bilinguismo, di cui l'opera di Putzer (1998a, 1998b) fa parte.

dell'Università di Firenze (LABLITA), ma è anche disponibile in modo indipendente nel sito del Laboratorio.
Il corpus di ca. 985.000 parole è pari a 30 sessioni registrate di varie lunghezze per circa 41:15:34 ore di registrazione. Le 30 sessioni sono state classificate secondo il canale: broadcasting, telefono, parlato naturale; secondo il contesto sociale: privato, familiare, pubblico; secondo la qualità dell'interazione: libero, regolato; e la struttura dell'evento comunicativo: monologo, dialogo, conversazione.
La lista di frequenza contiene il rango, la frequenza assoluta e la parte del discorso di ciascuno dei 6.182 lemmi. Un altro dato interessante è la Misura di Lessico Fondamentale. Si parte dalla teoria, che i lemmi più frequenti che saturano il 90% delle forme in un corpus di riferimento sono considerati una stima valida per l'identificazione del lessico fondamentale di una lingua. Il corpus Stammerjohann ha un lessico composto da solo 6.181 lemmi proiettati su 100.000 occorrenze elaborate. I lemmi più alti in rango, che saturano il 90% delle occorrenze nel corpus, identificano un lessico fondamentale di appena 1.163 lemmi. Questi dati sono stati confrontati con quelli del corpus di riferimento, il C-ORAL-ROM italiano.[33] Il corpus di riferimento è molto più ampio, contiene 306.600 occorrenze e 15.286 lemmi. Si nota che il numero maggiore di lemmi attestati nel corpus di riferimento determina una maggiore variazione di dominio semantico rispetto al corpus Stammerjohann. Parallelamente, in C-ORAL-ROM, anche il lessico che satura il 90% delle occorrenze è più ampio (2380 lemmi). Dalla stima si deriva che il corpus Stammerjohann non testimonia una variazione di domini e quindi un lessico sufficiente, per essere considerato un corpus di riferimento. Nondimeno, anche in un corpus con queste caratteristiche, è evidente la tendenza alla saturazione del lessico fondamentale su una porzione ridotta del lessico attestato nel corpus: circa 1/5 del lessico attestato in Stammerjohann e circa 1/6 del lessico attestato nel corpus C-ORAL-ROM.

2.2.32 Rinaldi P., Barca L., Burani C. (2004) Caratteristiche semantiche, grammaticali e di frequenza delle parole del "Primo Vocabolario del Bambino." In: *Psicologia clinica dello sviluppo,* 8, pp. 119-143.

Questa ricerca interessante, dedicata al linguaggio dei bambini, si basa sul questionario *Il Primo Vocabolario del Bambino* di Caselli e Casadio (1995). Si tratta dell'adattamento italiano del MacArthur CDI (*Comunicative Developmental Inventory*), un questionario adeguato a raccogliere dati per la valutazione della comunicazione e del linguaggio nei primi anni di vita. Lo strumento consiste in due schede distinte chiamate "Gesti e Parole", per bambini tra gli 8 e i 17 mesi, e "Parole e Frasi" per quelli tra i 18 e i 30 mesi. Esso permette di raccogliere dati sistematici sulla comunicazione e il linguaggio nei primi anni di vita, di valutarli e confrontarli con quelli di altri bambini.
La ricerca di Rinaldi *et al.* (2004) si basa sulle parole riportate nel manuale di Caselli e Casadio (1995), raccolti con circa 700 bambini italiani attraverso l'uso di queste schede. In un file excel (sito: http://www.istc.cnr.it/material/database/pvbvarless.htm), sono contenuti per 527 parole campionate dal *Primo Vocabolario del Bambino*, indici relativi a diverse variabili lessicali. Per ogni parola si indica la categoria grammaticale; la percentuale di bambini che a 30 mesi la produce (valori ricavati da Caselli e Casadio

[33] Questo corpus è stato realizzato nell'ambito del 5° programma quadro dell'Unione europea, che intende fornire alla comunità linguistica una collezione multilingue di risorse comparabili di parlato, per le principali lingue romanze: italiano, francese, portoghese e spagnolo. C-ORAL-ROM è l'acronimo romanzo per il titolo del progetto in inglese: "*Integrated Reference Corpora for Spoken Romance Languages*."

1995); l'immaginabilità, ovvero la facilità di una parola ad evocare un'immagine mentale. I valori di questa variabile sono stati raccolti mediante questionari sottoposti a 50 studenti universitari dai 20 ai 30 anni, i valori vanno da 1 (poco immaginabile) a 7 (molto immaginabile). Si precisa anche la "frequenza nello scritto infantile", tratta dal *Lessico Elementare* (LE) di Marconi *et al.* (1994) (v. 2.2.24), distinta in tre parti: frequenza totale nel corpus di lettura e di scrittura, frequenza delle parole scritte e delle parole lette dai bambini su un corpus di 500.000 occorrenze ciascuno. La frequenza nello scritto adulto è tratta dal CoLFIS (cfr. Laudanna *et al.* 1995 e v. sez. 2.2.25) ed è basata su un corpus di 3 milioni di occorrenze. La "frequenza nel parlato" è tratta dal LIP (v. 2.2.23), ed è basata su un corpus di 500.000 occorrenze. Una trattazione più dettagliata di queste variabili e del loro ruolo nel processing lessicale è contenuta in Rinaldi *et al.* (2004).
La presente base di dati comparativa ha utilità per la ricerca sui processi lessicali, e fornisce uno strumento per lo studio delle caratteristiche semantiche, grammaticali e di frequenza delle prime parole acquisite dai bambini.

2.2.33 Baldelli I., Mazzetti A., Falcinelli M., Servadio B. (2005) *Dizionario fondamentale della lingua italiana per stranieri,* Firenze, Le Monnier.

Il dizionario è composto di circa 5.000 lemmi:

> "scelti tra le parole di più alto uso delle liste di frequenza dell'italiano scritto e parlato, ma anche parole che, pur assenti o di bassa frequenza in tali liste, sono di assoluta 'necessità' per gli studenti stranieri e per gli stranieri che viaggiano in Italia"

come recita la prefazione. Similmente al *Vocabolario minimo*, uscito nel 1974 degli stessi autori Baldelli e Mazzetti (cfr. 2.2.12), il volume è nato presso l'Università Italiana per Stranieri di Perugia. Anche se l'opera non fa alcun riferimento al suo predecessore, vista la sua struttura, è plausibile che si fondi sul *Vocabolario* di Baldelli e Mazzetti, cui sembra che siano stati aggiunti ca. 3.000 vocaboli d'alta frequenza.
Ogni lemma è seguito dall'indicazione della categoria grammaticale, dalla trascrizione fonetica e dalle traduzioni in inglese, francese, tedesco, spagnolo e portoghese. I traducenti si riferiscono sempre al significato illustrato nella prima frase (ovvero all'accezione più "frequente"); in alcuni casi si dà una doppia traduzione per aiutare a capire le diverse accezioni. Seguono poi le forme irregolari, gli esempi (35.000 frasi in tutto), le locuzioni, i modi di dire, i proverbi (ca. un migliaio, secondo la pubblicità dell'opera). Tutte le forme degli articoli e dei pronomi personali sono indicati nell'ordine alfabetico e sono poi trattate nelle Schede Grammaticali (20 pagine), che si trovano in fondo al dizionario, insieme alla coniugazione dei principali verbi.
Trattandosi di un dizionario destinato ad apprendenti d'italiano L2, vi troviamo una serie di espressioni grammaticali, legate a contesti didattici, che pur non essendo di alta frequenza e di "assoluta necessità", a ragione possono essere inclusi in un vocabolario fondamentale per stranieri.

2.3 Tabella riassuntiva e comparativa

Possiamo sintetizzare i risultati del capitolo precedente (escluso tuttavia il § 2.2.11) nella tabella seguente.

	N. dei lemmi*	Defini-zioni, esempi	Parole gramm.	Ampiezza del corpus	Età del corpus	Criteri di selezione			
						Esperienza	Frequenza	Dispersione	Uso
Baldelli, Mazzetti (1974)	1.741	esempi, traducenti	+			+			
Baldelli et al. (2005)	5.000	esempi, traducenti	+			+	+		
Batinti – Trenta (1982)	6.933	lemmi			1920/1989	+	+	+	
Cannella (1991)	15.000 (4.300)	definizioni, esempi	+			+			
CoLFIS (1995)	90.000	formario	+	3.150.075	1992/1994		+		
Cresti, Moneglia (2004)	6.182 (1.163)	lemmario	+	100.000	1965		+		
D. I. L. (1978)	13.000	definizioni, esempi	-			+	+		
DAIC (1997)	20.000 (7.000)	definizioni, esempi	+	500.000	1947/1989	+ (disponibilità)	+	+	
De Mauro, Moroni (1997°)	7.000	definizioni, esempi	+		1947/1989	+ (disponibilità)	+	+	
DIB (1996)	15.000 (7.000)	definizioni, esempi	+	500.000	1947/1989	+ (disponibilità)	+	+	
FDI (1973)	5.014	lemmario formario	+	500.000	1920/1940		≥ 4	≥ 3 sottoinsiemi (≥ 4,88)	≥ 2,39
Gabrielli (1981)	10.000	definizioni, esempi	+			+			
Galli de' Paratesi (1981)	1.400	lemmario esempi	+			+			
Giovanelli (1966)	ca. 4500 (2.000)	esempi, traducenti	+			+			
KBS (1991)	691	lemmario	+	150.000	1987/1990		≥ 20		
Knease (1931)	2.097	lemmario	-	400.000	1850/1930		≥ 9	≥ 9 (testi)	
LE (1994)	6.095	lemmario	+	1.088.725	1987/1989		≥ 3,98	+	+
LIF (1971)	5.356	lemmario formari	+	500.000	1947/1968		≥ 3	≥ 3 sottoinsiemi (≥ 57,19)	
LIP (1993)	15.641 (7.213)	formari	+	496.335	1990/1992		≥ 1 (≥ 3)	5 generi 4 città	+
Migliorini (1943)	1.500 (500)	fraseol., traducenti	+			+			
Piccoli (1986)	10.000	definizioni, esempi	+			+			
Putzer (1998a, 1998b)	ca. 3.600 (1.800)	esempi, traducenti	-			+	+		
Reiske (1974)	1.592	esempi	+			+			
Rinaldi *et alii* (2004)	527	lemmario	+		1994 ca.		+		
Russo G. A. (1947)	3.173	lemmario	±			(+)	(+)	+	
Russo J. A. (1962)	452	idiomi	-	1..920 idiomi	1940	+		≤ 21 (testi)	
Savino (1995)	5.000 (2.000)	definizioni, esempi	+			+	+		
Sciarone (1977)	2.726	lemmario	+	1.500.000	1920/1974		≥ 40		
Skinner (1935)	2.962	traduc. con omonimi	±	19.800		+		≥ 10 (testi)	
Thompson (1927)	500	lemmario	-	100.000			≥ 30	+	+

VdB (1980)	6.690 (2.000)	lemmario	+	500.000	1947/1989	+ (disponibilità)	+	≥ 3 sottoinsiemi	≥ 1,78
VELI (1989)	10.000	lemmario formario	+	26.200.000	1985/1987		+	≥ 0,32	≥ 65

* Tra parentesi il numero di lemmi considerati "fondamentali".

2.4 La parola come segno linguistico nei lessici di frequenza e di base

La frequenza di una parola può variare moltissimo a seconda che si prenda in esame la parola stessa nella forma in cui è registrata nei dizionari (il lemma), le sue forme flesse, il significato principale o significati secondari. Nell'ambito della lessicografia dei dizionario di frequenza e di base esistono ed esistevano sin dai primi lavori, due opposte tendenze rispetto al trattamento dei lemmi. Da un lato gli statistici considerano il segno linguistico come entità monofacciale (sequenza grafica tra due spazi), anziché – con termine saussuriano – bifacciale e lo computano solo come "significante" e non come "significato." Ricordiamo che la Knease (1933: 13) dichiara:

> "As this study has as its aim only the establishment of a word list, the meanings of the words have not been considered."

Il fatto che molti dizionari di frequenza non prendono in esame il significato delle forme lemmatizzate, costituisce un grosso limite ai fini dell'insegnamento delle lingue straniere. In realtà, nella maggior parte delle liste di frequenza, ma anche in molti lessici fondamentali non si danno informazioni sui significati più usati delle parole. Né Thompson (1927), né Skinner (1935), né G. A. Russo (1947), né J. A. Russo (1962), né il LIF (1971), né il FDI (1973), né Sciarone (1977), né il VdB (1980), né Batinti-Trenta (1983), né il VELI (1989), né il KBS (1991), né Rinaldi *et al.* (2004) danno indicazione alcuna sui significati dei lemmi selezionati.
Soffermiamoci sul VdB. Due sono i cambiamenti notevoli avvenuti al suo interno. La prima è che nel 1983 alle parole è aggiunta la categoria grammaticale, mentre nella prima edizione il VdB ne era completamente privo. Questo costituisce un avanzamento, ma non risolve il problema più importante, ossia la mancanza dell'indicazione dell'accezione del lessema da ritenere fondamentale. L'accezionamento del VdB, inteso come indicazione della marca d'uso non per lemma, ma per ogni accezione, avviene per la prima volta nel *Grande dizionario italiano dell'uso* (1999). In questo modo possiamo sapere quali accezioni di un determinato lemma possono essere ricondotte al VdB e a quale fascia appartengono.
Gli altri lavori si limitano per lo più alla disambiguazione degli omografi grammaticali e (molto raramente) di quelli lessicali. Delle 32 liste prese in esame sono 12 (6 dizionari fondamentali e 6 di frequenza) che non permettono di reperire il significato degli elementi lessicali. De Mauro (2003: 144) aveva evidenziato la problematicità del mancato accezionamento del vocabolario di base:

> "La grande maggioranza delle parole del vocabolario comune e di base ha una pluralità di accezioni. Chi studia una lingua da un punto di vista statistico, ha potuto stabilire che quanto più una lingua è usata, tanto più numerose sono le sue accezioni. Così, ogni parola può essere fonte di equivoci."

I dizionari destinati ad uso didattico cercano di rimediare a questo problema. Baldelli-Mazzetti (1974) e Baldelli *et al.* (2005) precisano i significati della loro lista di 1.741 lemmi, ricorrendo a spiegazioni interlinguistiche con l'inglese, il francese, il tedesco, lo

spagnolo (e il portoghese nell'edizione del 2005). Similmente Migliorini (1943), Giovanelli (1966), Reiske (1974) e Putzer (1998a, b) forniscono traducenti in tedesco e contesti esemplificativi. Il Piccoli (1986) e Cannella (1991) danno una breve definizione ed esempi d'uso. Galli de' Paratesi (1980) dà esempi e contesti esplicativi. Il DIB (1996), il DAIC (1997a) e le *Prime parole* di De Mauro (1997b) riportano informazioni grammaticali, definizioni ed esempi d'uso. Tra i dizionari di frequenza solo quelli più recenti consentono di verificare i significati dei lemmi registrati. Il LIP (1993) riporta in due dischetti IBM tutti i testi delle registrazioni trascritte ortograficamente, ed i due corpora, da cui sono state tratte le liste del CoLFIS (1995) e di Cresti e Moneglia (2004), sono consultabili in Internet. Evidentemente i compilatori si sono resi conto che la reperibilità dei significati e dei contesti è una condizione fondamentale dell'attendibilità dei risultati.

Un altro aspetto problematico è la lemmatizzazione delle voci estratte dal corpus. Questa elaborazione è indispensabile per il calcolo della distribuzione di diversi elementi del vocabolario secondo la frequenza, poiché nessun conteggio potrà fornire risultati attendibili che manchi di fare una distinzione tra avverbio ed aggettivo, nome e verbo ecc. La lemmatizzazione nelle opere più recenti è, per lo più, automatica, quindi ha la caratteristica di fornire tutte le interpretazioni possibili di ogni forma, per quanto siano inverosimili nella realtà: p. es. la forma *venti* rimanda non solo al numero e al sostantivo *vento*, ma anche al raro verbo *ventare*. Da questo processo di lemmatizzazione si raccolgono anche lemmi che non si desiderano avere nel vocabolario. Pertanto, i criteri usati per la classificazione delle parole (a volte prive di un contesto) sono spesso divergenti e contraddittori. Grazie alla lemmatizzazione automatica nel VELI, ad esempio, sono considerati come lemmi separati (anziché uno solo) alcune forme: *e/ed, o/ od, qual/quale, alcol/alcool, un/uno/una, del/dello/della/dei/degli/delle, oramai/ormai* ecc.

Centrale è anche il trattamento degli omografi. De Mauro (1994) riferisce per l'italiano un tasso di omografi che varia dal 38% nei testi economico-finanziari al 46% nella lingua parlata. Nel LIF gli omografi ma non omofoni sono separati con l'aggiunta di un accento grafico; nel caso degli omografi appartenenti a classi diverse di parole è stato sufficiente indicare la categoria grammaticale; nel caso degli omonimi effettivi, in cui una stessa forma esprime due o più significati senza che la funzione morfosintattica ne sia modificata, gli autori hanno seguito le suddivisioni date dal vocabolario (senza ricorrere alle indicazioni che il contesto avrebbe potuto fornire). Il FDI (1973) nell'introduzione dedica una lunga sezione (pp. XXIII-XXXV) al trattamento degli omografi. Si distinguono i vari significati degli omonimi lessicali in base al contesto, mentre la categoria grammaticale degli omonimi sintattici è chiarita in base alla loro funzione svolta nella frase. Nel LE (1994) gli omografi sono stati lemmatizzati manualmente, quindi sono stati riportati separatamente gli omografi funzionali, cioè i lemmi graficamente identici ma con doppia categoria grammaticale, mentre non si è entrati nel merito dell'omografia semantica.

Esiste un numero indefinito di parole che possono essere classificate in più di una categoria grammaticale. Parole come *abbondante, brillante, divertente* ecc. a seconda dell'opinione del ricercatore possono essere classificate come aggettivi o participi presenti. Il FDI ha optato per la prima soluzione, il LIF per la seconda. Forme come *abituato, acceso, agitato* ecc. possono essere classificate come verbali (LIF) o come aggettivi (FDI). Il FDI impiega la distinzione tra uso sostantivale e aggettivale di *alto, amante, americano* ecc., mentre il LIF non fa distinzioni del genere come nomi oppure come aggettivi. Nell'inclusione delle parole nella lista di frequenza e nella determinazione del loro rango, i criteri di classificazione e la coerenza nella loro

osservazione appaiono determinanti. Se *seguente*, ad esempio, è considerata una forma del verbo *seguire*, verrà a mancare l'aggettivo, ma in compenso il verbo *seguire* occuperà un rango più alto. Nel LE (1994) la disambiguazione aggettivo-participio passato, aggettivo-avverbio è stata risolta assegnando manualmente al termine la categoria grammaticale di appartenenza.
Un'altra questione interessante è il trattamento delle locuzioni. Il VELI ha segnato "esplicitamente la locuzione, invece del vocabolo, se oltre la metà delle occorrenze di quella parola avvenivano in quella specifica locuzione" (p. 55). Per esempio la locuzione *in quanto* non appare nel VELI, poiché, pur essendo molto frequente, ciascuno dei vocaboli che la compongono (*in* e *quanto*) hanno una frequenza più che doppia di quella della locuzione stessa, infatti, appaiono come termini separati. Il LE (1994: 19) ha lemmatizzato come locuzioni le parole il cui significato è dato dalla "composizione stessa", e poi sono state suddivise in locuzioni preposizionali (*a seconda di*) e locuzioni avverbiali (*di solito*). Il LIF ha riunito solo alcune locuzioni che nei testi spogliati avevano una grafia unita o separata, del tipo *fin ché* (*finché*) e quelle in cui l'elemento semanticamente più importante non esiste più indipendentemente, come: *a galla, a malapena*. Nel LIP (1993: 95), invece, è stata adottata una definizione più larga dei lemmi polirematici. Si considerano espressioni polirematiche tutte le parti di enunciato il cui uso si è cristallizzato (*opinione pubblica*); il cui significato non può essere scomposto in quello dei suoi membri (*faccia di bronzo*); la cui funzione grammaticale non è deducibile dalla classe grammaticale degli elementi che la formano, ma è data dall'insieme dei suoi membri (*per quanto*). Pertanto il LIP registra ca. 6800 voci polirematiche.
Infine, bisogna notare che sono proprio i dizionari basati su criteri empirici che di norma non contengono le parole grammaticali, il che comporta la distinzione fra parole autosemantiche e parole sinsemantiche. L'eliminazione degli elementi grammaticali, pur sopprimendo l'estrema difficoltà di classificazione di questi elementi, ne pone un'altra: spesso è veramente difficile tracciare un limite netto tra elementi significativi ed elementi di relazione.
Perfino questa rassegna breve e schematica fa capire che il trattamento delle parole scelte è poco uniforme, il che, insieme alle divergenze dei criteri di selezione, spiega le discrepanze tra le varie liste.

2.5 Riassunto, prospettive

Alla fine di questo capitolo occorre rilevare, che esiste un numero consistente di realizzazioni lessicografiche dedicate alla statistica linguistica e alla selezione dei lessici di base. Le parole più utili, più essenziali per la conoscenza della lingua sono state selezionate in vari modi, per cui disponiamo di vari tipi di raccolte.
Nonostante i criteri di individuazione non siano sempre omogenei, c'è un'ampia convergenza nella determinazione di questo primo stato della lingua che ne costituisce lo zoccolo duro. Il numero delle parole fondamentali tende ad oscillare intorno alle 2.000. Tra un minimo di 1.400 parole di Galli de' Paratesi 1981 e un massimo di ca. 3.200 voci di Russo 1947 troviamo le opere di Knease 1931, Skinner 1935, Migliorini 1943, Giovanelli 1966, Baldelli e Mazzetti 1974, Reiske 1974, Sciarone 1977, De Mauro *et al.* 1980, Putzer 1998a.

I lavori con un numero inferiore di lemmi presentano alcune peculiarità: le liste di Thompson 1927 e di Katerinov 1991, basate entrambe su calcoli di frequenza, a causa della ristrettezza del corpus presentano soli 500 e 700 lemmi, rispettivamente. La lista di J. A. Russo 1962 è una raccolta di 500 locuzioni, ricavate da manuali didattici per l'italiano L2, mentre quella di Rinaldi *et al.* (2004) è una lista di frequenza dei termini usati dai bambini dai 18 ai 30 mesi. Pertanto nessuno dei quattro lavori menzionati può essere considerato rappresentativo del vocabolario fondamentale, bensì presentano una specie di lessico elementare che certamente non è adatta a soddisfare le esigenze comunicative di parlanti adulti.
I lavori con più di 3.200 di lemmi, sono dizionari scolastici dedicati a bambini con madre lingua italiana (Giovanelli 1966, D. I. L. 1978, Batinti e Trenta 1982, Cannella 1991, Gabrielli 1981, Piccoli 1986, Savino 1995, DIB 1996, DAIC 1997) oppure sono liste di frequenza (LIF 1972, FDI 1973, VELI 1989, LIP 1993, LE 1994, CoLFIS 1995, Cresti e Moneglia 2004). L'unico esempio di un dizionario fondamentale destinato a studenti stranieri con più di 3.200 lemmi, e precisamente 5.000 è Baldelli *et al.* (2005).
Mentre nel caso dei vocabolari compilati sulla base dell'esperienza non ci sono limiti pratici dell'ampiezza del vocabolario, per delimitare il lessico fondamentale sulla base delle liste di frequenza, occorre identificare i lemmi che hanno una frequenza stabile, quelli cioè la cui frequenza relativa corrisponde alla probabilità di occorrenza. Nei diversi dizionari di frequenza, il numero dei lemmi di frequenza attendibile varia secondo la grandezza del campione utilizzato. Ciò si spiega con la legge dei grandi numeri secondo cui, solo al crescere del numero delle prove (al crescere del campione corpus), la frequenza di un evento tende a diventare uguale alla sua probabilità di occorrenza.
Le prime due liste che, almeno parzialmente, riescono a superare le difficoltà derivanti dalla scelta del campione, il LIF e FDI, basati entrambi su un corpus di 500.000 occorrenze, differiscono pochissimo nei primi 1.000-1.500 lemmi: la percentuale dei lemmi comuni è 84,95% per i primi 1.000 (Katerinov 1991:16). In base alla legge dei grandi numeri però, dei circa 5.000 lemmi presenti nei due insiemi, meno di un terzo si può considerare rispondente alla definizione lessico fondamentale. L'elenco più ampio di lemmi, la cui frequenza può considerarsi indipendente dal tipo di testi analizzati, è quello offerto da Sciarone, che da un corpus di 1.500.000 occorrenze ha estratto oltre 2.725 lemmi. Questi lemmi con frequenza assoluta molto elevata si possono considerare a frequenza stabile (tale cioè, che la frequenza relativa sia vicina alla probabilità di occorrenza). Sembra quindi, che pur ampliando il campione adottato, la quantità di lemmi identificati come fondamentali, aventi una frequenza d'occorrenza stabile, non cresce in misura proporzionale al crescere del corpus, e continua ad oscillare attorno a 2.000. Allo stesso risultato si giunge se sulla scia di Cresti e Moneglia (2004) riteniamo fondamentale quella sezione dei lemmi di alta frequenza che satura il 90% di tutte le occorrenze: in un corpus di 300.000 occorrenze, il C-ORAL-ROM, le parole fondamentali sono 2.380, mentre nel LE, tratto da 1.000.000 di occorrenze, le parole che coprono il 90% di tutte le occorrenze sono 1.405 (si ricorda che si tratta di testi scritti da e per bambini). Nel LIF e nel LIP i primi 1.500 lemmi (quelli aventi una frequenza attendibile) coprono 88,21% e 89,07% del corpus, mentre i primi 2.000 (la fascia che il *Vocabolario di base* di De Mauro considera il lessico fondamentale) lo coprono per il 90,69% e 91,08%, rispettivamente (LIP 1993: 122). Sciarone, partendo dalla stessa idea che per capire un testo si deve disporre di una lista che lo copra per il 90%, per la propria lista calcola un valore di copertura di 88,73%, pertanto, per ottenere la copertura desiderata, occorrerebbe ampliare il lemmario, aggiungendovi altre 500 parole (Sciarone 1995: 70-72).

Possiamo dunque concludere che i dizionari fondamentali, basati su criteri in parte soggettivi e in parte costituiti con metodi più rigorosi, sembrano concordare nel fatto che la fascia del lessico fondamentale sia costituita da circa 2.000 parole. È quanto rivelato da De Mauro (1980), il primo studioso che in un sistematico lavoro ha cercato di definire il lessico fondamentale, integrando l'approccio frequenziale con quello basato sull'esperienza. Proprio questa integrazione dei due approcci diversi garantisce l'attendibilità dei lemmi riportati nel *Vocabolario di base della lingua italiana* (1980). Grazie anche alle ripetute verifiche cui è stato sottoposto, questa raccolta rappresenta la selezione più plausibile del lessico fondamentale dell'italiano. Essa sta alla base di numerosi dizionari scolastici (*Prime parole*, DIB, DAIC – si tratta dei dizionari che meglio si addicono alle esigenze degli apprendenti di italiano L2) e non (*Dizionario della lingua italiana per il terzo millennio*, Paravia, 2000; *Grande dizionario italiano dell'uso*, UTET, 1999), inoltre funge da vocabolario di riferimento per la redazione di testi di difficoltà controllata in Italia (cfr. 4.2). Per tutti questi motivi, costituisce una base valida per selezionare il lessico fondamentale da insegnare agli apprendenti di italiano L2 da un lato (cfr. 6.4), e per costruire prove d'esame con cui valutare la competenza in italiano L2 come vedremo nella sez. 5.2.

Grazie allo sviluppo delle tecniche di elaborazione automatizzata dei testi, recentemente si sono creati larghi corpora che permetteranno, tra l'altro, la creazione di nuovi dizionari di frequenza e/o di base. Uno di questi è il CoLFIS (cfr. sez. 2.2.25), il primo di cui è stata estratta anche una lista di frequenza, ma esistono una serie di iniziative volte alla creazione di database testuali della lingua scritta e orale. Senza voler fornire un quadro completo, tra i corpora dedicati alla lingua scritta occorre ricordare il *Corpus di Italiano Scritto contemporaneo* (CORIS) elaborato tra il 1998 e il 2001 all'Università di Bologna. Il corpus contiene 100 milioni di parole, tratte dalla narrativa prodotta negli anni Ottanta e Novanta e sarà aggiornato ogni due anni con nuovo materiale di controllo. Al Coris si aggiunge il *COrpus Dinamico dell'Italiano Scritto* (CODIS) che permette la selezione ed eventuale esclusione di sottocorpora considerati non rilevanti per specifiche ricerche. Il corpus è disponibile su cd-rom e per consultazione online. Ricordiamo ancora la *Letteratura Italiana Zanichelli* (LIZ) in cd-rom, disponibile dal 1993 e il *Tesoro della lingua italiana delle origini* (TLIO). Questo database testuale, nato nel 1995 e inaugurato nel 1998, è composto di circa 1.780 testi per circa 20 milioni di parole, tratte da scritti in lingua italiana prima del 1375, in prosa e in poesia.

Per la lingua parlata vanno menzionati i progetti di raccolta di materiale fonico spontaneo di lingua italiana, cui hanno partecipato il laboratorio di linguistica della Scuola Normale di Pisa, il CIRASS e l'Orientale di Napoli, il Politecnico di Bari e l'Università del Piemonte Orientale. Si tratta dei progetti AVIP (*Archivio delle Varietà di Italiano Parlato*, 1999), API (*Archivio di Italiano Parlato* 2001), CLIPS (*Corpora e Lessici di Italiano Parlato e Scritto*, 2003), IPar (*Italiano Parlato*, 2003).[34] Presso l'Università per Stranieri di Perugia è in corso la creazione del *Corpus di Italiano Televisivo* (CIT), mentre presso il Centro di Studi di Grammatica Italiana dell'Accademia della Crusca è stato elaborato un progetto per la realizzazione di *Lessici di Frequenza dell'Italiano Radiofonico* (LIR). Occorre ricordare ancora il progetto C-ORAL-ROM (2005), gestito dal Laboratorio Linguistico del Dipartimento di Italianistica dell'Università di Firenze, che prevedeva la costruzione di corpora di riferimento della lingua parlata per le principali lingue romanze. Attualmente è in fase di realizzazione anche un lessico di frequenza della *Lingua dei segni italiana* (LIS) (sui problemi metodologici relativi cfr. Russo T. 2005).

[34] Per una descrizione dei tre corpora AVIP, API e CLIPS cfr. Albano Leoni (2003).

Considerata questa fervente attività di raccolta di corpora orali e scritti, si prevede che nel futuro si moltiplicheranno le pubblicazioni di linguistica statistica, dedicate agli aspetti lessicali dell'uso linguistico. Ci auguriamo che delle nuove rivelazioni anche la lessicografia dei dizionari fondamentali possa trarre dei vantaggi.

3 I dizionari fondamentali e di frequenza della lingua ungherese (1933-2006)

Per il nostro obiettivo, ovvero per delimitare il lessico di base da proporre agli apprendenti ungheresi d'italiano L2 nei primi anni dello studio, a prima vista può sembrare superfluo un esame approfondito dei dizionari fondamentali e di frequenza della lingua ungherese. Movendo però dal principio che la lingua è fondamentalmente un mezzo di comunicazione, è importante precisare che un corso di lingua straniera realizza un duplice scopo: da un lato "esporta" il soggetto parlante nel Paese straniero, dall'altro "importa" la lingua e la cultura straniera nel Paese stesso dell'apprendente. Pertanto, se da un lato è vero che l'apprendimento di strutture linguistiche autentiche è necessario per l'uso appropriato della lingua, è altrettanto vero che occorre tener presente anche il valore d'uso degli elementi lessicali nel Paese stesso dell'allievo. Non è, infatti, opportuno attribuire un'importanza esclusiva alle strutture autentiche della L2, bensì occorre che gli autori dei manuali e gli insegnanti considerino anche l'uso di una parola anche nel Paese d'origine. E qui ci ricolleghiamo al concetto di disponibilità, poiché parole altamente disponibili per l'apprendente di una L2 non possono rimanere fuori del suo bagaglio lessicale.
Scopo di questo capitolo è di trovare quei lavori che potrebbero servire come liste di riferimento, per arricchire il lessico fondamentale dell'italiano L2 di quegli elementi lessicali che gli apprendenti ungheresi possono avere la necessità di esprimere.

3.1 Autori, finalità e criteri di selezione

La prima lista di frequenza dell'ungherese, proprio come la prima disamina quantitativa della lingua tedesca (Käding 1899), nasce negli anni Trenta per scopi stenografici (Nemes 1933), poco dopo la pubblicazione della prima lista di frequenza per l'italiano (Thomson 1927). In questo periodo escono altre due opere importanti di Cser (1939) e di Nemes (1941), segue poi una lunga fase, dagli anni Quaranta fino agli anni Ottanta, in cui gli studi sulla frequenza sembrano essere abbandonati dalla comunità scientifica. La pubblicazione della lista di Csirikné e Csirik nel 1986 ha rilanciato l'elaborazione delle statistiche lessicali, ma uno sviluppo intenso degli studi in materia si è verificato solo negli ultimi anni, dal 2000 in poi (Csirikné e Csirik 2003; Bácsi e Kerekes 2003, WordSword 2004).
La maggior parte delle liste è stata elaborata per esigenze pratiche: per sviluppare i sistemi stenografici (Nemes 1933 e 1941), per l'insegnamento dell'ungherese lingua materna (Cser 1939, Csirikné e Csirik 2002; Bácsi e Kerekes 2003), per sviluppare programmi di controllo ortografico e di analisi morfologica (WordSword 2004) e solo raramente per scopi linguistici (Csirikné, Csirik 1986; Füredi, Kelemen 1989). I corpora utilizzati intendono essere rappresentativi del linguaggio parlamentare (Nemes 1933), di quello infantile (Cser 1939, Csirikné, Csirik 2002; Bácsi, Kerekes 2003), giornalistico (Nemes 1941; Csirikné, Csirik 1986), letterario (Füredi, Kelemen 1989) o del linguaggio di Internet (WordSword 2004). Le prime liste sono state compilate in modo manuale

(Nemes 1933, Cser 1939, Nemes 1941), mentre dagli anni Ottanta in poi i compilatori si sono avvalsi di calcolatori elettronici.
Il criterio di selezione adottato è, in ogni caso, (salvo Aradi, Erdős e Sturcz 2000), la frequenza d'occorrenza assoluta o quella modificata da altri fattori. L'uno lessico fondamentale della lingua ungherese, infatti, si basa sull'esperienza degli autori.

3.2 Cronologia dei lessici fondamentali e di frequenza dell'ungherese

Presentiamo di seguito in ordine cronologico i dizionari di frequenza e l'unico lessico fondamentale della lingua ungherese. Per la redazione di questo capitolo ci siamo serviti di uno studio di Lengyelné Molnár (2006), in cui l'autrice passa in rassegna i dizionari di frequenza pubblicati in Ungheria. Tralasciamo, così come l'abbiamo fatto anche nel caso dei lessici italiani, i dizionari inversi ed altri lavori dedicati all'analisi frequenziale di strutture grammaticali, fonemi, lettere ecc.[1] e le concordanze di opere letterarie.[2]
Per ogni dizionario si forniscono i dati del corpus sottostante, i criteri di selezione, il numero di lemmi inclusi, i gruppi di destinatari ed altri dati rilevanti per dare un panorama delle divergenze e le convergenze delle liste pubblicate.

3.2.1 Nemes Z. (1933) *A magyar parlamenti nyelv leggyakoribb szavai*. Az Egységes Magyar Gyorsírás Könyvtára 66., Szeged.

La prima statistica lessicale ungherese si deve allo stenografo Zoltán Nemes. L'Associazione Nazionale degli Stenografi Ungheresi nel 1927 ha incaricato il segretario del dipartimento di stenografia del Parlamento di compilare la prima statistica dei suffissi e delle desinenze della lingua ungherese (*A magyar nyelv képző-, jel- és ragstatisztikája*. Az Egységes Magyar Gyorsírás Könyvtára 57., Szeged, 1932). Il lavoro aveva uno scopo pratico: semplificare ed abbreviare il logogramma delle parole più frequenti, per facilitare il lavoro degli stenografi del parlamento.
Questa statistica lessicale pubblicata in un secondo tempo, nel 1933, è stata ricavata dallo stesso corpus che fungeva come base per la compilazione della statistica delle desinenze. Questa lista contiene le 505 parole più frequenti del linguaggio parlamentare, tratte da una raccolta di 40.977 occorrenze, vale a dire 6.571 forme (Nemes 1933: 20). Il corpus di 100.000 sillabe è costituito da 200 brani di discorsi parlamentari, ciascuno di 500 sillabe.

[1] Il dizionario inverso della lingua ungherese, pubblicato nel 1969 a cura di F. Papp, contiene in ordine alfabetico inverso i 58.323 lemmi del Dizionario enciclopedico ungherese (*Magyar Értelmező Szótár* 1959-1962) e include dati statistici sulla frequenza dei fonemi, lunghezza e numero di accezioni delle parole. La prima statistica delle lettere e dei fonemi ungheresi si deve a Simonyi (1896). Szende (1973) nel suo lavoro (*Spontán beszédanyag gyakorisági mutatói*) analizza la frequenza e la distribuzione dei fonemi ungheresi, tratti da un corpus di parlato spontaneo di 18.000 occorrenze.
[2] Spicca tra questi lavori la concordanza di tutte le poesie del poeta Sándor Petőfi, pubblicato in quattro volumi tra 1973-1987 (a cura di Gáldi, e poi Soltész *et al.*). Il repertorio alfabetico contiene 22.719 parole usate, accompagnate da un breve contesto recante l'indicazione dei luoghi dove queste figurano. Altri autori la cui produzione completa è stata elaborata dal punto di vista lessicale sono Gyula Juhász (Benkő 1972), Csokonai (Jakab, Bölcskei 1993), Jókai (Balázs *et al.* 1994) Balassi (Jakab, Bölcskei 2000), Zrínyi (Beke 2004). Per le altre concordanze letterarie cfr. Lengyelné Molnár (2006).

Trattandosi di un lavoro volto a migliorare i sistemi stenografici, ogni forma è stata lemmatizzata in modo particolare: tutte le parole sono state analizzate in tema e desinenza (flessioni, suffissi), e i conteggi sono stati effettuati per ciascun tema e desinenza. Visto che nella stenografia il logogramma di un tema può essere utilizzato per abbreviare tutte le forme che iniziano con una determinata sequenza grafica, anche quando la derivazione diretta non è sostenibile dal punto di vista morfologico e/o etimologico, la lemmatizzazione a volte ha dato dei risultati assai strani. Ad esempio dal tema *áll-* si fa derivare *állam, állás, állandó*, da *terem-* abbiamo *természet, termel*, ecc. Nella lista figurano 26 temi che nella lingua non esistono come parola autonoma. In più, anche le voci composte sono state analizzate in costituenti: ad esempio la parola *magyar* ('ungherese') occorre 218 volte nei testi, di cui 177 volte come parola semplice, 37 volte primo elemento di un composto, 3 volte come secondo elemento e una volta come ultimo costituente (p. 21). Quindi tutte le parole composte e derivate sono analizzate nei loro costituenti, tranne che esse abbiano una frequenza pari o superiore a 25.
Le 505 voci con una frequenza assoluta da 25 in su, sono state disposte in ordine frequenziale (pp. 23-27) e alfabetico (pp. 28-41). Nella lista ogni parola è preceduta dal rango, poi segue la frequenza assoluta del tema e quella delle sue forme flesse e derivate, nell'ultima colonna si riportano le forme flesse e derivate aventi una frequenza superiore a 10. Al carattere particolare dei testi si deve l'estrema frequenza di voci come *primo ministro, governo, deputato* ecc. Sono riportate in una lista apposita anche le parole che più frequentemente figurano in composti (pp. 42-43). Per ogni parola è indicato il numero d'occorrenza totale, la frequenza come parola semplice e il numero d'occorrenza in composti, in cui la parola è il primo, secondo o l'ultimo elemento. Sempre dalla tipologia dei testi consegue che dopo i prefissi verbali, la parola che più frequentemente entra in composti è *ministro*.
Dai conteggi emerge che le prime 111 parole coprono più del 50% di tutte le occorrenze, e le 505 parole della lista rappresentano il 66,73% dei testi. Nemes (1933: 46) ne deduce che l'ungherese sia una lingua più ricca di quelle indoeuropee, poiché esse si avvalgono di un numero inferiore di parole per coprire la stessa percentuale di testi.

3.2.2 Cser J. (1939) *A magyar gyermek szókincse. Gyakorisági és korszótár*, Budapest, Magyar Pedagógiai Társaság.

Si tratta del primo lavoro statistico dedicato al lessico dei bambini ungheresi.[3] L'obiettivo è di dare un ausilio in particolare agli insegnati di lingua materna e straniera e agli autori dei libri didattici. In secondo luogo l'opera è dedicata a tutti coloro che si occupano di pediatria, psicologia infantile, psicopedagogia, di difesa dell'infanzia, infine agli addetti ai tribunali minorili.
Al fine di trovare i vocaboli appartenenti al lessico attivo dei bambini, l'autore ha condotto un esperimento associativo, proposto per la prima volta nel 1927 dall'americano Buckingham. A 4.483 ragazzi tra i 10 e 14 anni hanno chiesto di scrivere in 15 minuti il maggior numero possibile di parole che gli venivano in mente, evitando di ripetere le voci e di dare nomi propri, articoli e parole declinate o coniugate. In questo modo è nato un corpus di 400.501 occorrenze. Per l'enorme quantità del materiale, l'autore ha esaminato solo una parte della raccolta: esattamente 89.705 occorrenze, prodotte da

[3] Un altro lavoro dedicato alle caratteristiche fonologiche e morfolessicali del linguaggio dei bambini di tre anni è lo studio di Gósy (1984). Non si tratta di un dizionario di frequenza vero e proprio, ma l'opera include anche dati statistici sul linguaggio infantile.

1.000 bambini. Nella lista sono state incluse le voci scritte da almeno due ragazzi all'interno della stessa fascia d'età.
Nella prima parte dell'opera, i dati quantitativi sono stati comparati a quelli delle ricerche condotte per l'inglese americano (Dolch 1931) e il tedesco svizzero (Prescott 1929). Si costata che sebbene i bambini ungheresi in 15 minuti abbiano scritto meno parole dei ragazzi americani e svizzeri (e quindi, secondo l'autore sono più "lenti" dei loro coetanei stranieri), il numero di parole diverse era sensibilmente più alto presso i ragazzi ungheresi. In altri termini, nonostante loro abbiano scritto meno parole, esse erano più variegate, dimostravano maggiori differenze individuali e meno ripetizioni. L'autore sulla scia di Nemes (1933: 46) afferma che l'ungherese sia più variegato delle lingue straniere, poiché usa un maggior numero di parole diverse. L'idea è sostenuta dal fatto che mentre nel corpus di Cser le prime 377 parole, aventi una frequenza pari o superiore a 50, coprono soltanto il 64,6% di tutte le occorrenze, in Käding le prime 250 voci costituiscono il 75% del corpus.
È stata verificata la distribuzione delle 377 voci più frequenti tra i due sessi, e n'è stato ricavato che le parole menzionate più spesso dai maschi erano legate a temi come l'arte militare, le occupazioni, gli animali, la tecnica, mentre tra le femmine i temi più popolari erano l'uomo, i cibi, fiori, colori, abbigliamento, casa, arredamento. Da un sistematico confronto con Prescott (1929) emerge che le parole più frequenti nell'uso dei ragazzi svizzeri sono legate alle stesse tematiche, e in molti casi perfino la frequenza dei singoli elementi tende a coincidere.
Nella seconda parte del volume troviamo due liste:

a) la prima riporta in ordine alfabetico le 3.767 parole aventi una frequenza assoluta di 2 o superiore. Per ogni voce si indica la fascia d'età in cui il vocabolo appare per la prima volta, la frequenza assoluta tra ragazzi e ragazze rispettivamente e la frequenza assoluta cumulata.
b) La seconda lista contiene le 377 voci più frequenti, suddivise in tre fasce: parole con frequenza superiore a 200, voci con frequenza tra 100 e 200, infine vocaboli con frequenza tra 50 e 100.

Occorre rilevare che l'opera rappresenta tutt'oggi un punto di riferimento valido per chi voglia cimentarsi nello studio del lessico dei bambini ungheresi, non solo per la quantità di dati elaborati, ma soprattutto per il rigore metodologico con cui è stato condotto l'esame, in un'epoca in cui non si poteva usufruire dell'ausilio di calcolatori elettronici.

3.2.3 Nemes Z. (1941) *Szóstatisztika egymillió szótagot felölelő újságszövegek alapján.* Az Egységes Magyar Gyorsírás Könyvtára 190., Szeged.

È il primo dizionario di frequenza dedicato al linguaggio giornalistico. L'autore voleva estendere i suoi esami precedenti (Nemes 1932, 1933) ad un corpus più vasto, perciò ha raccolto 453.578 occorrenze (p. 12) (401.000 o 397.956 secondo Csirikné e Csirik 1986: VIII, XLI; 473.813 secondo Szende 1973: 5, ma sicuramente non 1.000.000 come lo afferma Fóris 2002: 117). Il numero di sillabe elaborate invece ammonta a un milione, come lo annuncia il titolo dell'opera. Il corpus è stato tratto da articoli di ciascuno dei 26 giornali pubblicati a Budapest nel periodo tra aprile 1933 e aprile del 1935 (p. 9). Per un periodo di due anni l'autore ha elaborato mensilmente una o due copie di ogni giornale, escludendo in questo modo le notizie di attualità. Il lavoro ha durato sei anni, poiché l'autore ha spogliato manualmente l'enorme corpus la cui sola sfogliatura richiedeva 14 ore intere, mentre per leggerla ci volevano 67 ore (Nemes 1941:

8). Per primo ha registrato ogni forma su una cartella, successivamente le ha ordinate alfabeticamente, infine le ha ricondotte ai loro costituenti.
La lemmatizzazione è stata condotta in maniera simile che in Nemes (1933): tutte le parole sono state analizzate nei loro costituenti, le forme flesse e derivate sono state ricondotte al tema, i composti sono stati analizzati nelle loro parti. Anche qui a volte troviamo soluzioni inattese: le parole *vető, vetendő, vetés, vetemült, vetkőz* ecc. sono state tutte ricondotte al tema *vet*. Grazie a questo tipo di lemmatizzazione le 31.078 parole diverse del corpus sono state ridotte a 16.571 voci, mentre senza la scomposizione delle parole complesse si sono ottenute 17.599 termini.
L'opera si articola nel modo seguente:

a) La parte principale del volume riporta in ordine frequenziale le parole aventi una frequenza d'occorrenza da 25 in su (pp. 49-79). Questa prima lista si suddivide in due parti: per le 731 parole con frequenza superiore a 100, si precisa la frequenza d'occorrenza come parola autonoma e quella in composizioni, come primo, secondo, terzo elemento, infine segue la frequenza totale. Per le altre 2.238 parole, con una frequenza superiore a 25, si indicano la frequenza assoluta e fra parentesi il numero di occorrenze come parola semplice. Si osserva che queste ca. 2.000 voci coprono l'87,4% del corpus.
b) La seconda è una lista alfabetica delle ca. 7.650 parole che figurano almeno 5 volte nel corpus (pp. 80-133). Nella prima colonna si riporta il numero di occorrenze come parola semplice e come costituente di una parola composta. Nelle due liste non sono stati inclusi i nomi propri e i numerali.

Accanto alle due liste principali nell'opera troviamo la statistica delle prime lettere delle parole ungheresi e straniere, il numero di sillabe, lunghezza delle parole, la distribuzione delle voci di origine straniera, ecc. con e senza la scomposizione delle parole composte (pp. 10-48). Altre tabelle (pp. 137-392) contengono le forme flesse e derivate più frequenti di ogni tema (pp. 137-261), la frequenza dei verbi prefissati (pp. 266-304), la distribuzione delle parole nelle varie fasce di frequenza secondo la struttura morfologica, numero di sillabe, i nomi propri più frequenti ecc.[4]
Concludendo, si tratta di un'opera pionieristica nel suo genere, poiché l'autore non solo ha elaborato manualmente una considerevole quantità di dati, ma li ha analizzati sotto ogni aspetto immaginabile, rendendo trasparente i metodi di selezione adottati, il che tuttora non è una pratica comunemente diffusa tra i compilatori dei lessici di frequenza (e tra i lessicografi in genere).

3.2.4 Csirikné Czachesz E., Csirik J. (1986) *Újságnyelvi gyakorisági szótár*. Szeged, Juhász Gyula Tanárképző Főiskola.

Il dizionario di frequenza del linguaggio giornalistico ungherese è uscito in formato trilingue: in ungherese, russo e inglese. È destinato per essere utilizzato nell'insegnamento della lingua materna, nella didattica dell'ungherese come lingua straniera, infine a linguisti, letterati e statistici che si interessano degli aspetti quantitativi del lessico ungherese (pp. VI-VII). Le liste sono state compilate da una coppia di coniugi, esperti di matematica e di pedagogia.

[4] Nemes dà anche un panorama internazionale degli studi sulle frequenze, menzionando che l'unica iniziativa italiana dell'epoca, l'impresa di Giovanni Aliprandi, non è stata portata a termine. Aliprandi intendeva compilare una lista di frequenza sulla base di un milione di sillabe, ma soltanto i risultati relativi alle lettere *a, e, i, o, u, b* sono stati pubblicati nel *Bollettino dell'Accademia Italiana di Stenografia,* XI, 7, 1935.

Il corpus è composto di giornali e riviste con una tiratura superiore a 200.000, diffusi su tutto il territorio nazionale (*Népszabadság, Füles, Ludas Matyi* ecc.). Di ogni giornale è stato analizzato un esemplare intero, ma sfortunatamente non si specifica a quale anno risale il materiale esaminato. Gli autori avevano l'intenzione di includere il maggior numero possibile di riviste destinate per bambini, ciononostante solo 3 (*Dörmögő Dömötör, Kisdobos*, *Pajtás*) dei 14 giornali esaminati sono pubblicazioni per ragazzi. L'intero corpus di 201.286 occorrenze è stato lemmatizzato manualmente, in ogni caso tenendo conto del contesto della parola, e facendo ricorso, nei casi dubbi, a *Magyar értelmező kéziszótár* (Budapest, Akadémiai Kiadó, 1972). Dall'intero corpus sono state estratte 25.852 parole.

La parte introduttiva (disponibile in ungherese, russo e inglese) è seguita da quattro tabelle (pp. XLI-XLIX).

a) La prima contiene i dati comparativi di alcuni dizionari di frequenza pubblicati delle lingue europee, ossia il numero totale delle occorrenze del corpus, il tipo di linguaggio, la frequenza relativa della prima parola e il valore di copertura delle prime 100 parole. Ne emerge che più il campione è variegato, minore è il valore di copertura delle prime cento parole (p. XVI). In Nemes (1941) e nella lista Csirikné e Csirik (1986), dedicati entrambi al linguaggio giornalistico, il valore di copertura delle prime 100 parole è di 39,08% e 37,21% rispettivamente, mentre in Nemes (1933), lista del linguaggio parlamentare, esso arriva a 48,69%.
b) La seconda è una tabella comparativa delle fasce di frequenza: mentre le prime 100 parole coprono il 37,21% dell'intero corpus, con le prime 500 possiamo coprire più della metà delle occorrenze, ossia 54,7%. Dalle prime 500 parole in poi il valore di copertura delle singole fasce diminuisce progressivamente (nella fascia 501-1.000 la copertura è di 8,67%), infatti, arrivando a 3.000 parole il valore di copertura si aggira attorno a 75%.
c) La terza tabella ci informa sulla distribuzione delle frequenze: a ogni numero d'ordine (rango) è associato il valore di frequenza assoluta (ovvero il numero delle occorrenze nel corpus) e quella cumulata (numero totale delle occorrenze delle parole da un determinato rango in su).
d) L'ultima tabella è dedicata alle frequenze delle singole parti del discorso: la classe più frequente è ovviamente quella dei sostantivi (30,89% di tutte le occorrenze sono sostantivi), seguiti a sua volta dai verbi (13,64%), dagli aggettivi (13,49%) e dagli avverbi (9,12%). Considerando invece la distribuzione dei lemmi della lista di frequenza, troviamo che i sostantivi ne costituiscono il 48,13%, gli aggettivi il 20,18%, i verbi il 16,10% e gli avverbi il 2,72%. Ciò vuol dire che nel caso dei sostantivi e aggettivi la grande varietà di lemmi è accompagnata da una minore frequenza d'occorrenza, nel caso dei verbi il numero dei lemmi e quello delle occorrenze tendono a coincidere, mentre sono gli aggettivi che con un esiguo numero di lemmi coprono un largo numero di occorrenze.

La prima parte del volume si conclude con l'appendice delle abbreviazioni e la bibliografia. Nella seconda parte del volume troviamo le liste di frequenza in ordine alfabetico (pp. 3-232) e di frequenza decrescente (pp. 233-312): ogni lemma è accompagnato dalla parte del discorso e dal valore di frequenza assoluta. L'ultima lista contiene le varie parti del discorso in ordine di frequenza decrescente: verbi, sostantivi, aggettivi, numerali, pronomi, avverbi, posposizioni, congiunzioni, interiezioni, articoli,

participi ed altro[5] (pp. 315-399), qui, accanto ad ogni lemma troviamo la frequenza assoluta.

3.2.5 Füredi M., Kelemen J. (1989) *A mai magyar nyelv szépprózai gyakorisági szótára* (1965-1977), Budapest, Akadémiai Kiadó.

Questo lavoro pubblicato nel 1989 è tuttora fondamentale per la descrizione statistica del lessico ungherese, essendo l'unico dizionario di frequenza basato su un largo corpus, compilato in conformità con i risultati più recenti di statistica linguistica. Similmente al lavoro di Csirikné e Csirik (1986) anche questo dizionario è uscito in formato bilingue, in ungherese e inglese, ed è stato realizzato in parte con l'ausilio di calcolatori elettronici. Il corpus include 508.008 occorrenze reperite in 258 opere letterarie di 200 scrittori, pubblicate tra il 1965 e il 1977, suddivise in quattro categorie: opere di prosa; drammi; testi giornalistici e dei mass media; testi scientifici e divulgativi.[6] Di ogni opera è stato selezionato un brano di ca. 2.000 occorrenze. Dalle 91.471 forme diverse riscontrate nel corpus, sono stati estratti 33.196 lemmi, di cui nel volume si è potuto pubblicare solo i 3.410 più frequenti. Un'importante novità dell'opera è di aver introdotto l'indice d'uso (chiamato frequenza modificata), proposto da Juilland e Chang Rodriguez (1964) nel *Frequency dictionary of Spanish Words* (cfr. 2.2.9). Questa formula è considerata un indicatore più affidabile del valore di frequenza assoluta, poiché essa modifica la frequenza assoluta tramite la dispersione, pertanto le forme aventi una distribuzione più omogenea, avranno un valore d'uso superiore rispetto alle forme mal distribuite.
Per il conteggio delle parole è stata considerata la forma grafica: una sequenza di lettere compresa fra due spazi o segni d'interpunzione è considerata una parola. Si è fatto eccezione per i verbi con "prefisso" verbale. Il punto di riferimento per la lemmatizzazione è stato il dizionario enciclopedico della lingua ungherese (*Magyar értelmező kéziszótár,* Budapest, Akadémiai Kiadó, 1972). Nell'elaborazione è stata presa in considerazione l'omonimia, ma non la polisemia. Ogni lemma è quindi identificato oltre che dalla forma grafica, dal codice di omonimia e dal codice di parte del discorso. L'opera contiene varie liste:

a) La prima lista è quella più dettagliata: contiene 3.410 "superlessemi" con i rispettivi lemmi e forme in ordine d'uso decrescente (pp. 1-353). I lemmi le cui forme derivate appartengono a più di una categoria grammaticale e/o hanno più di un significato sono riuniti sotto un c.d. superlessema, stampato in neretto. Sotto i lemmi (stampati in corsivo) sono elencate le rispettive forme (stampate in caratteri normali), ordinate anch'esse secondo il valore d'uso decrescente. Ad es. sotto il superlessema *épül*, troviamo tre volte il lemma *épül*: esso dà luogo a forme verbali *(épült, épül, épülhetett ecc.*), infiniti (*épülnie*) e participi (*épülő, épült*). Per ogni superlessema è specificato il rango, l'eventuale codice di omonimia, il codice di categoria grammaticale, il numero di sottoinsiemi (opere letterarie) in cui il esso è presente, la frequenza assoluta, la dispersione e l'uso (frequenza modificata). Per i lemmi ordinati sotto il superlessema si riportano gli stessi dati, tranne il rango, mentre per le forme si precisano la frequenza assoluta, la dispersione e l'uso.

[5] Nella categoria "altro" rientrano due lemmi: la particella interrogativa *-e* (simile all'uso interrogativo e dubitativo della congiunzione *se* in proposizioni indirette), e la particella *fog*, usato come una specie di ausiliare per esprimere azioni future.
[6] Si nota che rispetto alla datazione dei testi il dizionario è stato pubblicato relativamente tardi, quindi i suoi risultati rischiano di non essere più validi per la lingua ungherese contemporanea.

b) Segue la lista alfabetica dei 3.410 superlessemi dal valore d'uso più elevato (pp. 355-384). I dati elencati sono il rango, l'eventuale codice di omonimia, il codice di categoria grammaticale, la frequenza assoluta e l'uso. Una scelta originale, ma molto utile tra l'altro ai fini dell'insegnamento, è di aver pubblicato la lista delle 4.898 forme aventi una frequenza assoluta da 10 in su (pp. 385-424). Le forme sono numerate progressivamente, per ciascuno si precisa il rango del lemma cui appartiene, il codice di categoria grammaticale del lemma, la frequenza assoluta e il valore d'uso.
c) In appendice sono riportate altre liste (pp. 425-436): i codici degli omonimi; la lista degli omonimi nel cui caso il codice di categoria grammaticale non dà informazioni sufficienti sul significato; i codici delle parti del discorso (poiché la classificazione è più dettagliata di quella che generalmente troviamo nei dizionari); le frequenze dei grafemi nel corpus; la distribuzione delle vocali nei testi; la distribuzione delle parti del discorso; la distribuzione delle forme verbali finite; la distribuzione dei modi verbali, distribuzione delle forme verbali definite e indefinite (aventi un oggetto e non); la distribuzione dei superlessemi secondo il numero di brani esaminati (numero di superlessemi nuovi per ogni nuova opera analizzata).

Tenendo presente che una parte dell'enorme lavoro è stato effettuato manualmente, è lodevole la varietà dei dati, come anche i diversi criteri e il formato scelti per la loro presentazione. Questa lista, insieme al dizionario inverso di Ferenc Papp (1969), è alla base del thesaurus di Debrecen, raggiungibile on line (http://romanid.freeweb.hu/DT2/dt2.htm), tuttavia questa versione non fornisce dati di frequenza.[7]

3.2.6 Aradi A., Erdős J., Sturcz Z. (2000) *Küszöbszint, Magyar mint idegen nyelv.* Strasbourg, Európa Tanács.

Il *Livello soglia* per l'ungherese è il ventesimo nella serie dei vari livelli soglia, pubblicati per le lingue europee dal Consiglio d'Europa nel periodo 1975-2005. Nella prefazione J. L. M. Trim accentua l'importanza di quest'iniziativa di portata europea, che contribuisce a reintegrare l'Ungheria nella famiglia dei popoli europei dopo le vicissitudini degli ultimi cinquanta anni (p. 7). Nel passato recente si è creata una diaspora ungherese notevole in tutto il mondo, oltre che nei Paesi confinanti con l'Ungheria, perciò il *Livello soglia* per l'ungherese vede tra i suoi destinatari non solo gli apprendenti stranieri, ma anche i discendenti di famiglie ungheresi residenti all'estero.
L'opera ha due parti: la prima è dedicata alla descrizione funzionale del *Livello soglia*, mentre la seconda contiene il sistema delle forme linguistiche. La prima parte è stata elaborata in base al lavoro di Van Ek e Trim (1991), il *Threshold Level 1990*. La descrizione del sistema grammaticale, invece, è più approfondita rispetto al modello originale, per due motivi. Primo perché la ricca morfologia della lingua ungherese richiede una trattazione più dettagliata, secondo perché la descrizione dell'ungherese come lingua straniera non ha una tradizione paragonabile a quella delle altre lingue europee, pertanto gli autori desiderano informare l'utente di una serie di questioni grammaticali alle quali difficilmente potrà trovare una risposta nei libri didattici esistenti (p. 11).
Il libro è dedicato a turisti, studenti, lavoratori e commercianti che vogliono passare periodi più o meno lunghi in Ungheria. L'apprendimento del materiale linguistico del

[7] Sulla costituzione del thesaurus si veda Papp F. (2006: 254-285).

volume richiede ca. 360 ore di studio da parte di un discente con madrelingua europea, non residente in Ungheria (p. 13).
Punto di partenza della descrizione dettagliata è quindi la lista delle funzioni linguistiche (atti), che il discente deve essere in grado di compiere.[8] Segue poi l'inventario delle nozioni generali (concetti che compaiono in quasi tutte le situazioni) che corrisponde, per gran parte, all'inventario del *Threshold Level* (e quindi del *Livello soglia*). Esse sono suddivise nei seguenti sottogruppi: esistenza, spazio, tempo, quantità, qualità, nozioni mentali, rapporti, riferimenti deittici (pp. 67-91). Similmente le nozioni specifiche (concetti concreti da usare solo in determinate situazioni) seguono pressappoco le stesse categorie del *Threshold Level*: individuo e famiglia; casa e ambiente; vita quotidiana; tempo libero e divertimento; viaggi; rapporti con gli altri; salute e benessere; istruzione; compere; mangiare e bere; servizi; luoghi; lingua straniera; tempo (pp. 93-136). È interessante che manca la categoria "opinioni politiche" che invece è presente nel *Livello soglia* per l'italiano. Per ogni nozione generale e specifica si precisano le concrete forme linguistiche, indicando le voci che vanno apprese solo in modo passivo, ricettivo.
Le critiche rivolte ai precedenti lavori non sono state prive di conseguenze.[9] Si è reso conto che nelle reali situazioni comunicative le singole funzioni raramente si realizzano in entità discrete, perciò gli autori del *Livello soglia* ungherese hanno ritenuto opportuno inserire un capitolo dedicato ai modelli tipici di conversazione. Si è cercato di abbozzare le possibili combinazioni degli atti comunicativi in determinate situazioni (pp. 137-141).
Similmente al *Threshold Level 1990*, un capitolo è riservato alla comprensione dei testi prodotti dai mass media. Gli autori precisano che l'apprendente che abbia raggiunto questo livello, facendo uso delle strategie di compensazione ed eventualmente di un dizionario, deve essere in grado di capire l'essenza di qualsiasi tipo di testo, senza restrizioni relative al lessico. Le uniche restrizioni riguardano l'organizzazione del testo: esso deve avere una struttura chiara con delle informazioni esplicite (pp. 143-145).
Contrariamente a quanto visto nel caso del *Livello soglia*, anche le capacità di scrittura hanno un certo rilievo. Nonostante il *Livello soglia* sia una capacità di tipo prevalentemente orale, il discente deve essere in grado di completare moduli, scrivere alcuni tipi di lettere formali e lettere private (pp. 147-148).
Un intero capitolo è dedicato alle competenze socioculturali, poiché le funzioni linguistiche dipendono in certa misura dalle convenzioni sociali presenti in una determinata comunità linguistica. Si parla di esperienze universali, convenzioni sociali, regole di cortesia (pp. 149-158).
Consapevoli che nella realtà le conoscenze dell'apprendente non corrispondono ai requisiti posti da molte situazioni comunicative, gli autori richiamano l'attenzione sulle strategie di compensazione. Nelle situazioni più "difficili" dal punto di vista linguistico, il successo della comunicazione dipende dalla misura in cui il discente è in grado di attivare le proprie strategie di compensazione (pp. 159-162).
Il capitolo dedicato ai metodi dell'apprendimento non rientra strettamente nella sfera degli obiettivi previsti dal *Livello soglia*, tuttavia si è ritenuto opportuno richiamare l'attenzione su alcuni aspetti del processo di apprendimento. Il *Livello soglia* descrive le conoscenze linguistiche che corrispondono ai bisogni di un apprendente ideale, tuttavia ogni discente deve essere consapevole dei propri bisogni, selezionando tra i metodi e contenuti proposti quelli più adeguati (pp. 163-166). Nel capitolo dedicato al livello di

[8] Le funzioni linguistiche principali sono: richiedere e dare informazioni, esprimere e riconoscere attitudini, convincere, iniziare un contatto, tecniche della comunicazione, correzioni comunicative (pp. 35-66).
[9] Secondo le critiche avanzate da Ciliberti (1994: 104) "la comunicazione è un processo che si realizza in testi e non in insiemi di enunciati isolati. Le funzioni identificate dai sillabi comunicativi sono entità statiche, discrete e a-contestuate, sono in pratica delle idealizzazioni."

competenza sono descritti i requisiti qualitativi della comunicazione efficiente (pp. 167-169).
In appendice alla prima parte troviamo la lista alfabetica delle unità lessicali (pp. 171-196). Purtroppo non si precisa il carico degli esponenti, che secondo le nostre stime si aggirano attorno a 2.200 unità. Segue un elenco delle liste (aperte) delle nozioni concrete (p. 197) e un indice alfabetico delle funzioni linguistiche e delle nozioni generali e specifiche (pp. 198-205).
La seconda parte dell'opera è una vera e propria grammatica descrittiva della lingua ungherese. Nei singoli capitoli sono trattate le caratteristiche dei fonemi e lettere, le parole, le espressioni polirematiche, la frase semplice e complessa, l'intonazione (pp. 207-377). In appendice alla seconda parte è riportata una tabella riassuntiva delle coniugazioni.
Questo lavoro testimonia che, nonostante diverse critiche, è possibile applicare il modello del *Livello soglia* a vari tipi di lingue (come il greco, russo, lettone, lituano, celtico, arabo, maltese), perfino a quelle non indoeuropee (l'ungherese è stata la seconda dopo l'estone in questa serie). L'approccio di fondo, per cui si rinvia alla sez. 5.1.1, si è rivelato utilizzabile, poiché il concetto nozionale-funzionale che n'è alla base, non parte dalla struttura formale delle lingue, ma si concentra sugli atti comunicativi che l'utente deve compiere nelle varie situazioni quotidiane.

3.2.7 Csirikné Czachesz E., Csirik J. (2002) *10-16 éves tanulók írásbeli szókincsének gyakorisági szótára*, Budapest, BIP.

In questo dizionario di frequenza del lessico di scrittura gli autori analizzano 2.170 composizioni scritte da ragazzi dai 10 ai 16 anni, selezionate casualmente tra 8.670 lavori. Il corpus completo conta 600.000 occorrenze, ed è raggiungibile anche on line, all'interno della database Szeged Korpusz (http://www.inf.u-szeged.hu/projectdirs/hlt/). L'opera è nata come prodotto di 'scarto' di due progetti: nel 1998 presso l'Università di Szeged si è condotto un'indagine nazionale sulle competenze espressive di studenti dai 10 ai 16 anni; e una ricerca di linguistica computazionale del Gruppo di Ricerca sull'Intelligenza Artificiale dell'Università di Szeged (SZTE), dell'Accademia Ungherese delle Scienze (MTA) e della società MorphoLogic Kft. sull'analisi morfologica automatica dei testi.
Il dizionario fornisce varie liste: la prima (pp. 19-103) contiene tutti i lemmi del corpus con le loro frequenze assolute, complessivamente 20.300 parole diverse. La seconda è una lista alfabetica delle 1.000 voci più frequenti, accompagnati dall'indicazione di categoria grammaticale e frequenza assoluta (pp. 104-116). Seguono le liste di frequenza secondo categoria grammaticale (pp. 117-153): 602 aggettivi, 106 congiunzioni, 88 interiezioni, 169 numerali, 599 sostantivi, 144 pronomi, 622 avverbi, 84 postposizioni, 3 articoli, 589 verbi, 14 abbreviazioni. Per ogni lemma si precisa la frequenza assoluta ma non il rango. Per le categorie più numerose (aggettivi, sostantivi, avverbi, verbi) si riportano solo le 600 voci più frequenti, mentre per le altre classi sono riportate tutte le voci trovate (congiunzioni, interiezioni, numerali, pronomi, posposizioni, articoli, abbreviazioni).
Il libro è un ottimo ausilio per analizzare il lessico in una certa fascia d'età, poiché nella seconda parte del volume le parole più frequenti sono ripartite anche secondo le classi scolastiche (pp. 148-248). Per ogni anno scolastico (quarta, sesta, ottava e decima classe) oltre la lista alfabetica dei primi 500 lemmi si trova una suddivisione delle voci per categoria grammaticale. Accanto a ciascun lemma figura il codice della categoria

grammaticale e il valore di frequenza assoluta (senza indicazione di rango). Il numero delle occorrenze e delle forme diverse è riportata in apposite tabelle (pp. 251-253). L'intero corpus di 601.135 parole registra 20.300 forme diverse. Il numero di forme diverse registrate nelle singole fasce d'età aumenta progressivamente: mentre gli alunni della quarta classe si avvalgono di sole 7.534 forme, quelli della sesta classe di 8.798 forme, quelli dell'ottava classe usano 10.530 forme e i ragazzi della decima classe hanno usato 12.179 forme diverse. I ranghi con le rispettive frequenze assolute si trovano alla fine del volume (pp. 254-259).
Si tratta della statistica lessicale più recente che grazie alla sua completezza e alle soluzioni adottate, si presta come strumento efficace per l'analisi del linguaggio dei ragazzi.

3.2.8 Bácsi J., Kerekes J. (2003) Az első osztályos olvasókönyvek szóanyagából készült gyakorisági szótár: "Van szó". In: *Módszertani közlemények*, 43, 2, pp. 53-58.

Questo dizionario è nato come risultato secondario dei lavori del software *Beszédmester*, volto a migliorare le abilità comunicative di bambini sordi e con difetti di pronuncia. Contiene le 2.000 parole più frequenti dei 13 libri lettura, utilizzati nel primo anno della scuola elementare, scelti dalla lista proposta dal Ministero dell'Istruzione (p. 54). L'intero lessico dei libri è stato registrato in forma elettronica, successivamente si sono compilate tre liste (p. 56).

a) La prima elenca in ordine di frequenza decrescente tutte le 27.293 forme (sequenze di grafemi) contenute nei libri di lettura. Sono state incluse anche sequenze che pur non facendo parte del lessico della lingua ungherese, sono indispensabili per l'insegnamento della lettura (*am, ib, pa* ecc.).
b) Per la seconda lista gli autori hanno isolato 12.226 forme, escludendo le parole grammaticali come prefissi verbali, congiunzioni, interiezioni, ausiliari, articoli e le sequenze prive di significato. Tutte le forme sono state ricondotte ad un lessema, e riportate in ordine di frequenza decrescente.
c) Per la terza lista sono state scelte le 1.953 parole con una frequenza non inferiore a 10. Queste costituiscono il database del software, che riconosce, analizza e visualizza i fonemi pronunciati dai bambini.

Giacché l'obiettivo del lavoro non è stato quello di compilare una lista di frequenza, sono stati pubblicati solo i dieci verbi, nomi propri, nomi di animali e di piante più frequenti. Nella lista accanto ad ogni parola troviamo il suo rango nella lista di frequenza decrescente, la frequenza assoluta e relativa. Il titolo dello studio mostra appunto che le due parole più frequenti sono il verbo *van* (c'è) e il sostantivo *szó* (parola).
Il software *Beszédmester*, realizzato con il sostegno finanziario del Ministero dell'Istruzione in collaborazione del Dipartimento d'Informatica dell'Università degli Studi di Szeged, dalla Facoltà di Magistero Juhász Gyula e dalla Scuola dei Non Udenti di Kaposvár ha richiesto la collaborazione di informatici, matematici, pedagoghi. Il programma è in vendita dal 2006, e i primi riscontri sono abbastanza promettenti. Gli autori prevedono inoltre, di pubblicare in Internet un dizionario di frequenza vero e proprio, che "corrisponda a tutti i requisiti della lessicografia" (p. 58).

3.2.9 *Dizionario di frequenza WordSword* (2004)

Uno dei più grandi lavori ungheresi di statistica lessicale è il progetto *WordSword* (Szószablya) del Centro di Formazione e di Ricerca dei Media (Média Oktató és Kutató Központ) dell'Università Tecnica di Budapest. Il progetto è stato avviato nel marzo del 2003, e attualmente è consultabile sul sito: http://mokk.bme.hu. Il lavoro comprende lo sviluppo di un programma di controllo ortografico e d'analisi morfologica e la creazione del dizionario di frequenza *WordSword*. La base del dizionario è il più vasto corpus elettronico dell'ungherese: la raccolta è considerata rappresentativa della lingua contemporanea scritta, poiché si basa su 3,5 milioni di siti web, comprensive di 1.486 milioni di occorrenze, scaricati con l'aiuto dei motori di ricerca Origo e Vizsla tra dicembre 2002 e l'autunno del 2003.[10]

Innanzi tutto si sono eliminati i testi ricorrenti più volte e i file privi di testi utilizzabili. Successivamente i testi sono stati sottoposti ad un controllo linguistico per mezzo del programma di controllo ortografico *HunSpell*. La qualità del materiale raccolto presentava oscillazioni notevoli, perciò con il filtraggio automatico dei siti sono stati separati 4 corpus diversi:

a) L'intero corpus di 3,5 milioni di pagine si può considerare una raccolta rappresentativa della lingua dei siti ungheresi: essa comprende 1.486 milioni di parole e 19,1 milioni di forme.
b) Con una soglia massima di 40% di parole non riconosciute dal programma *HunSpell*, sono stati scartati i documenti scritti in una lingua diversa dall'ungherese, così sono rimaste 3,125 milioni di pagine, 1.310 milioni di occorrenze e 15,4 milioni di forme.
c) Adoperando una soglia di 8% di parole sconosciute scompaiono anche i siti senza accento grafico, ma rimangono i testi scritti in un linguaggio settoriale e in gergo: sono rimaste 1.918 milioni di pagine con 928 milioni di occorrenze e 10,9 milioni di forme.
d) Con la soglia minima del 4% sono state accettate solo pagine contenenti meno errori ortografici di un qualsiasi testo pubblicato, ottenendo 1,221 milioni di pagine con 589 milioni di occorrenze e 7,2 milioni di forme. Quest'ultimo corpus sarebbe rappresentativo della lingua ungherese standard.

La lemmatizzazione delle forme è stata effettuata tramite *HunStem*, un programma che produce anche informazione morfologica. Tuttavia i risultati ottenuti sono approssimativi, perché allo stato attuale non sono state risolte le ambiguità di lemmatizzazione e d'omonimia.

Il dizionario tratto dal corpus d), contenente ca. 600 milioni di occorrenze, può essere scaricato dal web (http://mokk.bme.hu/resources/webcorpus/index_html). Nella lista di frequenza ogni lemma è preceduto dal rango, seguono poi i dati di frequenza: i quattro numeri indicano la frequenza per l'intero corpus (a) e quella per gli altri tre corpus (b, c, d) costituiti con la soglia del 40, 8 e 4% rispettivamente.

Sono state messe a disposizione due liste: una è ordinata alfabeticamente, l'altra contiene i lemmi in ordine di frequenza decrescente. In più si possono scaricare le 100.000 parole più frequenti (con o senza i dati di frequenza), che dovrebbero coprire il 95% dei testi.

Ad un esame più attento dei dati si rileva che finora non si è riusciti ad eliminare completamente le parole inglesi dal corpus integrale, e nemmeno la lemmatizzazione delle forme è risolta. Nonostante queste lacune, la lista WordSword è unica sia per la quantità di dati elaborati, sia perché è liberamente accessibile.

[10] Emerge la domanda se il web può essere considerato un corpus di riferimento rappresentativo. Chiari (2007: 54-57) argomenta che grazie all'alto tasso di errori e pagine amatoriali il web dà una rappresentazione molto sbilanciata delle caratteristiche di una lingua, pertanto ci vuole molta cautela nell'interpretazione dei risultati.

3.2 Tabella riassuntiva e comparativa

Possiamo sintetizzare i risultati del capitolo precedente nella tabella seguente.

	N. dei lemmi	Defini-zioni, esempi	Parole gramm.	Ampiezza del corpus	Età del corpus	Criteri di selezione			
						Esperienza	Frequenza	Dispersione	Uso
Aradi et al. (2000)	2.200	esempi	+			+			
Bácsi, Kerekes (2003)	1.953		-	12.226	2001-2002 ca.		≥ 10		
Cser (1939)	3.767		-	89.705	1938 ca.		≥ 2		
Czachesz, Csirik (1986)	25.852		+	201.286	1980-1985 ca.		≥ 5		
Czachesz, Csirik (2002)	20.300		+	600.000	1998		+		
Füredi, Kelemen (1989)	3.410		+	508.008	1965-1977		≥ 10	+	+
Nemes (1933)	505		+	40.977	1930 ca.		≥ 25		
Nemes (1941)	2.969 (7.650)		+	453.578	1933-1935		≥ 25 (≥ 5)		
WordSword (2004)	100.000		+	589.000.000	2002-2003		+		

3.3 Riassunto, prospettive

In Ungheria molti dizionari di frequenza sono stati estratti dalle concordanze di opere letterarie, mentre poche sono le statistiche basate su un corpus rappresentativo della lingua standard. Il dizionario di frequenza per eccellenza è stato considerato per lungo tempo quello di Füredi e Kelemen (1989), basato su un corpus di 500.000 occorrenze, tratte da opere letterarie. Si tratta del primo dizionario di frequenza moderno della lingua ungherese, perché esso ha utilizzato per la prima volta i risultati della statistica linguistica (introduzione dell'indice di dispersione e d'uso, selezione di un corpus rappresentativo, elaborazione automatizzata). È da notare però, che siccome i testi elaborati risalgono a più di un trentennio fa, i risultati della ricerca sono poco rappresentativi della lingua ungherese del XXI secolo. La realizzazione più recente di statistica linguistica, la lista *WordSword* purtroppo non costituisce un'alternativa per i ricercatori, giacché essa non è un dizionario di frequenza vero e proprio, ma una lista di forme grafiche (tra cui anche parecchie parole inglesi) non lemmatizzate, provviste di dati di frequenza. Sembra che il lavoro si trovi alle fasi iniziali, pertanto nello stato attuale difficilmente si presta come strumento valido per le ricerche linguistiche.
L'unica raccolta può o meno rappresentativa della fascia fondamentale del lessico ungherese è il lavoro di Aradi *et al.* (2000). Pare tuttavia, che poca attenzione, da parte dei linguisti ed insegnanti, è stata rivolta alla versione ungherese del *Livello soglia*. L'unica recensione di Kövérné Nagyházi è uscita quattro anni dopo la pubblicazione del volume, nel 2004, e non pare che esistano libri didattici dell'ungherese L2, basati sulla proposta di Aradi e i suoi collaboratori. In assenza di altre opere, sarà questa lista che prenderemo come punto di riferimento per integrare il lessico fondamentale di italiano L2, destinato ad apprendenti ungheresi.
Come nel caso della lingua italiana, anche per l'ungherese ci sono sempre più iniziative volte a costituire larghi corpora di riferimento che possono servire come base per la redazione di liste di frequenza. Il più illustre è il Corpus nazionale ungherese (*Magyar Nemzeti Szövegtár*), raggiungibile sul sito http://corpus.nytud.hu/mnsz/, contenente ca.

150 milioni di parole, raccolte dall'Istituto di linguistica dell'Accademia ungherese delle scienze tra il 1997 e il 2002. Il corpus rappresenta la lingua ungherese contemporanea scritta, ed è detto nazionale, poiché i materiali derivano da cinque grandi aree geografiche in cui si parla l'ungherese: l'Ungheria, la Transilvania, l'ex Ungheria meridionale (oggi appartenente alla Croazia e alla Serbia), l'ex-Ungheria settentrionale (oggi parte della Slovacchia) e la Rutenia subcarpatica (l'Ucraina occidentale). Sono state prese in esame 75 milioni di occorrenze di testi giornalistici, 15 milioni di parole di opere letterarie (tratte dall'archivio dell'Accademia digitale di letteratura) e 20 milioni di parole estratte dal database della Biblioteca elettronica ungherese. Il linguaggio amministrativo è rappresentato da materiali ministeriali e municipali, mentre per lo stile informale sono stati esaminati i testi di vari forum on-line del sito www.index.hu (20 milioni di occorrenze per ciascuno dei due sotto-insiemi). Il corpus è accessibile a tutti dopo una registrazione. La raccolta, pur non contenendo dati di frequenza, informa sul numero delle occorrenze della parola nel corpus, e fornisce tutti i contesti in cui la parola appare, e le loro fonti, potrà quindi servire come base per la creazione di nuove liste di frequenza. Sui dati ricavati dal corpus si basano i dati di frequenza del Dizionario enciclopedico ungherese (*Magyar Értelmező Kéziszótár*, 2003).
L'altra raccolta importante è il Corpus storico della lingua ungherese (*Magyar Történeti Szövegtár*, disponibile sul sito www.nytud.hu/hhc/). Il corpus, costruito tra il 1985 e il 2005, raccoglie testi letterari, scientifici, divulgativi e giornalistici dal periodo 1772-2000 e registra 27 milioni d'occorrenze. Il Grande dizionario della lingua ungherese (*A magyar nyelv nagyszótára*), di cui il primo volume e il supplemento sono stati pubblicati nel 2006 (con notevole ritardo rispetto alla data prevista) conterrà ca. 110.000 lemmi, che saranno ricavati, fra l'altro, dal suddetto corpus.
Concludendo, la costituzione di vasti corpora permetterà nel futuro la compilazione di diversi dizionari di frequenza, dedicati ai linguaggi settoriali e non. L'incremento del numero delle ricerche è più che promettente, pertanto si prevede che nel futuro saranno pubblicati nuovi dizionari di frequenza, tra cui alcuni probabilmente si fonderanno sul Corpus nazionale ungherese.

4 Alcune applicazioni dei lessici di frequenza e di base

I lessici di frequenza e di base, e in generale i dati di frequenza ricavati da larghi corpora, sono fonti importanti per varie discipline. Fin dall'inizio le applicazioni hanno interessato tutti i campi disciplinari: dagli studi di stenografia, crittografia, stilometria, filologia, linguistica teorica e tecnologia delle lingue, attraverso la terminologia, traduzione computerizzata dei testi, ricerche di psicolinguistica e sociologia, fino alle scienze della comunicazione e al marketing, per non parlare della didattica delle lingue. Genericamente si può affermare che i dizionari di frequenza possono essere usati come strumenti linguistici di pieno diritto in ogni caso, in cui un gruppo ristretto di parlanti non può esprimere un giudizio sicuro, ed è quindi meglio fare affidamento su dati offerti da un corpus rappresentativo della lingua che si vuole descrivere.
In questa breve rassegna non possiamo che tracciare in grandi linee alcune delle applicazioni lessicografiche, ci soffermiamo sull'apporto che i lessici di frequenza e di base possono dare alla scrittura di testi di facile lettura, infine tratteremo qualche applicazione informatica che fa uso di dati di statistica lessicale.

4.1 Lessicografia

> "Il carattere e l'estensione delle definizioni, la distinzione dei vari lemmi e significati, la scelta e la frequenza degli autori citati, la misura e il taglio delle esemplificazioni, l'inclusione di voci d'uso rarissimo o di tipo troppo arcaico, l'accoglimento di voci tecniche e scientifiche, la dosatura fra antico e moderno, l'ambito e i limiti dell'indicazione etimologica sono problemi assai delicati che non sempre è agevole mantenere ancorati ad una soluzione canonica, giacché ogni volta aspirano a conformarsi secondo la personale vicenda delle singole parole."

Così ragiona Battaglia nella Prefazione al *Grande dizionario della lingua italiana* (Torino, UTET, 1961-2002), agli inizi degli anni Sessanta. Dalla pubblicazione di quest'opera, singolare nel suo genere, sono passati quasi cinquant'anni, durante i quali la lessicografia, nonché i settori connessi, come la linguistica dei corpora e la lessicologia, hanno compiuto sviluppi notevoli. I dizionari (elettronici) avanzati possono contenere una gran quantità di informazioni linguistiche in più rispetto a un dizionario tradizionale: oltre alla pronuncia e la categoria grammaticale, possono includere indicazioni sulle frequenze d'uso, informazioni lessicali, morfo-sintattiche e semantiche, esempi e collocazioni tratti da corpora, ecc. D'altro lato, nei dizionari moderni la selezione dei dati inclusi è frutto di una metodologia più coerente, formulata in armonia con criteri prestabiliti. Di seguito si analizzeranno alcuni dizionari d'uso, per verificare in quale misura ricorrono a dati statistici ricavati da corpora di riferimento. Vedremo in particolare secondo quali criteri avviene la selezione e la definizione dei lemmi, l'eventuale inclusione dei dati di frequenza e la scelta delle marche d'uso, l'ordinamento delle accezioni, la registrazione delle polirematiche e delle collocazioni in alcuni dizionari d'uso.

4.1.1 La selezione dei lemmi

Nella selezione dei lemmi i vocabolari d'uso solitamente puntano alla lemmatizzazione più completa possibile delle parole della lingua. Tra gli argomenti di vendita dei vocabolari, infatti, occupa uno spazio costante il riferimento al numero delle parole ed accezioni contenute nell'opera, nonché la loro utilità e indispensabilità per chiarire tutti i possibili dubbi linguistici di ogni possibile utente: il Devoto-Oli 2008 dichiara, ad esempio, di essere "un vocabolario di tutti e per tutti"; ecc. A questo proposito emergono due problemi: da un lato è comunemente noto che il solo numero dei lemmi non è segno dell'alta qualità del dizionario in tutti i suoi aspetti; d'altro lato, più universale è il dizionario, più cresce la quantità delle informazioni offerte, e quindi il pericolo che non dia una spiegazione (esatta) alle concrete domande degli utenti.[1] Siamo concordi con De Mauro e Ferreri (2005: 299) nel ritenere che

> "occorrerebbe ormai dare per acquisito che l'obiettivo di un dizionario non può essere quello di adeguare la effettiva quantità di lessemi utilizzati nei testi e discorsi."

Queste considerazioni tuttavia stentano a farsi strada tra editori, autori di pubblicità di dizionari e giornalisti che, nel tentativo di offrire al pubblico un dizionario con qualche lemma in più rispetto agli altri concorrenti, spesso gonfiano le cifre dei lemmi nei loro dizionari. Molte volte anche il titolo delle opere lessicografiche è fuorviante per quanto riguarda la quantità di lemmi: numeri dichiarati e termini spesso equivoci costituiscono un problema nella determinazione dell'estensione dei lemmari (cfr. anche sez. 1.2.3). La questione si pone non solo nel confronto tra lingue diverse, ma anche all'interno di una stessa lingua e spesso persino di una stessa casa editrice.[2]

Fonte	**Numeri dichiarati**
Devoto-Oli 2008	100.000 parole d'autore
Dizionario della lingua italiana per il terzo millennio (De Mauro 2000)	140.000 lemmi + 30.000 polirematiche
Dizionario Italiano Sabatini Colletti (DISC, 1997)	138.000 voci (108.000 nel CD)
Grande dizionario della lingua italiana (Battaglia, 21. voll., 1961-2002)	185.000 lemmi
Grande Dizionario Italiano dell'Uso (GRADIT, voll. 6 + 1, 1999)	251.000 lemmi + 130.000 polirematiche
Grande dizionario italiano Garzanti 2008	250.000 lemmi ed accezioni
Zingarelli 2008	140.000 voci + 375.000 significati

Numero di lemmi contenuti in alcuni dizionari d'uso.

Puntando all'universalità, i vocabolari d'uso indicano un numero spropositato di lemmi in essi elencati: dal GRADIT (*Grande Dizionario Italiano dell'Uso*) con i suoi 251.000 lemmi, cui devono aggiungersi le 130.000 espressioni polirematiche, si scende pian piano fino ai 100.000 registrati dal Devoto-Oli 2007. La tabella mostra che i dizionari non adottano terminologie comparabili. Quel che è rappresentato è chiamato *lemma, voce, parola d'autore* e anche *accezione, polirematica, significato*. Notiamo che sarebbe opportuno fissare il termine di riferimento, giacché:

> "il carico memoriale è ben diversamente gravato se deve tendere al solo significato principale, al significato con il suo carico di accezioni, ai lemmi intesi come unità singola o al lemma con tutti i suoi sotto-lemmi, a parole singole o a raggruppamento di più membri con un significato unitario" (De Mauro, Ferreri 2005: 303).

[1] Rubando la metafora a De Mauro (2005: 32), come una carta geografica, anche il dizionario risulta utile se da un lato è di dimensioni ridotte rispetto a ciò che rappresenta, d'altro lato sono chiare le scelte e le regole seguite nella selezione e nella rappresentazione dei fatti lessicali.

[2] Sulla confusione terminologica, creatasi sul mercato dei dizionari monolingui e bilingui della lingua tedesca, si veda Földes (2004). Lo studio conferma che le denominazioni utilizzate nei titoli delle opere, spesso non sono indicatori affidabili della quantità (e della qualità) del materiale incluso.

È vero, infatti, che a volte il numero delle accezioni può essere sensibilmente superiore a quello dei lemmi, pertanto quest'ultimo sembra essere un indicatore più valido della mole dell'opera. Ad esempio nello Zingarelli non è chiaro se l'universo di riferimento sia rappresentato dalle 140.000 voci o dai 375.000 significati. Ancora più confondente è l'indicazine data dal Garzanti: non è chiaro, infatti, quale informazione vuole fornire la somma dei lemmi e delle accezioni. In più, solitamente non è precisato se sotto la stessa etichettatura di parola o lemma si computano sia singole unità lessicali, sia agglomerati di più di un vocabolo. L'unica opera lessicografica ad aver immesso a lemma tutte le espressioni lessicali complesse (polirematiche) è il GRADIT e un altro dizionario ad esso correlato (De Mauro 2000) (cfr. anche sez. 4.1.4). Possiamo concludere che l'uso equivoco delle designazioni limita le possibilità di comparazione.
Tornando alla selezione dei lemmi, consultando la prefazione d'alcuni dizionari d'uso, si rileva che comunemente non è esplicitato il criterio adottato per individuare le parole da includere (così nel Battaglia, nel DISC e nel Devoto-Oli). È ovvio, ed è stato additato in varie sedi, che una lemmatizzazione completa di tutte le voci, accezioni e locuzioni della lingua di cultura è infattibile per motivi pratici e teorici. Nel passato i lessicografi hanno cercato di affrontare questo problema ricorrendo nella selezione a lavori precedenti, integrandoli con altri dizionari e liste di parole, omettendo le voci obsolete o troppo rare. Nel caso dei dizionari le cui fonti sono esclusivamente altri vocabolari, inchieste condotte dai compilatori o schede predisposte, questo procedimento è l'unico possibile, e il suo valore lessicografico dipende dalla chiarezza con cui sono stati definiti i singoli passi della selezione. A titolo d'esempio, la base di tutte le versioni del Devoto-Oli è stata la primissima edizione del 1970, che nelle successive edizioni è stata aggiornata ed ampliata (l'ultima versione è l'edizione 2008 a cura di Serianni e Trifone). Come nel dizionario originario non sono stati precisati i principi della scelta dei lemmi, così neanche nelle successive edizioni è chiarito, con quali criteri sono stati selezionati i lemmi nuovi ed esclusi quelli obsoleti: la pubblicità parla di "lemmi provenienti dai giornali e dalla televisione." Nella prefazione dell'edizione 2004-2005 (p. V) si tiene tuttavia a precisare, che:

> "per la prima volta gli alterati sono stati selezionati non sulla base della competenza del lessicografo (o sul modello di altri dizionari precedenti), ma verificando attraverso testi reali, disponibili in rete, quali fossero effettivamente in circolazione, magari in accezioni particolari."

Del resto, però, non si specificano le fonti utilizzate per integrare e rivedere le informazioni grammaticali (reggenze, fraseologia), il lessico scientifico e le etimologie, menzionati dalla pubblicità.[3]
Il Garzanti 2004-2005 dichiara di aver adottato, nella selezione, un "criterio funzionale": sono state incluse tutte le voci del lessico comune; le voci antiche che "hanno avuto un effettivo rilievo nella tradizione storico-culturale"; termini tecnici della medicina, biologia, fisica, ecologia, informatica, diritto ecc.; voci regionali, dialettali e gergali che "hanno consistente diffusione nella lingua comune"; voci e locuzioni d'altre lingue, entrate in italiano in tempi recenti; locuzioni latine tuttora "vive e significative nell'uso attuale" (p. V). Ciononostante non è esplicitato se esisteva un database, cui questi criteri "funzionali" sono stati adoperati.
Un interessante e ben documentato procedimento di selezione si riscontra nel GRADIT. Il database iniziale è stato costituito da una serie di dizionari, tra cui il Battaglia, lo Zingarelli e il Devoto-Oli ed altri (cfr. De Mauro 2005: 43). Dalla base di dati è passato nel lemmario del GRADIT ogni vocabolo che sia presente in tutte le opere del database,

[3] Anche molto vaga sembra la dichiarazione degli autori secondo cui "ciascuna [voce] è registrata nella forma più comune e moderna" (p. V).

in più, ogni vocabolo del Battaglia, a condizione che non fosse obsoleto, e quindi presente solo nel Battaglia. L'iniziale base di dati è stata integrata con una serie di dizionari a destinazione commerciale e/o scolastico e dizionari di neologismi, apparsi dopo il 1989 (per l'elenco completo cfr. De Mauro 2005: 45-46). A tutto ciò si sono aggiunte decine di migliaia di schede lessicografiche, predisposte da Edoardo Sanguineti, tratte dalla lettura di testi italiani d'ogni secolo e genere. Sono state tenute fuori dal lemmario tutte le parole occasionali, parole novecentesche d'uso troppo raro e vocaboli presenti solo in scritti anteriori alla metà dell'Ottocento. Si tratta, quindi, di un dizionario composto con criteri scientifici rigorosi, puntualmente documentati in De Mauro (2005). Di questi procedimenti coerenti giova anche il più ridotto *Dizionario della lingua italiana per il terzo millennio* (De Mauro 2000), composto sulla base della base di dati del GRADIT.

In relazione al problema della selezione dei lemmi, emergono in realtà due questioni fondamentali: quanti e soprattutto quali lemmi dovrebbero essere inclusi. È ovvio che con principi come fornire "tutte le voci dell'uso corrente" (*Dizionario medio italiano 2007*, Milano, Garzanti) o compilare "un vocabolario di tutti e per tutti" (*Devoto-Oli* 2008) non si possono prendere decisioni chiare. L'unico caso in cui è possibile garantire una specie di selezione 'ideale' è quando si ricorrono a corpora chiusi. Questo procedimento finora era adottato in prevalenza ai dizionari storici e alle concordanze letterarie, ma nel tentativo di privilegiare gli usi più frequentemente testimoniati nella lingua contemporanea, la scelta di includere o escludere determinati lemmi in molti moderni dizionari è frutto di analisi sulle frequenze delle forme nei corpora. Notiamo che una selezione fondata su dati di frequenza permette di operare scelte coerenti anche riguardo all'inclusione dei nomi propri, sigle ed abbreviazioni, nomi geografici, forme flesse irregolari, termini tecnici e dialettali, polirematiche, derivati, neologismi, parole obsolete, marchionimi ecc., la cui selezione risente, nella maggioranza dei casi, delle preferenze individuali dei compilatori. I dati di frequenza andrebbero poi integrate in modo da disporre di liste complete nel caso delle serie chiuse (colori, numeri ecc.) e ogni volta quando lo scopo preciso, che il dizionario si è prefisso, lo richiede.[4]

Il primo dizionario ricavato da corpus fu il *Collins Cobuild English Language Dictionary* (London, Harper Collins, 1987[1], 1995[2], 2001[3], 2003[4], 2006[5]). I lemmi sono stati estratti da oltre 524 milioni d'occorrenze della *Bank of English*, costituita dall'editore Harper Collins e dall'Università di Birmingham dagli anni Ottanta in poi. Su questo corpus si basano una serie di opere (*Advanced Learner's English Dictionary,* Glasgow, Harper Collins, 2001[3]*; Concise Learner's Dictionary,* London, Harper Collins, 2003[2] ecc.), accompagnate da cd-rom, in cui è incluso l'accesso a una parte della *Bank of English*, la cosiddetta *Word Bank* (composta di testi scritti e parlati con cinque milioni di occorrenze) che ha costituito la base della prima edizione e dei successivi aggiornamenti. Il criterio decisivo della selezione dei lemmi è stata, appunto, la frequenza d'occorrenza, pertanto solo lemmi ricorrenti con una frequenza prestabilita sono stati ammessi nel dizionario.

Notiamo che per adesso lo Zingarelli è l'unico dizionario italiano, che (almeno) nei suoi aggiornamenti ricorre a dati lessicali ricavati da corpora. Nella compilazione si è fatto uso sistematico del *Corpus Italiano Zanichelli*, un database costituito dalle principali opere classiche della letteratura italiana (pubblicate nel cd-rom LIZ 4.0, curato da Stoppelli e Picchi, 2001) da annate e periodici, testi provenienti dalle fonti più svariate,

[4] Inoltre, Bergenholtz (1990: 775) richiama l'attenzione sul fatto che tradizionalmente alcuni ambiti, come i termini della linguistica, sono quantitativamente più numerosi di quelli degli altri campi, ad esempio le voci relativi alla sessualità e non solo. Con una lemmatizzazione incentrata sull'analisi statistica di corpora parlati, oltre che scritti, tali lacune potrebbero essere eliminate.

compreso Internet. Questi procedimenti offrono vantaggi decisivi, poiché permettono una valutazione più precisa dell'uso corrente della lingua: si pensi all'inserimento di numerose parole della lingua parlata o di linguaggi settoriali presenti negli scritti giornalistici (economia, nuove tecnologie ecc.).
Pare che in Italia, non sia stata ancora creata una banca di dati linguistici basata su un corpus continuamente aggiornato e incrementato, dal quale si possano trarre i materiali per ricerche lessicologiche o per confezionare i più svariati prodotti lessicografici. Che questa sia ormai una via obbligata della linguistica e della lessicografia moderne, emerge anche dal fatto che il dizionario enciclopedico della lingua ungherese (*A magyar nyelv nagyszótára*, Budapest, Akadémiai Kiadó, 2006), attualmente in fase d'elaborazione, si basa su ampi corpora. Si prevede che l'intera opera, di cui finora sono usciti il primo volume e il supplemento, conterrà 110.000 lemmi. Il dizionario si fonda su un corpus esteso che include tre tipi di raccolte: il corpus storico della lingua ungherese (*Magyar Történeti Szövegtár*), contenente 27 milioni d'occorrenze, ricavate da testi scritti tra il 1772 e il 2000; un archivio cartaceo di 6 mila lemmi, raccolti tra la fine del XIX secolo (quando nel 1898 emerse per la prima volta il progetto del dizionario) e l'inizio degli anni Ottanta; più una raccolta di testi registrata su cd (per i dati bibliografici delle opere racchiuse nei tre corpora cfr. Ittzés 2006: 51-797). Come primo passo, i lemmi sono stati estratti dal database elettronico in base alla loro frequenza, insieme alla data della prima e l'ultima attestazione. Questo procedimento ha permesso di valutare l'attendibilità del corpus: si è arrivati alla conclusione che in base al solo corpus storico non si poteva realizzare il dizionario, pertanto l'hanno integrato con l'archivio cartaceo e una raccolta di testi registrati su cd (Ittzés 2006: 20, I vol.). Ogni lemma e ciascun'accezione è attestata nel corpus ampliato.
Analogamente ben documentata è l'elaborazione della seconda edizione aggiornata del dizionario enciclopedico ungherese (*Magyar Értelmező Kéziszótár,* 2003): il lemmario della versione precedente è stato arricchito con termini ricavati da una serie di dizionari pubblicati dagli anni Novanta in poi (cfr. pp. V-VII) e dal *Corpus nazionale ungherese*, comprensivo di 150 milioni di occorrenze, rappresentativo della lingua ungherese parlata in Ungheria e nei Paesi confinanti (cfr. anche 3.3).
Sembra quindi che, per garantire la coerenza e l'omogeneità della selezione (e della presentazione) del materiale, sia indispensabile ricorrere (anche) a dati statistici ricavati da ampi corpora.

4.1.2 La definizione lessicografica

Oggetto delle nostre considerazioni è la definizione lessicografica che tuttora solo in parte risente dei risultati ottenuti dalla linguistica quantitativa. Senza entrare nei dettagli sui vari procedimenti definitori, occorre rilevare che i due difetti principali da cui il lessicografo deve guardarsi nella formulazione delle definizioni, sono la circolarità diretta (cioè di definire la parola A tramite la parola B, definita a sua volta con A) e l'uso di parole non elencate come lemma. Si avverte che quest'ultima regola vale non solo alla parola come lemma, ma anche ai suoi sensi specifici usati nelle definizioni, perché spesso è il caso che, sebbene la parola sia riportata nel dizionario, quel senso specifico che è usato nella definizione, non è elencato.[5] Al fine di evitare definizioni

[5] Gli autori del dizionario enciclopedico ungherese (*Magyar Értelmező Kéziszótár,* 2003) tengono a rilevare nella prefazione che nelle definizioni si usano esclusivamente parole che sono registrate come lemma, tuttavia la effettiva presenza di tutti i sensi utilizzati nelle definizioni non emerge come criterio.

evidentemente circolari ed incomprensibili, alcune opere lessicografiche fanno uso del cosiddetto vocabolario definitorio, in altre parole *defining vocabulary*.
Il *defining vocabulary* si riferisce al metodo di costruire le definizioni del dizionario non da tutte le parole della data lingua, ma ricorrendo ad un insieme relativamente ristretto di parole selezionate per tale scopo. La definizione lessicografica, a differenza delle altre informazioni metalinguistiche sulla parola è (dovrebbe essere) un'informazione 'familiare' che appartiene ad un discorso di livello usuale. Tale linguaggio 'naturale' dovrebbe garantire al lettore medio la comprensibilità della definizione. Ad esempio West (1935) nel suo lavoro d'importanza fondamentale ha selezionato su basi logiche 1.490 parole per definire 24.000 elementi del vocabolario inglese. Per selezionare il *defining vocabulary*, di solito i dizionari di frequenza costituiscono un primo punto di partenza, poiché la 'familiarità', uno dei criteri principali che si pone nei confronti delle parole definitorie, spesso corrisponde a un certo grado di frequenza. Mentre la frequenza e la familiarità possono dare suggerimenti iniziali per la selezione, Neubauer (1990: 903) osserva che l'utilità lessicografica, vale a dire la capacità definitoria di una parola ha un ruolo altrettanto rilevante in questo processo.
Le opinioni sull'uso del *defining vocabulary* sono controverse. La motivazione principale per la loro applicazione è di evitare di usare nella spiegazione del lemma parole ancora più difficili del lemma da spiegare, il che certamente frustra non solo gli apprendenti di una lingua straniera, ma anche gli utenti di madrelingua. Eppure non tutti (Rudolf e Hartmann 1990, Cardinale 1994) sono concordi nel ritenere che la restrizione del vocabolario delle definizioni a un nocciolo centrale faciliti automaticamente la comprensione da parte dell'utente, poiché queste parole ad alta frequenza sono anche quelle più polisemiche ed idiomatiche. Pertanto, usando un vocabolario ristretto le definizioni perdono necessariamente della loro accuratezza. Proprio per questo diventa problematica la definizione dei termini tecnici, ciononostante non mancano dizionari scientifici, in cui le definizioni sono formulate con un vocabolario definitorio controllato.[6] Si afferma poi, che la spiegazione del significato di un termine può giovarsi della competenza dell'utente relativa a tutto il linguaggio, proprio perché l'acquisizione della competenza linguistica avviene in maniera globale. Secondo Cardinale, perfino nei dizionari fondamentali appare artificiosa l'ipotesi, fatta propria dal dizionario di francese di 5.000 parole di Matoré (1963), di far entrare nelle definizioni solo il vocabolario di 5.000 parole, descritto con una perfetta identità tra metalinguaggio e linguaggio-oggetto, che ha imposto agli autori un enorme esercizio di semplificazione e "predispone i lettori ad un innaturale accostamento alla lingua attraverso una gradualità che non appartiene alla realtà." Nonostante tutte le critiche, quest'approccio è diventato particolarmente importante con la divergenza dei dizionari monolingui generali in dizionari monolingui per parlanti nativi e dizionari monolingui per apprendenti stranieri della lingua, i cosiddetti dizionari pedagogici (*learners' dictionary*).
Un esempio di un progetto professionale condotto con l'ausilio di un *defining vocabulary* è il *Longman's American Idioms dictionary* (2000) che definisce i 4.000 idiomi più comuni dell'inglese americano ed effettivamente spiega molti aspetti della cultura americana, usando esclusivamente 2.000 parole nelle loro accezioni più comuni con i prefissi e suffissi più semplici. Non è questo l'unico dizionario della casa editrice che fa uso del Longman *defining vocabulary*, si vedano ad esempio, oltre ai dizionari pedagogici anche il *Longman Dictionary of American English* (2001), il *Longman Dictionary of Contemporary English* (2005[2]), comprensivi rispettivamente di 52.000 e di

[6] Si vedano ad es. Graham E. (1965) *Science Dictionary in Basic English*, London; Flood W. E., West M. (1962) *An Elementary Scientific and Technical Dictionary*, London, pubblicato per la prima volta nel 1952 come *An Explaining and Pronouncing Dictionary of Scientific and Technical Words*, London.

200.000 termini. Questi dizionari sono stati ricavati dal *Longman Spoken and Written English Corpus* di 40 milioni di occorrenze. Chiari (2007: 95) osserva che in questi dizionari basati su corpora, anche i canoni di definizione dei lemmi, tradizionalmente un po' ellittici, e le spiegazioni assumono il carattere di esempi rappresentativi molto più immediati e meno simili a voci d'enciclopedia.[7]

Nei dizionari monolingui Neubauer (1990: 900) osserva le seguenti strategie riguardo al controllo del vocabolario:

a) nessun controllo sulle parole usate nelle definizioni;
b) si cerca di usare un 'linguaggio semplice' nelle definizioni, ma senza restrizioni esplicite;
c) le parole usate nelle definizioni sono selezionate da una lista specifica contenente il c.d. *defining vocabulary*, cioè le parole prescelte per essere usate nelle definizioni, in più alcune parole non contenute nel *defining vocabulary* sono usate come rinvii e terminologia (definizione di tipo misto);
d) tutte le parole occorrenti nelle definizioni sono elencate nel dizionario e fanno parte di un *defining vocabulary* compilato esplicitamente per questo scopo;
e) tutte le parole occorrenti nelle definizioni fanno parte di un *defining vocabulary,* compilato esplicitamente per questo scopo, in cui sono marcati tutti i sensi usati e solo i sensi elencati nel dizionario sono usati nelle definizioni.

Per illustrare in che modo l'uso di un *defining vocabulary* influisce sulle definizioni, riportiamo di seguito degli estratti da vari vocabolari che hanno in comune d'essere dizionari monolingui d'italiano in un volume. Per la comparazione ho scelto l'entrata *macaco*, una voce appartenente al linguaggio della zoologia che potrebbe risultare sconosciuto perfino per i parlanti nativi.

a) Definizione senza controllo di vocabolario

> **macàco** o (raro) **macàcco** [port. macaco, vc. proveniente dall'Angola; 1704 ca.] s. m. (pl. -chi) **1** Genere di scimmie dei Cercopitecidi con coda pendente non prensile, callosità nelle natiche e arti anteriori non più lunghi dei posteriori (Macacus). SIN. Maimone. **2** (f. -a) (fig.) Uomo goffo e sgraziato, o sciocco, stupido. || **macacóne**, accr.
>
> Estratto da: Zingarelli (2006: 1030)

Questo è un chiaro esempio dell'indifferenza verso la difficoltà del vocabolario al fine di fornire informazioni precise, facendo uso della terminologia corrispondente (una simile definizione la potremmo ricavare da un trattato di zoologia). Tuttavia per un apprendente straniero, che apparentemente non fa parte del gruppo dei destinatari di questo dizionario, non è di grande aiuto l'informazione che il macaco appartenga alla famiglia dei Cercopitecidi, né il fatto di essere mandato da *macaco* a *callosità, prensile* e *natica* che probabilmente gli sono sconosciuti quanto *macaco*. È plausibile inoltre, che il vocabolario utilizzato in questa definizione causerebbe problemi di comprensione perfino per molti parlanti nativi di limitata competenza linguistica, o semplicemente poco esperti di zoologia. Sembra giustificato porre la domanda, quali siano le differenze tra un articolo d'enciclopedia e un'entrata del dizionario. Non è chiaro se l'utente consulterà un dizionario generale monolingue per reperire informazioni scientifiche e terminologiche del genere.

[7] Notiamo con Neubauer (1990: 904) che le considerazioni fatte per il vocabolario usato nelle definizioni valgono anche per quello degli esempi. Benché la funzione centrale delle frasi sia quello di informare sulle caratteristiche sintattiche (piuttosto che semantiche) dei lemmi, a ragione l'utente dovrebbe essere in grado di capire le parole usate negli esempi, il che non è possibile se essi contengono un vocabolario totalmente incontrollato.

b) Definizione in un "linguaggio semplice"

macaco [ma-cà-co] non com., *s. m.* [pl. *-chi*] **1** genere di scimmie grigio-brune dal corpo tozzo, con coda breve e talora del tutto assente e visiera sopraorbitale molto vistosa (*ord*. Primati) **2** (*fig*.) persona sciocca o goffa.
Dal port. *macaco*, da una voce africana.

Estratto da: Garzanti 2005 (cd-rom)

Il dizionario Garzanti si propone di offrire all'utente "gli strumenti per interpretare con sensibilità e rigore lo stato della lingua, i suoi mutamenti, le permanenze e le novità." Gli autori hanno fatto a meno della terminologia scientifica, tuttavia parole come *sopraorbitale* e *visiera* sicuramente non facilitano la comprensione per gli utenti. C'è da dubitare, inoltre che l'appartenenza dei macachi all'ordine dei Primati costituisca parte integrante di una definizione lessicografica.

c) Definizione di tipo misto

ma|cà|co s.m. **1a TS** zool. com., nome comune delle scimmie del genere Macaca **1b OB TS** zool. macaca **2 CO** fig., scherz., uomo goffamente brutto e sciocco **Varianti**: macaco

Polirematiche
macaco mulatto loc.s.m. **TS** zool.com., scimmia del genere Macaca (Macaca mulatta o Macacus rhesus) diffusa in India e nota perché su di essa furono condotti gli esperimenti per determinare il fattore RH del sangue

ma|cà|ca s.f.**1 TS** zool., scimmia del genere Macaca, di statura media e corporatura robusta, diffusa in Asia e Africa comunemente detta macaco | con iniz. maiusc., genere della famiglia dei Cercopitecidi **2 CO** fig., donna goffamente brutta e sciocca

Estratto da: De Mauro 2000 (versione on-line)

Il dizionario di De Mauro si propone di "unire la semplicità definitoria con una certa essenzialità." Facendo uso primariamente delle parole del vocabolario comune (ca. 50.000 termini compresi da chiunque abbia un livello d'istruzione mediosuperiore), per definire i lemmi più complessi si ricorre anche a termini tecnici, spiegati a loro volta con parole appartenenti al vocabolario comune. Un'altra caratteristica importante è che tutte le parole utilizzate nelle definizioni compaiono a lemma, ad eccezione di alcuni raggruppamenti tassonomici. Nella trattazione dei nomi di piante e animali molto conosciuti "si è preferito indicare anzitutto i tratti legati all'esperienza comune rispetto a quelli legati alla posizione nella tassonomia scientifica", come recita la prefazione.
Nel caso delle parole appartenenti al Vocabolario di base (parole fondamentali – FO, di alto uso – AU e di alta disponibilità – AD), De Mauro (2000) riporta una definizione rispondente all'utilizzo quotidiano e al senso più prossimo con cui lo si adopera, separando con una barra e introducendo la marca d'uso tecnico-specialistico solo quando introduce la definizione specifica.

mam|mì|fe|ro s.m. **1 AU** animale che partorisce la prole e la alleva nutrendola con il proprio latte; anche agg.: *la balena è un animale m.* **2 TS** zool., vertebrato omeotermo della classe dei Mammiferi, con due paia di arti atti alla locomozione, cute provvista di peli, caratterizzato da respirazione polmonare, circolazione sanguigna doppia e completa, riproduzione vivipara e dalla presenza di ghiandole mammarie nella femmina | pl. con iniz. maiusc., classe del phylum dei Vertebrati, cui appartiene anche l'uomo

Estratto da: De Mauro 2000 (versione on-line)

L'ordinamento delle definizioni in base alla marca permette di esplicitare il significato in modo da aderire al senso in cui la maggioranza dei parlanti adopera il vocabolo. Osserviamo con Ferreri (2006: 138) che "la marca con la forza degli usi di un'intera comunità impone un ripensamento sulle definizioni approntate per il dizionario" (v. infra).

d) Definizione tramite un vocabolario di base

> **macaco** (ma.ca.co) s.m. scimmia di medie dimensioni tipica dell'Asia orientale; ha una folta pelliccia e talvolta può essere priva di coda: *Il macaco del Giappone lava il cibo prima di mangiarlo*.
> [**G**] plur. **macachi**.
> [**E**] dal portoghese **macaco**, da una voce africana **kaku kaku** che imita il grido dell'animale.
>
> Estratto da: De Mauro (1996: 767)

La definizione evita ogni tecnicismo, collocandosi perfettamente al livello di un preadolescente (allievi delle elementari e medie, cui il dizionario è destinato) o di un adulto di limitata competenza linguistica. Si evitano le parole che non figurino nel lemmario costituito di 15.000 voci, di cui 7.000 appartenenti al *Vocabolario di base* (cfr. anche 2.2.27). Si è rinunciato a raggruppamenti tassonomici, all'uso di termini tecnici, e si è preferito indicare anzitutto i tratti legati all'esperienza comune. Gli esempi privilegiano, infatti, gli aspetti nozionali (qual è il luogo di provenienza tipica del macaco? il Giappone), mentre un dizionario per adulti riterrebbe superflua l'esemplificazione per un lemma del genere. Notiamo poi, che contrariamente agli altri tre dizionari, qui non è inclusa l'accezione figurata del termine.
Concludendo, vogliamo rilevare l'utilità dell'uso di *defining vocabulary* non solo nei dizionari destinati agli studenti e/o apprendenti stranieri, bensì anche in quelli pensati per parlanti di madrelingua, giacché esso contribuisce alla riduzione della circolarità, al contempo migliora il grado di coerenza e comprensibilità delle definizioni e, in fin dei conti, del dizionario stesso.

4.1.3 Dati di frequenza e marche d'uso

La tradizione lessicografica specifica con un apposito apparato di notazioni gli usi marcati dei significati delle parole. Fondandosi sul grado di distanziamento da una supposta normalità, il lessicografo assegna la marca d'uso per lo più in base alla propria coscienza linguistica. Le scelte attuate nelle descrizioni lessicali appaiono quindi assai soggettive, non sostenute da dati oggettivi. Questo tipo di procedimento empirico è spesso guardato con un certo sospetto, a volte è ritenuto prescientifico o addirittura ascientifico. In alcuni dizionari moderni invece, la marcatura si basa sulle informazioni relative alle frequenze d'uso e può avere due forme: marche d'uso o dati quantitativi estratti da corpora (Chiari 2007: 85). Per illustrare il valore informativo dei dati di frequenza, si cita Arcaini (1978: 272):

> "l'informazione in 'aprire' è ben diversa da quella poniamo di 'salassare' o 'solcare', e ciò in base alla sua probabilità d'occorrere in certi contesti. Questa probabilità concorre dunque a definire 'aprire' rispetto ad altri termini."

La frequenza quindi *fa parte* della definizione di un lemma. Inoltre, i dati statistici sono importanti non solo perché informano sulla frequenza d'uso delle parole, ma anche

perché sono un importante punto di riferimento nella classificazione dei vari significati dei lemmi polisemici, come vedremo nella sezione successiva.
Dizionari famosi rieditati negli ultimi anni e perfino il DISC (1997) prendono in considerazione il criterio della frequenza che permette di evidenziare alcune migliaia di parole, appartenenti al vocabolario di base. Il vocabolario fondamentale può assolvere ai più elementari bisogni comunicativi, perciò è giusto contrassegnarlo con particolari caratteri grafici che permettono una rapida individuazione. Si tratta di un'indicazione particolarmente utile che consente all'utente di scegliere, in determinate circostanze comunicative, vocaboli più noti e diffusi rispetto a quelli meno noti.
Nell'edizione curata da Serianni e Trifone del Devoto-Oli (2004-2005, pp. V-VI) si legge:

> "L'indicazione delle diecimila parole fondamentali (evidenziate in azzurro) nasce dal confronto tra numerosi repertori (lessici di base, lessici di frequenza, dizionari che contengono un numero ridotto di lemmi, dizionari che segnalano i vocaboli di uso più comune). Il lemmario derivante da tale comparazione è stato naturalmente integrato e modificato anche sulla base dell'esperienza e della sensibilità linguistica del lessicografo."

Nella *Guida all'uso* del dizionario Sabatini-Coletti (DISC) alla voce *La "disponibilità" delle parole* (p. XIII) si trova quanto segue:

> "Hanno ricevuto il contrassegno di alta disponibilità (un fondino grigio sotto il lemma) circa 10.000 lemmi, scelti sulla base delle liste di frequenza oggi esistenti per la lingua italiana (da quella di Bortolini-Tagliavini-Zampolli 1972, al LIP di De Mauro-Mancini-Vedovelli-Voghera 1993), intervenendo però su di esse con valutazioni personali per espungere o integrare, considerando, della singola parola, almeno la più comune delle accezioni."

Più chiaro ed esplicito Zingarelli (2006, p. 8) che recita:

> "I circa 4500 lemmi di maggiore frequenza d'uso (l'italiano fondamentale) sono preceduti da un piccolo rombo (◊): queste parole sono state scelte sulla base dei lessici di frequenza (fra cui il *Lessico di frequenza della lingua italiana* a cura di U. Bortolini, C. Tagliavini e A. Zampolli, Garzanti, 1972, il *Lessico di frequenza dell'italiano parlato* di T. De Mauro, F. Mancini, M. Vedovelli, M. Voghera, Etaslibri, 1993 e il *Lessico elementare* di L. Marconi, M. Ott, E. Pesenti, D. Ratti, M. Tavella, Zanicheli, 1994) e in base ad analisi statistiche del *Corpus Italiano Zanichelli*."

I tre dizionari condividono l'esigenza di indicare all'interno del lemmario un nucleo di parole tra le più frequenti e usate. Tutti fanno riferimento ai lessici di frequenza, alle liste d'uso ecc., ciononostante né le scelte quantitative, né le soluzioni terminologiche adottate sono omogenee (cfr. Ferreri 2006). La tabella seguente evidenzia l'oscillazione.

Zingarelli: italiano fondamentale	4.500 lemmi
DISC: parole d'alta disponibilità	10.000 lemmi
Devoto-Oli: parole fondamentali	10.000 lemmi

Comparazione dei sottoinsiemi dei lemmari

Sorprende anzitutto la differenza numerica delle parole isolate, da 4.500 a 10.000, ma colpiscono ancora di più le diversità della terminologia adoperata. La condivisione di una cifra – diecimila – (DISC e Devoto-Oli) non significa che le due fasce coincidano: i due insiemi di 10.000 parole, come vedremo oltre, non nascondono solo differenze di poco rilievo, ma anche difformità sostanziali, dovute a procedure diverse di selezione. L'utilizzo dell'aggettivo 'fondamentale' non esprime lo stesso concetto nelle opere che lo adottano (Zingarelli e Devoto-Oli). Ricordiamo, infatti, che la tradizione di isolare il vocabolario fondamentale dal resto dei lessemi in Italia è strettamente legata alla messa a punto del *Vocabolario di base* (VdB) (De Mauro *et al.* 1980), caratterizzato da una puntuale ripartizione in tre fasce. In questa proposta, le designazioni *vocabolario di base*,

parole *fondamentali*, *d'alto uso* e *d'alta disponibilità* non sono assolutamente intercambiabili, poiché definiscono fasce di lessico che fanno riferimento a lemmi ben definiti in base alla frequenza e dell'uso, per le prime due fasce, e alla disponibilità psicologica per parlanti con livelli d'istruzione medio-bassi per la terza fascia (cfr. 2.2.15).
Ciascuno dei tre dizionari fa riferimento ai lessici di frequenza, anche in questo caso, però, sono necessarie ulteriori chiarimenti. Lo Zingarelli parla di 4.500 lemmi tratti dai lessici di frequenza, soglia che non corrisponde a quella dei lessici di riferimento: il LIF e il LIP includono 5.000 vocaboli, mentre il LE contiene 6.000 lemmi. Emerge quindi, un primo problema da chiarire: come e perché sono state eliminate 500 parole del LIF e LIP e 1.500 del LE. Per il resto, tutti i dizionari rinviano a vocabolari pedagogici o di base, senza indicare tuttavia le fonti di riferimento. Nel Devoto-Oli e nel DISC sui dati oggettivi domina la sensibilità linguistica del lessicografo, come viene peraltro dichiaratamente espresso (v. sopra). Lo Zingarelli è l'unico che, secondo quanto esplicitato, non ammette tra i suoi criteri di selezione la valutazione soggettiva.
Le varie designazioni dei dizionari si potrebbero considerare un elemento non essenziale, che non modifica la struttura delle selezioni effettuate. Tuttavia il confronto diretto evidenzia differenze notevoli che effettivamente non è facile ricondurre a precise intenzioni dei lessicografi, dal momento che la formulazione dei criteri adottati per selezionare i sottoinsiemi dei lemmari è insufficiente.
Per illustrare quanto detto, si riportano i dati risultanti da una comparazione da *f* a *filosofo* dei lemmi appartenenti all'italiano fondamentale nel Devoto-Oli, nel DISC e nello Zingarelli. Riportiamo i lemmi presenti in solo uno dei tre dizionari, nella colonna a destra confronto con le marche d'uso presenti in De Mauro (2000).[8]

N. di lemmi	Devoto-Oli (2004-2005)	DISC (1997)	Zingarelli (2006)	De Mauro (2000) (marche d'uso)
	fa (s.m.)			AD TS
		fabbisogno		CO
			fabbricante (s.m.)	CO
	facilmente			CO
		falsario		CO
		falsificare		CO
		famigerato		CO OB
		fanalino		CO
		fanciulla		AU
		fantapolitica		CO TS
	fantasticare			CO
	fante (s.m.)			CO TS
	faraona			CO TS
	fascina			AU
		fasullo		CO
	fatalmente			CO
	fato			LE
	fattispecie			TS
	fatturare			TS
	fatturato (s.m.)			TS
	fazioso			CO
	fecondazione			TS
		femminilità		CO
		femminismo		CO
		femminista		CO
			fenicio	TS
	fenomenale			CO
		feriale		AD
	fermamente			CO
	fermento			CO TS
		ferramenta		CO
		ferrarese		AU
		fertilizzante		CO

[8] Il confronto con i dizionari De Mauro (1999 e 2000) è reso obbligatorio dalla sistematica adozione delle marche d'uso per l'intero lemmario.

	fesseria		CO
festeggiamento			CO
	fetente		CO
	fettina		CO
fiancata			CO OB TS
	fica		CO OB
		figurina	CO
filamento			CO TS
		filante (agg. e s.m.)	CO TS
	filastrocca		AD
filiale (agg. e s.f.)			CO
	filippino		CO
filone			CO TS

Comparazione da *f* a *filosofo* dei lemmi appartenenti all'italiano fondamentale nel Devoto-Oli (2004-2005), nel DISC (1997) e nello Zingarelli (2006). Nella colonna a destra confronto con le marche d'uso presenti in De Mauro (2000).

Nel piccolo campione esaminato si trovano complessivamente 218 lemmi, appartenenti all'italiano 'fondamentale', di cui 193 nel Devoto-Oli, 192 nel DISC e 98 nello Zingarelli, il che dimostra una buona consistenza numerica. L'equivalenza numerica dà l'apparenza illusoria di una sovrapposizione completa degli insiemi lessicali: in realtà, ci sono molte zone di sovrapposizione ma anche vistose differenze. Vocaboli elencati nel DISC come di 'alta disponibilità' sono assenti nell'insieme delle 'parole fondamentali' del Devoto-Oli; di contro, quest'ultimo include lemmi non marcati nel DISC. I lemmi *fabbisogno, falsario, famigerato, fantapolitica, ferramenta, fertilizzante* si trovano segnalati tra i diecimila nel DISC ed esclusi dai 10.000 fondamentali del Devoto-Oli, tra i quali figurano invece *fantasticare, fante, fatturare, fecondazione, fiancata* che l'altro non segnala come frequenti. In alcuni casi perfino lo Zingarelli, il cui lemmario è molto più ridotto (4.500 contro i 10.000 lemmi del DISC e del Devoto-Oli) contiene alcuni termini che non figurano in nessuno degli altri due dizionari. Né il DISC, né il Devoto-Oli marca *fabbricante, fenicio, figurina, filante* che sono presenti nello Zingarelli, in cui la mole del lemmario non raggiunge nemmeno la metà degli altri due.
Il dizionario Devoto-Oli immette fra le parole fondamentali vocaboli quali *facilmente, fantasticare, fatalmente, fazioso, fenomenale, filiale* e alcuni termini tecnici, come *fatturare* e *fatturato*. Controllando la marcatura nel GRADIT si nota, infatti, la presenza della marca TS, giacché l'ambito nel quale circolano i due lemmi è tipicamente commerciale. Altri lemmi, come *fante, faraona, fermento, fiancata, filamento, filone*, nel GRADIT sono marcati con un doppio contrassegno: la marca TS è affiancata da CO.

fer|mén|to s.m. **1 TS** biochim., sostanza in grado di indurre un processo fermentativo: *fermenti acetici*, *lattici* | ⇒enzima **2 CO** fig., spec. colloq., stato di particolare tensione, di inquietudine latente: *gli animi erano in f. per l'imminente battaglia*, *in città c'è f. per l'arrivo del presidente* | fervore creativo | anelito a nuovi valori e ideali che promuovano il mutamento di una condizione interiore o di un ordine costituito: *fermenti di rivolta* **3 TS** st.relig., nei primi secoli cristiani, la particola dell'eucarestia che il papa mandava alle chiese titolari della città come simbolo dell'unità ecclesiale

Polirematiche
fermento lattico loc.s.m. **TS** biochim., fermento o enzima che degrada il lattosio in acido lattico e anidride carbonica, utile in certe diete e terapie di affezioni intestinali.

Estratto da: De Mauro 2000 (versione on-line)

Come si vede dall'articolazione in tre accezioni distinte, di cui due marcate TS, è difficile immaginare che siano fondamentali tutte le accezioni, giacché rinviano ad alcuni sensi molto particolari, inconsueti nell'uso quotidiano. Poiché ciascuno dei tre dizionari assegna la marca all'entrata senza distinguere tra le varie accezioni, non sorprende che

molte volte anche significati tecnici sono marcati come 'fondamentali'. Le tre accezioni della parola differiscono sostanzialmente nella loro sfera d'uso, infatti, due sotto-significati su tre sono usati esclusivamente nell'ambito specialistico, ciononostante il Devoto-Oli non esita ad assegnare la marca 'fondamentale' all'entrata nel suo insieme.
D'altro lato mancano nel Devoto-Oli parole appartenenti alla fascia AU (*fanciulla, ferrarese*) ed AD (*feriale, filastrocca*), presenti nel DISC, di contro è stato incluso *fato*, di carattere nettamente letterario.
La selezione effettuata nel DISC sembra più equilibrata. Tendenzialmente sono omesse le parole marcate TS, tuttavia anche qui ci sono delle lacune. Manca il vocabolo *fascina*, marcato come AU, di contro è presente *fantapolitica*, sulla cui appartenenza al vocabolario fondamentale c'è da dubitare.

> **fan|ta|po|lì|ti|ca** s.f. **CO TS** lett., tipo di narrativa che ha per soggetto l'anticipazione ipotetica di fatti politici e bellici trattati spec. in chiave satirica e paradossale| **CO** *fare della f.*: azzardare ipotesi politiche inconsistenti, assolutamente fantasiose e utopistiche
>
> Estratto da: De Mauro 2000 (versione on-line)

È curiosa la scelta degli autori di includere tra le parole di base termini come *femminilità, femminismo, femminista*, oppure l'aggettivo *famigerato*, marcato CO OB. Un approccio meno purista rispetto al Devoto-Oli è testimoniato dall'inclusione nel DISC di *fica*, termine marcato CO OB e volgare da De Mauro.
Tra le parole 'fondamentali' dello Zingarelli, sorprende la presenza di *fenicio* (TS) indicante oggetti e persone appartenenti alla Fenicia, e di *filante* (CO TS) che, usato come sostantivo, indica l'alterazione del vino bianco, causata da batteri.
Gli esempi discussi rendono evidente che la coscienza linguistica del lessicografo, per quanto esperto, non può sostituire il rilevamento statistico degli usi di una comunità. Si è dimostrato che

> "per quanto attento e colto, il singolo studioso non può cogliere il comportamento medio di un'intera comunità di parlanti che si rivela nella sua oggettività con spogli e registrazioni di entità sufficienti a garantire l'attendibilità dei risultati" (Ferreri 2006: 127).

Una marcatura più analitica e più complessa si trova nella serie di dizionari di De Mauro, in particolare nel GRADIT, *Grande dizionario italiano dell'uso* (1999, UTET), nel *Dizionario della lingua italiana per il terzo millennio* (2000, Paravia) e nel *Dizionario dei sinonimi e contrari* (2002, Paravia).[9] Con un'operazione insolita nella tradizione lessicografica italiana (ed estera), De Mauro, ideatore e direttore del GRADIT in collaborazione con G. Lepschy ed E. Sanguineti, marca sia l'intero lemmario di ca. 250.000 lemmi, sia le accezioni interne. Sappiamo che non tutte le accezioni di una parola hanno la stessa frequenza: alcune sono assolutamente periferiche, rispetto ad alcuni sensi più adoperati. Dal punto di vista della marcatura ciò implica l'adozione di marche separate per ciascun gruppo di sensi, altrimenti si perderebbe appunto la significatività delle marche. Si è visto tuttavia, che gli autori delle tre opere esaminate (Devoto-Oli 2004-2005, DISC 1997 e Zingarelli 2006) non sembrano preoccuparsi di questo fatto: marcano il lemma una sola volta, estendendo la validità dell'indicazione a tutte le accezioni incluse nel lemma. Di contro De Mauro marca separatamente ogni accezione (cfr. sopra la marcatura di *fermento*), mentre nei tre dizionari esaminati la

[9] Oltre alle opere citate (De Mauro 2000 e 2002), ricordiamo che i tre dizionari pedagogici di De Mauro e Moroni (1996, 1997a, 1997b) adottano nei loro lemmari la marcatura relativa alle tre fasce del *Vocabolario di base*: FO, AU, AD. Nel DIB (1996) ad esempio, una mezza luna posta prima del lemma indica le parole d'alta frequenza; altri simboli sono la luna piena, per le parole fondamentali e il quarto di luna per le parole d'alta disponibilità. Trattandosi di dizionari di mole ridotta, non si marcano separatamente le accezioni.

marcatura separata è riservata agli omonimi che figurano come lemmi autonomi nel dizionario.[10]
L'operazione di marcatura nel GRADIT si fonda sull'incrocio di quattro diversi tipi d'informazioni provenienti dalla statistica linguistica, da test psicologici e prove di comprensione, dalle aree di provenienza delle parole e dai campi del sapere cui sono legate (cfr. Ferreri 2006).

Frequenza e uso	FO, AU, BU, OB
Provenienza	DI, RE, ES
Appartenenza	TS, LE
Comprensibilità	AD, CO

Criteri e marche adottati nel GRADIT.

Il criterio connesso alla frequenza determina le due fasce del VdB: le parole fondamentali (FO) e quelle d'alto uso (AU) da un lato, e di basso uso (BU) e obsolete dall'altro (OB); il criterio della provenienza delle parole prima di stabilizzarsi nella lingua standard, genera esotismi (ES), dialettalismi (DI), regionalismi (RE). Dall'appartenenza ad un campo del sapere deriva la marcatura (TS) delle parole tecnico-specialistiche, con la specificazione relativa al tipo di sapere – si tratta del settore più ampio con ca. 130.000 lemmi –, e dei vocaboli d'uso esclusivamente letterario (LE). Il quarto criterio è quello della comprensibilità, che dà luogo a parole del vocabolario comune (CO) e di alta disponibilità (AD). Le marche d'uso introdotte nel GRADIT e riprese in altre opere dello stesso autore, si fondano su molteplici verifiche, e, secondo la Ferreri (2006: 128), "non hanno margini sporchi". Questo tipo di marcatura è difatti unico, perché rende visibili le stratificazioni delle parole, giacché parte dagli usi che i parlanti ne fanno.[11]
Quanto detto implica che le fonti sulla cui base si procede all'attribuzione delle marche siano rese esplicite, permettendo in questo modo la verificabilità dell'attribuzione delle marche. Nel GRADIT i testi presi come base del lemmario e le fonti consultate per giungere all'attribuzione delle marche, sono elencati meticolosamente in De Mauro (2005: 38-42), per permettere ulteriori controlli.
Occorre menzionare ancora due dizionari ungheresi. Ambedue includono, accanto alle marche d'uso, dati statistici sull'uso dei singoli lemmi, ricavati da larghi corpora. Nella versione aggiornata del dizionario enciclopedico della lingua ungherese (*Magyar értelmező kéziszótár*, Budapest, Akadémiai Kiadó, 2003) è stata inserita, per la prima volta nella tradizione lessicografica ungherese, la frequenza dei singoli lemmi, ricavata dal corpus nazionale della lingua ungherese (*Magyar Nemzeti Szövegtár*) contenente ca. 150 milioni di parole (cfr. 3.3). Il dizionario non dà il valore di frequenza assoluto, ma classifica le parole secondo una graduatoria di cinque fasce. Per ogni lemma s'indica la fascia di frequenza cui appartiene: nel primo gruppo rientrano i lemmi con ranghi 1-2.000, seguono quelli tra 2.001-10.000, tra 10.001-30.000, 30.0001 e 60.000, mentre all'ultimo gruppo appartengono i lemmi con ranghi superiori a 60.000). Gli autori rilevano che nonostante la puntuale annotazione del corpus, non è possibile una distinzione tra le varie accezioni dei lemmi polisemici, giacché "i moderni sistemi di trattamento automatico del linguaggio non sono in grado di elaborare il significato delle parole" (p. XIV). I dati di frequenza, infatti, si riferiscono sempre all'intero lemma, senza distinguere tra la frequenza d'uso delle varie accezioni. Mentre gli omonimi funzionali

[10] Ad esempio il DISC marca come 'fondamentale' l'entrata *fa* come avverbio temporale, ma non *fa* come sostantivo.
[11] Di contro, nel Devoto-Oli, ad esempio, le marche d'uso sono "ispirate al criterio dell'uso e della coscienza linguistica contemporanea" (p. V).

sono stati conteggiati automaticamente, per separare le frequenze degli omonimi assoluti, si è proceduto all'analisi manuale di campioni ridotti, in modo da poter giungere ad una stima abbastanza precisa sulle frequenze dei vari significati.[12] Gli autori prevedono di pubblicare una parte dei dati di frequenza in un volume dedicato (p. XV).
L'altro lavoro di particolare importanza è il grande dizionario della lingua ungherese, uscito nel 2006 a cura di Ittzés (*A magyar nyelv nagyszótára*, Budapest, Akadémiai Kiadó). Esso si basa sul corpus storico della lingua ungherese (*Magyar Történeti Szövegtár*) con 27 milioni di occorrenze, disponibile sul sito www.nytud.hu/hhc/ (cfr. 3.3). Dal corpus è stata estratta una lista di frequenza dei lemmi, contenente anche la data della prima e l'ultima attestazione. Ogni entrata è seguita (oltre alla marca grammaticale e un codice che rinvia al paradigma morfologico) da un cerchietto, destinato ad indicare l'appartenenza ad una delle sei fasce di frequenza assoluta, da 0 a 100.000. Il cerchio è diviso in sei parti, e la misura in cui il cerchio è colorato di nero, mostra la frequenza d'occorrenza del lemma nel corpus storico di riferimento. Il cerchio vuoto segna che il lemma non figura nel corpus, quello pieno indica una frequenza d'occorrenza superiore a 100.000. Se un sesto del cerchio è riempito, ciò indica che il lemma registra 1-9 occorrenze, con due sesti si segnano i lemmi aventi 10-99 occorrenze, con tre sesti i lemmi appartenenti alla fascia 100-999, con quattro sesti la fascia 1.000-9.999, con cinque sesti si indicano i lemmi aventi 10.000-99.999 occorrenze. Analogamente a quanto visto nel caso del dizionario enciclopedico della lingua ungherese, per gli omonimi s'indica separatamente la frequenza di ciascun'entrata, mentre nel caso delle parole polisemiche non si precisa a quale significato si riferisce l'indicazione di frequenza. Per stabilire le frequenze degli omonimi si è ricorso ad un programma di analisi morfologica, Humor, sviluppato per l'annotazione del corpus.
Tutti questi esempi servono ad evidenziare che, analogamente a quanto abbiamo visto nella sezione dedicata alla selezione dei lemmi, la prassi lessicografica moderna non può fare a meno di dati statistici obiettivi, derivanti dall'analisi di larghi corpora che attestino gli usi reali della lingua da parte dei parlanti. Solo in questo modo è possibile garantire l'attendibilità dei dati di frequenza e delle marche d'uso proposti dal dizionario.

4.1.4 Accezioni, collocazioni ed espressioni polirematiche

I compiti fin qui elencati non esauriscono il lavoro del lessicografo che voglia sfruttare i risultati ricavati dall'analisi delle frequenze. Una questione importante è legata alla polisemia che genera entro uno stesso segno numerose accezioni o famiglie di sensi. La polisemia penetra gran parte del lessico di una lingua, e in particolare le parole d'uso più frequente. Nel cap. 1 abbiamo visto, che gli studi statistici confermano la correlazione tra frequenza di una parola e sviluppo di sensi entro un medesimo significato.[13] Nel GRADIT ad esempio, sono registrate oltre 50.000 parole con più di un'accezione (numero che corrisponde proprio al vocabolario comune), mentre il numero di accezioni e sotto-accezioni arriva ai massimi con verbi generici come *prendere* (48) e *fare* (43). È stato anche rilevato che non tutte le accezioni una parola sono ugualmente frequenti. L'uso di dati statistici, ricavati da corpora, serve anche ad ordinare e

[12] In un campione selezionato in modo casuale, contenente tante volte 100 frasi quanti significati l'omonimo aveva hanno separato manualmente i vari significati, per stimare con una precisione del 10% l'effettiva frequenza dei lemmi.
[13] Per significato si intende il valore complessivo che una parola ha in una data lingua, mentre per senso intendiamo la determinazione particolare che una parola ha in una particolare situazione (cfr. De Mauro 2005: 81).

rintracciare usi nuovi e diversi da quelli tradizionalmente riferiti. I corpora servono proprio nella fase di studio e analisi dei significati, dei sensi, delle collocazioni e delle locuzioni idiomatiche complesse che caratterizzano ogni lingua.
Consultando la prefazione di alcuni dizionari d'uso, possiamo rilevare che il Garzanti (2005) e il Devoto-Oli (2004-2005) non esplicitano l'ordine con cui le diverse accezioni sono presentate. Lo Zingarelli (2006) dà una spiegazione abbastanza vaga (p. 9):

> "L'ordinamento delle accezioni all'interno della voce segue di solito un criterio che vuol essere logico e storico al tempo stesso: precedono cioè le definizioni dei significati propri od originariamente più in uso e seguono quelle dei significati figurati, estensivi, specifici ecc."

Il DISC invece, adottando una soluzione pionieristica, già nel 1997 tiene a precisare nella prefazione:

> "L'ordine delle accezioni rispetta la frequenza e l'importanza di queste, ma viene modificato quando l'accezione più comune o importante è, in modo ancora visibile, uno sviluppo figurato di altra meno comune."

Entrambi gli esempi testimoniano di una qualche importanza attribuita alla frequenza nell'ordinamento delle accezioni.
Nel *Grande dizionario italiano dell'uso* anche per la registrazione e per la scansione delle accezioni sono stati dichiarati i criteri: le accezioni sono state ordinate secondo un criterio storico-cronologico, ma in ogni caso, in cui l'accezione avvertita come basilare si sarebbe trovata in fondo alla struttura della voce, il criterio storico è stato abbandonato a favore di un ordinamento che colloca ai primi posti le accezioni più frequenti nell'uso (De Mauro 2005: 81-82).
Attenzione alla frequenza d'uso sembra mostrare anche il grande dizionario della lingua ungherese (*A magyar nyelv nagyszótára*). L'ordinamento delle accezioni segue un criterio logico, più raramente cronologico, cercando ogni volta di dare precedenza alle accezioni più frequenti, più neutrali e diffusi (cfr. Ittzés 2006: 35 I. vol.).
In alcuni dizionari basati su corpus, come quello del progetto COBUILD (*Collins Cobuild English Dictionary for Advanced Learners*, 2003), le singole accezioni sono elencate in ordine di frequenza, ricavata dal corpus.[14] Questo tipo di procedimento, proprio perché basato su un criterio quantitativo omogeneo, offre vantaggi decisivi rispetto ai dizionari tradizionali. Chiari (2007: 95) nota che il *Cobuild* è stato il primo dizionario inglese a citare come accezione di senso principale del termine *gay*, il significato di 'omosessuale' prima del senso 'vitale e allegro'. Per riportare un altro esempio, la versione elettronica del *Longman Dictionary of Contemporary English* (2005) permette la ricerca anche selezionando le accezioni più frequenti (scritti o parlati, a scelta) di una parola. In altri casi le informazioni sulla frequenza e sulla dispersione possono anche non essere date all'utente, ma, nei moderni dizionari rimangono alla base della progettazione stessa della voce del lemma e delle sue gerarchie interne.
Dal punto di vista lessicografico è d'importanza fondamentale la sistematica combinabilità degli elementi lessicali, le combinazioni preferenziali e il tipo di contesto che richiedono. In alcuni casi le parole possono occorrere con gran frequenza in molti contesti diversi, in altri le parole occorrono per lo più in un determinato complesso di elementi. Vi sono dunque diversi gradi d'intensità nel legame tra due o più parole che co-occorrono in un testo. Le combinazioni di parole relativamente libere, i cui componenti però co-occorrono con una certa frequenza, si chiamano collocazioni. Tipiche

[14] Per l'inglese esistono una serie di dizionari elettronici che permettono di effettuare ricerche avanzate su categoria grammaticale, frequenza, derivati e basi, sinonimi e note d'uso, pronuncia. Si vedano il CDE *Cambridge International Dictionary of English* (2001), il *Cambridge Advanced Learner's Dictionary* (2003) e l'OED *Oxford English Dictionary* (2004).

collocazioni sono ad esempio: *compilare un modulo, prescrivere una ricetta* ecc. Individuare le collocazioni è importante soprattutto in un dizionario bilingue, in cui i traducenti dei singoli elementi non sempre trasmettono il significato dell'espressione.
Le polirematiche, di contro, sono espressioni composte di più di una parola, che tuttavia si comportano semanticamente e spesso anche morfologicamente come un solo lessema, p. es. *scala mobile, essere al verde* ecc.
Le più diffuse procedure automatiche usate per estrarre le collocazioni e le polirematiche sono misure di associazione statistica, come il *mutual information* (informazione reciproca) e lo *Z-score* (punteggio Z).[15] Il primo sistema compara la co-occorrenza effettiva di una coppia di parole con il valore di co-occorrenza che le due parole avrebbero se si trovassero vicine casualmente. Lo *Z-score*, invece, prende come unità di confronto la frequenza d'occorrenza di una parola, comparata alla frequenza attesa di co-occorrenza con altre parole, divisa per la deviazione standard della parola co-occorrente (Chiari 2007: 77-78).
Quanto alle collocazioni, è un dato di fatto che in un dizionario è impossibile registrare tutte le collocazioni di un lemma, pertanto il lessicografo è obbligato ad effettuare una selezione. Vediamo ora come la prassi lessicografica attuale affronta la questione. Esaminando cinque dizionari d'uso (Devoto-Oli 2004-2005, DISC 1997, De Mauro 2000, Garzanti 2005, Zingarelli 2006), si trovano complessivamente 59 collocazioni diverse con l'aggettivo attributivo *fresco*, di cui riportiamo alcuni nella tabella sottostante.

	Zingarelli 2006 (25)	DISC 1997 (30)	De Mauro 2000 (32)	Devoto-Oli 2004-2005 (20)	Garzanti 2005 (24)
acqua	x			x	
aria		x	x		x
atleta		x	x		
bevanda					x
bibita		x	x		
brezza			x		
caffé	x				
camera					x
carnagione		x			
carne	x	x	x	x	x

Collocazioni del lemma *fresco* in alcuni dizionari d'uso.

Il numero delle collocazioni contenute nei singoli dizionari varia da 20 (Devoto-Oli) a 32 (De Mauro), mentre solo 7 (*carne, fronte, frutta, latte, pane, pesce, uovo*) figurano in tutti e cinque i dizionari. Restano 28 contesti (ca. la metà del totale) che sono presenti esclusivamente in uno dei dizionari. De Mauro ad esempio, propone *zona fresca*, *piselli freschi*, il DISC invece riporta *carnagione fresca* e *informazioni fresche*. È sorprendente che unicamente il Devoto-Oli riporta *pasta* tra le collocazioni di *fresco*, sebbene la *pasta fresca* appartenga strettamente alla realtà italiana.
Sembra quindi, che la selezione delle collocazioni sia poco sistematica, o addirittura casuale. La lessicografia basata su corpora, invece, può giovarsi in modo particolare della possibilità di estrazione e selezione automatica delle collocazioni dai testi. Elencare solo quei contesti d'uso che hanno una certa frequenza nel corpus di riferimento del dizionario, osservando nella compilazione dell'elenco l'ordine di frequenza, permette di evitare le collocazioni banali e casuali e trovare quelle davvero caratteristiche. Questa soluzione indurrebbe il lessicografo ad una registrazione sistematica, e, nello stesso

[15] Cfr. De Mauro (2006: 19): per la collocazione dei singoli lemmi occorre tener presente, fra l'altro, "la frequenza delle occorrenze, la distribuzione nei testi, insomma la stratificazione statistica del lessico e di ciascuna parola in corpora rappresentativi della lingua."

tempo, permetterebbe all'utente di interpretare in modo veloce e chiaro il lemma in questione.
Nei dizionari COBUILD una delle maggiori novità, dovute all'utilizzazione del corpus, è appunto, l'estrazione delle collocazioni più frequenti e più autentiche. Non solo i dizionari contengono le collocazioni registrate con la maggiore frequenza nel corpus, ma il sito della casa editrice (http://www.collins.co.uk/Corpus/CorpusSearch.aspx) offre un accesso diretto alla versione demo del *Wordbank*, il corpus da cui il dizionario è stato tratto, in cui è possibile effettuare una ricerca sulle possibili collocazione di una parola. Dopo aver digitato la parola ricercata, si può decidere se la misura di co-occorrenza statistica deve essere calcolata in base al *mutual information* o lo *Z-score* (cfr. sopra). Con la parola *white*, ad esempio, è saliente che la collocazione ben più frequente è *house*, equivalente di *Casa bianca*, con una frequenza assoluta di 22.752 nell'intero corpus di 56 milioni occorrenze.
Passiamo ora alle polirematiche. Ricordiamo, che dal punto di vista semantico la caratteristica più importante delle polirematiche è, appunto, l'imprevedibilità, ovvero la non deducibilità del suo significato a partire dal valore dei singoli componenti. Pertanto, una registrazione coerente di queste espressioni è di particolare importanza. Ciononostante la premessa del Devoto-Oli (2004-2005) e del Garzanti (2005) non dà alcun riferimento sul trattamento delle espressioni complesse. Lo Zingarelli e il DISC precisano nelle loro prefazioni che le locuzioni idiomatiche sono riportate all'interno di ciascun'accezione, precedute da una barra verticale. Il DISC aggiunge ancora:

> "Trovano posto tra le locuzioni comuni, ma sono evidenziate con carattere neretto corsivo, le unità polirematiche di contenuto o con valore sostantivale non messe a lemma, ma che si possono considerare pressoché lessicalizzate."

In realtà, il GRADIT è stato il primo dizionario a dare autonomia ai lessemi complessi. Esso include, accanto alle 250.000 entrate principali, oltre 130.000 sottolemmi polirematici, cui il settimo volume di supplemento aggiunge 3.400 nuovi lemmi semplici e ca. 500 polirematiche.[16] Analogamente il Dizionario italiano per il terzo millennio (De Mauro 2000) registra 130.000 lemmi costituiti da una sola parola e 30.000 espressioni polirematiche.
I lemmi che figurano in locuzioni polirematiche dispongono, dopo la fine della trattazione come lemma autonomo, di una sezione dedicata alle polirematiche, mentre un piccolo gruppo di espressioni polirematiche, per lo più locuzioni latine o di lingua straniera (marcate ES), le cui parole costitutive non sono incluse come lemma semplice, formano un lemma autonomo. Le polirematiche del primo tipo sono lemmatizzate sotto tutti gli elementi che le compongono (tranne preposizioni, articoli e le congiunzioni *e* ed *o*), la definizione è svolta sotto il primo elemento pieno della polirematica, mentre sotto gli altri compare un rinvio al lemma sotto il quale è svolta la definizione. Le polirematiche con definizione sono ca. 63.000 (oltre 1.753 date come lemmi autonomi), quelle con rinvio oltre 67.000.
In De Mauro (2005: 88-91) sono esplicitati i tre criteri in base ai quali le unità polirematiche sono state individuate:

a) l'esistenza di uno specifico sovrappiù semantico, cioè la non ricostruibilità del loro significato in base alla semplice somma dei significati dei singoli componenti monotematici;

[16] Come risulta chiaramente dal GRADIT, le polirematiche sono in espansione nell'italiano contemporaneo a causa di fattori diversi, come i linguaggi tecnico-scientifici, i forestierismi, i regionalismi (it. sett. *levati su* 'alzati') e anche il parlato, che preferisce forme semanticamente trasparenti.

b) la più o meno forte cristallizzazione lessicale e sintattica, ovvero che tendenzialmente non ammettono variazioni lessicali e strutturali interne;
c) la presenza significativa in linguaggi tecnico-specialistici.

Quest'ultime, ovvero le polirematiche tecnico-specialistiche costituiscono circa l'87% di tutte le polirematiche definite.
La prassi lessicografica ungherese, tuttavia, sembra seguire la soluzione tradizionale, ovvero di registrare le polirematiche sotto il lemma che ne costituisce il componente semanticamente più importante. Nel grande dizionario della lingua ungherese (*A magyar nyelv nagyszótára*, 2006) i lemmi composti da più di un elemento si limitano ai verbi riflessivi, costituiti dal verbo e dal pronome riflessivo: *megmagkacsolja magát* (cfr. Ittzés 2006: 41-43 vol. 1), mentre le parole che figurano esclusivamente in polirematiche sono registrate nel nominativo singolare, con rinvio al lemma sotto il quale l'espressione idiomatica è registrata. All'interno dei lemmi, ogni espressione idiomatica è esemplificata con frasi estratte dal corpus. Alla fine di ciascun lemma si rinvia a tutte le espressioni polirematiche in cui il lemma figura.
Il dizionario enciclopedico della lingua ungherese (*Magyar értelmező kéziszótár*, 2003) registra sistematicamente le espressioni polirematiche dotate di un sovrappiù semantico: sotto le 70.000 entrate si trovano ca. 15.000 espressioni idiomatiche.
Pare quindi che i nuovi dizionari tendano a dare autonomia sia a prefissi e suffissi sia alle polirematiche, innovando in questo la nostra tradizione lessicografica.

Abbiamo visto che l'ordinamento delle accezioni, la selezione delle collocazioni e delle espressioni polirematiche sono operazioni delicate che ogni volta dovrebbero basarsi sugli usi di una comunità linguistica. Per effettuare scelte oggettive ed omogenee, c'è bisogno di una gran quantità di dati autentici, il che consente di produrre un dizionario più fedele alla lingua in uso in un determinato periodo. I dizionari basati su corpora e sui dati di frequenza che se ne possono estrarre, rappresentano una "rivoluzione lessicografica", con le parole della Chiari (2007: 96), sia dal punto di vista della pianificazione del lavoro lessicografico, sia dal punto di vista dell'utente. I dati di frequenza ricavati dai corpora permettono di effettuare scelte coerenti in vari campi: nella selezione delle voci da includere o escludere, nella compilazione delle definizioni con un lessico controllato ed esempi autentici, mediante l'inclusione di marche d'uso di tipo statistico, nella selezione e ordinamento delle collocazioni e delle espressioni idiomatiche, nell'ordinamento delle accezioni di senso, nella selezione delle voci tecniche. D'altra parte l'utente ha la possibilità di fruire di questa maggiore aderenza agli usi, e, in alcuni casi, anche di accedere a materiale non direttamente elaborato nel dizionario per migliorare la sua comprensione degli usi.
Notiamo, infine, che l'estrazione delle collocazioni e delle polirematiche non serve solo in lessicografia, ma anche in vari altri campi: nella traduzione automatica serve per individuare traducenti cristallizzati ed arricchire le banche dati terminologiche, nel trattamento automatico del linguaggio, per operare corrette analisi sintattiche, ma anche nell'*information retrival*, nella disambiguazione dei sensi di una parola e nel riconoscimento e nella sintesi del parlato, come vedremo di seguito.

4.2 Stilometria e scrittura di testi di difficoltà controllata

I primi testi analizzati dal punto di vista statistico sono state opere letterarie: le prime concordanze furono applicate a studi biblici e risalgono a tempi remoti. All'analisi delle concordanze si sono dedicati numerosi studiosi con obiettivi diversi: per individuare le peculiarità delle scelte lessicali di un autore in un testo; per determinare le caratteristiche linguistiche (lessicali, morfo-sintattiche) che differenziano testi di tipologie diverse; per predisporre strumenti di statistica che agevolano l'attribuzione di un'opera ad un autore o ad un altro (cfr. De Mauro e Chiari 2005: 373). L'iniziatore dell'analisi lessico-statistica di testi italiani fu padre Roberto Busa, che nel 1949 avviò la preparazione delle concordanze dell'opera di S. Tommaso d'Aquino con gli elaboratori a schede perforate dell'IBM. Nell'*Index Thomisticus,* la sintesi del lessico di Tommaso occupò i primi dieci volumi (su 56) per complessive 11.500 pagine. Il successo fu immediato: dagli anni '50 e '60 tutte le concordanze sono automatizzate, qui ricordiamo solo i lavori di Rosiello (1965) ed Alinei (1968-1978). In ambito internazionale, pionieristico fu il lavoro di Guiraud (1954). Il cuore della sua teoria è che i valori numerici degli aspetti quantificabili del testo informano sulle scelte lessicali e grammaticali dell'autore. È centrale l'idea di suddividere le parole usate in fasce di frequenza ben delimitate; per ricavare le specificità di un testo, basta esaminare in che misura esso devia dalla media trovata in altri testi.
In un secondo periodo è nata l'esigenza di trovare strumenti per la misurazione degli indicatori stilistici o stilemi in un testo, normalmente in funzione di una specie di norma, rispetto al quale le eventuali deviazioni possono essere valutate. Gli studi di stilometria hanno, infatti, l'obiettivo di creare strumenti per la diagnosi delle differenze individuali negli stili verbali. L'indice più famoso è il rapporto tipo-replica che mette in rapporto il numero delle occorrenze (repliche) con il numero di parole differenti (tipi). Si prevede un rapporto di proporzionalità diretta tra la variazione lessicale e la quantità di tipi: maggiore è il numero di tipi, maggiore è la varietà lessicale. Fondamentali sono stati su quest'argomento i lavori di Estoup (1907). Tra gli aspetti statistici del linguaggio oggi più studiati è l'estensione (o ricchezza) del vocabolario, ampiamente analizzata fin dagli anni Trenta e Quaranta, quando appaiono i primi dizionari di frequenza e di base.[17]
Un importante campo d'applicazione dei dizionari fondamentali, sviluppatosi dagli studi di stilometria, è la scrittura di testi di difficoltà controllata. Il primo a dichiarare che la leggibilità di un testo è un concetto misurabile, è stato lo studioso americano Rudolf Flesch che nel 1948 ha preparato il suo indice, tarato sull'inglese americano. L'indice calcola la leggibilità di un testo dal numero medio di parole per frase e dal numero di sillabe per 100 parole, ed è compreso tra 0 (leggibilità minima) e 100 (leggibilità massima).[18] La formula di Flesch deve la sua diffusione proprio alla sua semplicità: fu, infatti, uno degli strumenti più utilizzati nei primi studi sulla leggibilità dedicati a testi in lingua italiana e dalla metà degli anni Settanta ad oggi la formula di Flesch, nell'adattamento di Vacca del 1972, è stata usata in varie sedi e diversi obiettivi (cfr. Thornton 1992, Piemontese 2005).[19]

[17] Nell'editoria moderna le concordanze e gli indici delle frequenze sono strumenti indispensabili in varie fasi del lavoro filologico, come in quella della collazione dei testimoni, dell'individuazione delle varianti, della definizione delle parentele tra codici attestanti una tradizione, fino alla realizzazione dell'edizione critica in senso stretto. Un programma di analisi testuale quantitativa, capace di produrre concordanze, liste di frequenza e statistiche, che al contempo tratta correttamente particolarità editoriali è IntraText, sviluppato da Èulogos SpA (cfr. Mastidoro e Amazzoni 2005: 417-429).
[18] La formula di Flesch per la lingua inglese: *Ease of Reading* = 206,835 - 0,864 S - 1,015 W, in cui S = sillabe di 100 parole e W = media di parole per frase. La formula di Flesch adattata per la lingua italiana da Vacca: Facilità di lettura = 206- 0,65 S - W (Piemontese 2005: 388-389).
[19] Va menzionata la collana dei Libri di base degli Editori Riuniti, avviata nel 1980 sotto la direzione di Tullio De Mauro. L'obiettivo dichiarato era di soddisfare le esigenze culturali di un pubblico molto vasto, con almeno la licenza di scuola media. Al fine di produrre testi di alta comprensione nacque il lavoro di

In seguito all'applicazione sistematica della formula emersero una serie di critiche sulla sua validità. Da un lato la formula è stata progettata per l'inglese ed è, quindi, tarata sulla struttura morfologica e sillabica di questa lingua; l'altro inconveniente è rappresentato dal problema del conteggio delle sillabe. Questo tipo di calcolo si mostra particolarmente complesso nell'ambito della lingua italiana, poiché esso non è completamente formalizzabile mediante regole di portata generale, se non ricorrendo a stime di tipo statistico. Nel campo della sillabazione le difficoltà maggiori sono prodotte dall'accentazione e dalla presenza di dittonghi. Si è anche scoperto che ci sono aspetti importanti dei testi che influiscono ampiamente sulla loro comprensibilità, ma non sono considerati dalla formula. Per stabilire se un brano ha una buona leggibilità, infatti, occorre tener presente anche la diversità lessicale e il tipo di parole usate.
Alla metà degli anni Ottanta, un gruppo di ricercatori e studenti dell'Università degli Studi di Roma "La Sapienza" ha sviluppato un nuovo indice di leggibilità, tarato direttamente sulla lingua italiana. Sulla base dei risultati delle ricerche del Gruppo universitario linguistico e pedagogico (GULP) è stata ideata la formula *Gulpease*. Per calcolare l'indice si devono considerare la lunghezza delle parole in lettere (e non in sillabe) e la lunghezza delle frasi in parole. Come la formula di Flesch, anche il *Gulpease* presenta una scala di valori compresa fra 1 e 100, ma diversamente dalla formula di Flesch, il valore dell'indice *Gulpease* è proiettato su una scala di leggibilità in rapporto al grado di scolarizzazione del lettore: elementare, media e superiore (cfr. Piemontese 2005: 393). Questa scala permette di valutare la difficoltà di un testo rispetto a tre popolazioni di riferimento. Per esempio, un testo con indice *Gulpease* 60 è molto difficile per chi ha la licenza elementare, difficile per chi ha la licenza media, facile per chi ha un diploma superiore.
Agli sviluppi successivi della formula *Gulpease* e alla messa a punto di un lemmatizzatore che utilizza il *Vocabolario di base* (De Mauro 1980, cfr. 2.2.15) sono state dedicate le ricerche di due tesi di laurea: l'una di M. Amizzoni (1991) e l'altra di N. Mastidoro (1991). I due studi hanno permesso di realizzare in Èulogos SLI una versione informatizzata della formula *Gulpease*, e di creare un programma d'analisi lessicale di un testo a partire dal *Vocabolario di base* e di memorizzazione di altri lemmi in vocabolari logici in modo da arrivare all'analisi anche di parole di linguaggi settoriali (cfr. Mastidoro 1992). Da questi lavori mirati alla realizzazione di un algoritmo capace di calcolare la leggibilità dei testi e di analizzarne il lessico, è nato l'attuale analizzatore integrato, della Éulogos *Censor* (http://www.eulogos.net/it/censor/). Il programma analizza la leggibilità di un testo secondo l'indice *Gulpease*, inoltre confronta le parole del testo con i vocaboli del VdB, fornendo dati utili per individuare le parole da sostituire o da spiegare. L'analisi condotta da *Censor* restituisce oltre i dati di leggibilità, dati statistici e una riproduzione del testo in cui ogni occorrenza è presentata con caratteri diversi, secondo il livello di VdB. Il servizio funziona per posta elettronica ed è gratuito.
Il calcolo delle forme di leggibilità e il confronto sistematico con lessici di base, permette di scegliere in modo più efficace i testi da inserire nelle pubblicazioni didattiche, negli esami di certificazione e non solo, come vedremo di seguito.

4.2.2 Il mensile *due parole*

ricerca sulle liste di frequenza, che portò alla pubblicazione del *Vocabolario di base,* in appendice a *Guida all'uso delle parole* (v. sez. 2.2.15). La leggibilità dei testi è stata verificata per mezzo della formula Flesch, adattata da Vacca.

Un esempio di applicazione sistematica e professionale dell'indice *Gulpease* e dei criteri della scrittura controllata è *due parole* (www.dueparole.it), "mensile di facile lettura". La rivista è redatta dal Gruppo H, costituito da collaboratori e studenti della cattedra di Filosofia del linguaggio dell'Università di Roma "La Sapienza", guidati da Maria Emanuela Piemontese e Tullio De Mauro.[20]

Accanto ai destinatari originari di *due parole,* le persone con forme lievi di ritardo mentale che costituiscono un 2,2% della popolazione (Piemontese 1994: 275), troviamo studenti "normodotati" ma poco avvezzi alla lettura, ragazzi dislessici; studenti stranieri che studiano l'italiano in corsi elementari in Italia e all'estero; i frequentanti di corsi di alfabetizzazione per adulti; stranieri immigrati di recente in Italia; anziani; persone sorde e sordomute. Del gruppo dei cosiddetti "lettori dimenticati" fanno parte tutti coloro che, avendo problemi specifici di lettura e di comprensione, hanno bisogno di un'attenzione supplementare da parte di chi fa comunicazione, informazione, formazione.[21]

Leggibilità e comprensibilità dei testi, infatti, nascono dalla combinazione d'essenzialità, chiarezza e precisione. Pertanto, uno dei criteri principali di redazione è stato l'uso di un lessico specifico, contenente 2.000 entrate, selezionate in base ad alcuni dizionari di base e di frequenza dell'italiano, in particolare il LIF, il VELI, il LIP e il VdB. Quest'ultimo è tutt'oggi lo strumento fondamentale per controllare e migliorare la leggibilità di un testo secondo criteri scientifici. I redattori hanno delimitato le 2.000 parole del lessico di *due parole* a partire dal vocabolario fondamentale del VdB, aggiungendovi ca. 400 termini d'alto uso, 130 parole di alta disponibilità e 180 termini non appartenenti al VdB. Il *Lessico di Due Parole* (LDP) è stato pubblicato in appendice a Piemontese (1998).

L'uso di un lessico ben preciso non vuol dire però che un testo ad alta leggibilità sia chiuso a nuove parole, ma significa che quando si usa un lemma non di base e soprattutto quando esso è importante per capire il testo, bisogna spiegarne il significato, usando nella spiegazione solo lemmi presenti nel VdB.[22]

> "Un discorso costruito tenendo d'occhio le parole dell'elenco, le 'parole di base', può anche contenere parecchie parole fuori elenco, purché spiegate con le parole di base" (De Mauro 1980: 138).

Un testo altamente comprensibile non ha solo questa caratteristica: le frasi devono essere brevi, ma ben collegate fra loro sia dal punto di vista linguistico che concettuale; sono da evitare le astrazioni e gli usi metaforici della lingua. Inoltre, la redazione applica anche altri criteri di leggibilità, attinenti alla grafica (caratteri grandi, le illustrazioni devono essere esplicative del testo senza interromperlo, ecc.) e al modo di disporre i contenuti (riquadri di spiegazione, testo a nuova riga per ogni frase, ecc.).

---

[20] I dati di questo capitolo sono stati ricavati da una conversazione con la Prof. Piemontese. Per ulteriori informazioni cfr. Piemontese (1998), Piemontese e Saponaro Cioffi (1994).

[21] Parallelamente alla pubblicazione di *due parole*, alla fine degli anni Ottanta e in particolare agli inizi degli anni Novanta sono usciti i primi giornali di facile lettura in Europa: dal 1985, in Belgio è pubblicato in lingua fiamminga il giornale *Wablieft* "Che cosa dici?"*, disponibile* anche on line: www.wablieft.be. Nel 1986 per iniziativa del governo svedese è uscito *8 Sidor. Nättlästa Nyheter* "8 pagine. Notizie di facile lettura", settimanale destinato a persone adulte con danni intellettuali lievi (http://8sidor.lattlast.se). Dal 1987 in Danimarca esce il giornale *På Let Dansk* "Il danese facile da leggere" (www.paaletdansk.dk). Nel 1990 in Finlandia è uscito il quindicinale *Selko-Uutiset* "Notizie chiare" anch'esso per iniziativa governativa (http://papunet.net/selko). Date le caratteristiche sociolinguistiche della Finlandia, fu realizzata una versione svedese dello stesso giornale dal titolo *LL-Bladet* "Giornale di facile lettura" (http://www.papunet.net/ll/ll-bladet). Dal 1990, in Norvegia alcuni giornalisti pubblicano il settimanale *Klar Tale* "Parlare chiaro" (www.klartale.no). Dal 1990, l'associazione belga *Lire et Écrire* e la Comunità francese realizzarono nel Belgio francofono il mensile *L'essentiel* (http://cours.funoc.be/essentiel/).

[22] Sulle informazioni scientifiche in *due parole* cfr. Piemontese (1994).

Il mensile aveva uno scopo triplice: si volevano definire e sperimentare tecniche di redazione di testi di alta leggibilità e comprensibilità; s'intendevano produrre testi scritti adeguati ai bisogni di persone con difficoltà di lettura e di comprensione; ed era presente l'obiettivo di insegnare agli studenti e studentesse della cattedra di Filosofia del Linguaggio tecniche di analisi e di scrittura di testi di difficoltà controllata.
In seguito ad un lungo periodo di sperimentazione e di formazione dei redattori, tutti volontari, il primo numero è uscito nell'aprile del 1989, ed ha avuto un'accoglienza positiva, perché dopo quasi un anno il periodico aveva 630 abbonati spontanei. Il primo numero è uscito in formato tabloid con 8 pagine che presentavano, oltre alla notizia più importante del mese in prima pagina, notizie di spettacoli, politica italiana ed estera, sport e cultura. La tiratura era di 3.000 copie mensili, inviate gratuitamente a insegnanti, operatori socio-sanitari, genitori di disabili, persone sorde, ma anche a molti giovani con problemi di comprensione. Tra gli abbonati, diventati alla fine del 1994 un migliaio, vi troviamo scuole medie ed elementari, centri di formazione, associazioni di volontari, cittadini privati, docenti ecc. Tuttavia si è notata la disattenzione a livello istituzionale per *due parole*, ad eccezione dell'Università di Roma "La Sapienza" che inizialmente lo finanziava e forniva l'appoggio istituzionale. La pubblicazione del mensile, consentita da un piccolo finanziamento dell'università e dagli abbonamenti, è il risultato dell'impegno volontario e gratuito dei redattori. Dopo otto anni di pubblicazione a stampa, *due parole* è stato sospeso nel 1997. Le pubblicazioni si sono interrotte, ma la redazione non ha mai smesso di lavorare per riprendere il giornale in altre forme. Alcuni redattori di *due parole* fondarono nel 1998 "Parlar chiaro. Associazione per la semplificazione della comunicazione di interesse pubblico". Obiettivo dell'Associazione è stato quello di contribuire alla diffusione della cultura della chiarezza e della semplicità linguistica in Italia. L'Associazione nel dicembre 2002 ha ripreso la pubblicazione del periodico, anche in formato elettronico ad accesso gratuito. L'ultimo numero ad oggi disponibile però è uscito nel maggio 2006, sembra quindi, che nonostante l'enorme successo e l'entusiasmo dei redattori, senza il sostegno finanziario delle autorità, non sia stato possibile portare avanti il progetto.

4.2.3 *Codice di stile*

Parlando dei campi d'applicazione delle tecniche di semplificazione linguistica, Piemontese (1998: 229) intuì gli esiti futuri di tali iniziative.[23]

> "l"idea che anche in Italia si possa arrivare a imporre per legge il rispetto di standard minimi di leggibilità e comprensibilità, almeno per certi tipi di testo, come quelli di interesse pubblico […] non appare più del tutto peregrina."

Frutto del lavoro di una équipe di linguisti, guidati da Sabino Cassese, il *Codice di stile* è stato pubblicato nel 1993 su incarico del Dipartimento della Funzione Pubblica della Presidenza del Consiglio dei Ministri. È la prima pubblicazione di questo genere prodotta in Italia da organi dello Stato, e prende esempio da testi analoghi già circolanti in altri Paesi (come il *Manual de Estilo del Lenguaje Administrativo* in Spagna nel 1990). L'opera scaturisce dalla legge 7 agosto 1990, n. 241, intitolato "Nuove norme in materia di procedimento amministrativo e di diritto di accesso ai documenti amministrativi". La

[23] De Mauro (1994: 114) addita che in 37 dei 51 Stati degli USA le leggi prevedono sanzioni per tutti gli addetti a uffici pubblici e perfino per gli stessi legislatori. A tutti è stato richiesto di rivolgersi al pubblico rispettando standard minimi di leggibilità, fissati a partire dalle scale elaborate da Flesch. In alcuni Stati disposizioni del genere si vanno estendendo anche a scritture private, come i contratti assicurativi.

legge prevede la semplificazione dell'azione amministrativa e il diritto d'accesso dei cittadini ai documenti amministrativi. Per adempire tali requisiti non basta semplificare i modi dell'azione amministrativa, per una reale trasparenza e accessibilità, ovvero per una reale democraticità è indispensabile semplificare e unificare il linguaggio con cui tale agire si esprime. Il *Codice di stile* rappresenta il primo passo di un progetto, finalizzato alla redazione di un vero manuale del linguaggio amministrativo e alla messa a punto di percorsi formativi specifici per il personale delle pubbliche amministrazioni. Scopo del *Codice di stile* è dare alcuni elementi di base per sensibilizzare la Pubblica Amministrazione sul problema della leggibilità, fornendo anche indicazioni pratiche ed esempi.
Nella parte prima del *Codice di stile* è illustrato lo stretto rapporto tra diritti dei cittadini e chiarezza dei testi, attraverso l'analisi del linguaggio delle leggi e delle Pubbliche Amministrazioni. Quest'aspetto del rapporto Stato–cittadino costituisce, secondo gli autori, per l'Italia un primato negativo. Nella parte seconda segue l'ambito d'applicazione del *Codice* e le concrete raccomandazioni per la semplificazione del linguaggio e per un uso non sessista e discriminatorio, infine esempi di moduli e documenti pubblici, riscritti secondo le regole del *Codice*. In appendice si riportano alcuni dati sui livelli di alfabetizzazione in Italia, è inoltre allegata la decima edizione del *Vocabolario di base* di De Mauro *et al.* (1989), quale strumento di riferimento per la semplificazione lessicale e per il controllo della leggibilità dei testi.
Da quest'iniziativa nasce nel 2003 il *Manuale di scrittura amministrativa*, curato da F. Franceschini, pubblicato dall'Agenzia delle Entrate in collaborazione con il Dipartimento di Studi Italianistici dell'Università di Pisa, scaricabile interamente dal sito dell'Agenzia. Il volume sintetizza i contenuti teorici ed esempi significativi dei testi riscritti durante una serie di corsi organizzati presso l'Agenzia delle Entrate nel periodo 2001-2002 dal titolo "Scrivere nella pubblica amministrazione".
Considerando la situazione in Ungheria, non possiamo non ripetere le parole di Piemontese (1994: 277).

> "la disattenzione totale per questo ordine di problemi alla comunicazione mette in discussione la sostanziale democraticità dell'ordinamento di un Paese che, contraddittoriamente, da una parte riconosce e proclama i diritti di una cospicua fetta di cittadini, d'altra parte, poi, non fa tutto il necessario o il possibile perché tali diritti siano pienamente rispettati."

Sorprende quindi, che mentre per l'inglese è stata elaborata una dozzina di formule di leggibilità diverse, ancora nel primo decennio del terzo millennio, non esiste una formula sviluppata appositamente per la lingua ungherese, tanto meno esistono pubblicazioni pensate per un pubblico con difficoltà di lettura, per non parlare della semplificazione dei testi amministrativi.

4.3 Applicazioni informatiche

La relazione tra statistica linguistica e linguistica computazionale è sempre stata molto stretta. Lo studio delle caratteristiche quantitative delle lingue naturali ha fornito strumenti fondamentali per lo sviluppo delle applicazioni della linguistica computazionale. L'utilizzo di alcuni metodi di tipo statistico ai dati linguistici ha permesso la creazione di strumenti quali lo *Statistical Natural Language Processing,* i sistemi di estrazione di informazioni, la traduzione automatica, ecc. Si tratta di

applicazioni *dictionary-based* (Chiari 2007: 85), in cui il termine è usato per descrivere repertori contenenti liste di parole che servono a svolgere determinate operazioni: disambiguare la categoria grammaticale, riconoscere o produrre una forma fonetica ecc. In questo senso il dizionario non contiene definizioni dei significati e degli usi di una parola, ma è un repertorio di lemmi o forme flesse associate a una serie di informazioni linguistiche supplementari, disponibili all'accesso delle applicazioni computazionali. Il termine *dizionario macchina* indica appunto la relazione stretta tra repertorio lessicale e applicazioni computazionali che svolgono operazioni facendo ricorso a tale repertorio. Una specie di *dizionario macchina* è il lessico di frequenza che, nato inizialmente come strumento di pura investigazione linguistica, è entrato a far parte di numerose applicazioni computazionali di trattamento automatico del linguaggio. Le applicazioni più diffuse sono il riconoscimento e la sintesi vocale, la correzione ortografica, il *parsing* sintattico e il *tagging*, la traduzione automatica, e in generale tutte le applicazioni che incorporano un vocabolario.

4.3.2 Trattamento automatico del linguaggio

Il *Natural Language Processing* (NLP), detto in italiano il trattamento automatico del linguaggio (TAL) è uno dei settori fondamentali della linguistica computazionale. L'obiettivo di questo approccio, legato strettamente all'intelligenza artificiale, è lo svolgimento di compiti linguistici produttivi o ricettivi da parte di sistemi informatici. Il dominio principale, su cui si applica è, conformemente alla tradizione generativa, la sintassi, tuttavia negli ultimi anni sono emersi alcuni altri campi di ricerca, come l'estrazione di conoscenze da documenti testuali o il reperimento di specifici dati estratti da grandi database testuali.
Dall'integrazione dei programmi tradizionali con i metodi statistici è nato un nuovo campo di ricerca della linguistica computazionale, il c.d. *Statistical Natural Language Processing*, volto a migliorare le capacità di produzione o di analisi di frasi, servendosi di dati statistici. Il presupposto di base è che parlando o ascoltando, si ricorre, inconsciamente, a considerazioni sulle frequenze, ad esempio sulla frequenza con cui troviamo due parole una accanto all'altra in sequenza. I metodi statistici sono particolarmente vantaggiosi in diversi settori della linguistica computazionale, come il *tagging* grammaticale e il *parsing* sintattico.

Il *parsing* sintattico. Il *parsing* è un processo d'analisi che associa ad una frase di una lingua naturale una struttura che descrive la frase da un certo punto di vista (Chiari 2007: 101). Questo processo è considerato il nucleo dell'analisi delle lingue naturali. Il termine *parsing* indica una segmentazione linguistica: esistono *parser* morfologici, sintattici (il tipo canonico), semantici ecc. L'input al solito è costituito da una frase, cui si applicano alcune regole che permettono di produrre l'analisi della sua struttura. Per ottenere l'output desiderato, è necessario consultare un dizionario macchina, in cui sono presenti le parole della lingua e la loro categoria grammaticale (v. sopra), e il confronto con una grammatica di riferimento che definisce le regole di buona formazione di una frase in una data lingua. Un *parser* sintattico individua quindi le relazioni tra i costituenti della frase, le loro gerarchie ed attribuisce ad ogni forma della frase un'etichettatura di categoria grammaticale. Il primo *parser* sintattico per l'italiano (ATN-SYS) è stato implementato nel 1976 presso l'Istituto di Linguistica Computazionale del Consiglio Nazionale delle Ricerche a Pisa. Tra i suoi componenti principali troviamo, accanto ad una grammatica

complessa, un largo dizionario di 17.000 entrate: un sottoinsieme del *Dizionario Macchina Italiano* (DMI), selezionato sulla base del LIF.
Il *parsing* serve a numerosi scopi, una delle sue applicazioni più interessanti è l'annotazione automatica di grandi corpora. Il progetto più conosciuto di corpus sottoposto a *parsing* sintattico è il *Penn Treebank* dell'Univeristà di Pennsilvania, divenuto un punto di riferimento per la valutazione dell'efficienza degli strumenti di *parsing*.
Mentre il *parsing* sintattico di tipo tradizionale opera con regole definite (*rule-based parsing*), esistono tuttavia forme di *parsing* probabilistico, in cui alle regole grammaticali sono aggiunte informazioni relative alla probabilità d'occorrenza misurata sulla base di un corpus di riferimento annotato. Questo tipo di *parsing* usa algoritmi statistici per determinare quali etichette e gerarchie sono più probabili in dati contesti (cfr. Chiari 2007: 104).

Il *tagging* grammaticale. Il POS (*part of speech*) *tagging*, ossia l'etichettatura per categorie grammaticali, è, rispetto al *parsing*, un'operazione notevolmente più semplice. Il *tagger* non individua le categorie sintattiche, ma semplicemente la parte del discorso che una parola ricopre in una data frase (cfr. Chiari 2007: 106). Anche il *tagger* si serve di un dizionario che ha lo scopo di specificare quali categorie grammaticali possono corrispondere a ciascuna forma grafica. La disambiguazione delle forme omografiche può essere di due tipi: esistono *tagger* basati su regole e *tagger* di tipo probabilistico. I primi accedono ad una grammatica in cui sono formalizzate le regole di formazione dei diversi possibili sintagmi di una lingua, mentre i secondi si basano su statistiche di frequenza delle parti del discorso e delle loro sequenze.[24] Per fornire i dati sulle frequenze, si usano dei corpora testuali annotati in modo manuale, chiamati *training corpora*, giacché servono ad "allenare" il *tagger* a produrre delle predizioni corrette. Allenare vuol dire: considerare una percentuale significativa di testi del tipo che si sta analizzando, cui si attribuisce manualmente la marca grammaticale al fine di insegnare al programma quale marca è più probabile in determinati contesti. Chiari (2007: 108) rileva che i *tagger* probabilistici usati per l'analisi di corpora inglesi, raggiungono un grado di copertura corretta di etichette pari al 97-99% delle occorrenze.[25]
Tagger e *parser* probabilistici sono stati sviluppati per l'annotazione del corpus LIP. De Mauro *et al.* (1993: 64-76) hanno sviluppato un programma di analisi morfo-sintattica, basato su criteri statistici, in cui l'assegnazione delle categorie grammaticali si basa sulle frequenze con cui determinate sequenze di classi grammaticali sono state osservate in un particolare corpus campione.
Uno dei software più potenti di analisi morfo-sintattica dei testi è INTEX, ideato in Francia nel 1993 da Max Silberztein, il cui modulo italiano è stato costruito nel 2004 dall'Università di Salerno. INTEX è un analizzatore automatico di testi che consente di effettuare operazioni di *tagging* e *parsing*, *information retrieval*, di costruire grammatiche, dizionari elettronici, concordanze, ecc.[26]

[24] Il processo di disambiguazione morfologica, adottato nell'analisi del *Webcorpus* ungherese, è presentato in Kornai *et al.* (2006). L'analisi morfologica si basa sui dati statistici del paradigma di flessione nominale. In base alla probabilità d'occorrenza delle singole desinenze, il modello è in grado di predire quanto grande deve essere il corpus in cui una determinata forma occorre almeno una volta.
[25] Il *British National Corpus*, costituito da 100 milioni d'occorrenze è stato etichettato tramite il *tagger* probabilistico CLAWS (*Constituent Likelihood Automatic Word-tagging System*); per individuare il campo semantico delle occorrenze è stato costruito un sistema di *tagging* semantico chiamato ACASD (*Automatic Content Analysis of Spoken Discourse*).
[26] Il modulo ungherese è stato elaborato dall'Istituto di Linguistica dell'Accademia Ungherese delle Scienze, ed è disponibile dal 2004 sul sito (http://corpus.nytud.hu/INTEX/). Un altro programma d'analisi morfologica, sviluppata su basi euristiche dalla società Morphologic appositamente per la lingua

Estrazione automatizzata di informazioni. Il tagging e il parsing sono strumenti polifunzionali. Applicazioni di questo tipo sono, infatti, usate anche in una serie di settori di enorme interesse commerciale ed applicativo, cui si accenna brevemente: l'*information retrieval* ed *extraction*, il *text mining*, la *summarisation*.
L'indicizzazione automatica serve a produrre analisi rapide dei testi raccolti, attraverso l'identificazione delle parole chiave di un testo, ricorrendo ad analisi sintattiche e probabilistiche.
All'individuazione delle parole chiave è strettamente legato l'*information retrieval*, che permette di individuare, da un insieme ampio di documenti, quelli che soddisfano i nostri criteri di ricerca. In Italia Antonio Zampolli iniziò, già nel '69-'70, un progetto con la Camera dei Deputati per la creazione del *Dizionario Macchina Italiano* (DMI) da usarsi per sistemi di *information retrieval* sulle leggi. Ormai sono molto diffusi i software di recupero di informazioni in grado di effettuare ricerche su grandi collezioni di testi sulla base di richieste formulate come singole parole o come frasi: l'esempio più comune può essere quello dei motori di ricerca sul web. L'*information extraction* invece permette di cercare e selezionare i contenuti dei documenti appartenenti ad un insieme. In maniera schematica si può dire che l'*information retrieval* s'interessa al documento nella sua globalità, in quanto permette di scegliere, da un insieme ampio di documenti, quelli che soddisfano i nostri criteri di ricerca, mentre l'*information extraction* seleziona le informazioni specifiche all'interno del documento (cfr. Bolasco, Bisceglia, Balocchi 2004: 31).
In questi ultimi anni, suscita molto interesse nel filone statistico dell'analisi dei dati testuali, il *text mining*: si tratta di una delle applicazioni indirizzate ad aziende e istituzioni, le quali, dovendo interagire con enormi masse di materiali testuali spesso disponibili in rete, hanno il problema di selezionare, all'interno di queste fonti smisurate, i dati di loro interesse. Il *text mining* s'intende come attività di scavare nei testi, intesi come miniere di dati, in cui cercare informazioni di rilievo. Si tratta soprattutto di procedure di categorizzazione, di classificazione e di selezione intelligente di testi o sub-testi per estrarre da basi documentali (Bolasco 2005: 335). Applicazioni tipiche sono: smistamento di posta elettronica, ricerche su basi documentali giuridiche, nell'editoria, nelle telecomunicazioni, ecc.
La *summarization* consente di produrre automaticamente riassunti di testi o rapporti estratti da dati strutturati.[27] Tali sistemi accettano come input un testo in linguaggio naturale (un articolo scientifico), insieme con le richieste dell'utente, e produce come output un sommario del tipo specificato. L'estrazione automatizzata si basa su dati statistici.[28]

Il correttore ortografico. Una delle applicazioni elementari dei dati di frequenza ricavati di larghi corpora è la programmazione dei strumenti di correzione ortografica. La maggior parte dei correttori ortografici funziona mediante la comparazione con un

ungherese, è il software Humor, utilizzato per l'annotazione del corpus storico della lingua ungherese (*Magyar Történeti Szövegtár*).
[27] Una ricerca in corso per l'ungherese è quella di Lengyelné Molnár (2006), volta a creare un programma di estrazione automatizzata off-line, che permetta agli scienziati di generare automaticamente riassunti di studi pubblicati nel proprio campo di ricerca. La base dell'applicazione sarà una lista di frequenza che contiene le radici delle parole più rilevanti di un determinato campo scientifico.
[28] Secondo la legge di Zipf l'ambito delle parole significative non coincide con i termini più frequenti nei testi. Ad esempio in un eserciziario di matematica le parole *esercizio* ed *esempio*, pur avendo un alto numero d'occorrenza, non sono significative dal punto di vista informativo. Il lessico specifico di un determinato settore si trova tutt'al più nella fascia intermedia delle liste di frequenza.

dizionario di riferimento, inteso come repertorio in cui figurano anche le forme flesse delle parole. Il correttore spesso non dispone di un dizionario completo, ma solo di una lista di radici e una serie di regole di scomposizione morfologica che permettono di controllare la plausibilità della forma (Chiari 2007: 111).
Il problema più complesso è quello dei suggerimenti per la correzione. Per proporre la parola giusta, il correttore usa, oltre allo spoglio del dizionario di riferimento, algoritmi che prevedono una serie di combinazioni di lettere alfabetiche legittime e frequenti. Le analisi statistiche di tipo probabilistico permettono di individuare i candidati migliori a rappresentare sequenze possibili, probabili e frequenti e di analizzare delle tipologie d'errore più frequenti. Per la correzione degli errori grammaticali si ricorre al *tagging* che permette di disambiguare le categorie grammaticali e individuare ad es. un mancato accordo. Rimangono tuttavia problemi irrisolti, come la mancanza di una parola dal dizionario di riferimento, o quando il termine errato effettivamente esiste, perciò non è riconosciuto.
Un'interessante ricerca per aumentare il grado di copertura dei sistemi di controllo ortografico è stata condotta da Halácsy *et al.* (2004). La base dei sistemi primitivi di controllo ortografico è la lista delle forme corrette, predisposte in ordine di frequenza. Nell'ungherese, per ottenere un grado di copertura del 50%, bastano 3.000-6.000 forme (prendendo come base di riferimento il *Webcorpus* ungherese). Questo metodo però ha i suoi limiti, giacché la maggior parte delle forme ricorre una sola volta nei testi, si tratta quindi di *hapax*, la cui inclusione nelle liste predisposte richiederebbe un enorme lavoro, d'altra parte non migliorerebbe in misura corrispondente il grado di copertura del programma. Pertanto l'analisi morfologica è particolarmente importante nel caso di una lingua come l'ungherese, di cui è impossibile immagazzinare tutte le forme flesse esistenti. È opportuno che il sistema sia in grado di analizzare le flessioni, quindi nel programma di controllo ortografico va implementato anche un analizzatore morfologico. Prima si procede dunque all'immissione delle radici più frequenti, mancanti nella lista, successivamente si analizzano le nuove radici con l'aiuto di HunStem, un programma di *tagging* morfologico. In accordo alla legge di Zipf è stato verificato che dopo aver introdotto le 100.000 radici più frequenti, la percentuale di forme scoperte era del 5,6%, nel caso di 1 milione di radici 3,2%, con 10 milioni 1,8% e con 100 milioni 1%. Si conclude quindi, che l'esattezza dei sistemi di controllo ortografico dipende in primo luogo dal numero di radici incluse, ma con l'ampliamento dell'inventario, a causa della legge di Zipf, l'esattezza del programma aumenta in misura decrescente.

Il sistema T9. Un'altra applicazione basata su dati di frequenza è T9, sviluppata dalla società Tegic, per facilitare l'inserimento delle parole tramite la tastiera del cellulare. Utilizzando un vocabolario integrato, la tecnologia T9 è in grado di prevedere quale parola l'utente sta scrivendo e propone di completarla, risparmiando la pressione dei tasti corrispondenti alle lettere mancanti. In base ai testi da noi di solito inviati, è fatta una scelta previsionale circa la parola più probabile fra le diverse possibili con la combinazione digitata. La nuova versione, XT9 è pensato per essere utilizzato con tutti gli apparecchi elettronici, palmari, computer e riconosce anche testi parlati e scritti a mano sullo schermo dei palmari. Sulla stessa idea si basa iTAP, sistema per l'immissione facilitata dei testi nei cellulari, sviluppato da Motorola.

Traduzione automatica dei testi. I progressi del trattamento automatico del linguaggio (in particolare il *tagging* e il *parsing*), negli anni Settanta hanno dato nuovo impulso ai progetti di traduzione automatica. Al semplice confronto lessicale sono affiancate componenti d'analisi morfologica, sintattica e disambiguatori semantici. Il primo dei

grandi sistemi di traduzione fu *Systran* (acronimo di *System Translation*), sviluppato per tradurre dal russo all'inglese, successivamente integrato con nuove coppie di lingue.
In questo campo si sono impegnati soprattutto gli organismi internazionali, come l'Ue, la Nato, l'Onu, che oltre al promuovere progetti di traduzione automatica, hanno elaborato banche dati terminologiche, contribuendo in tal modo alla standardizzazione delle traduzioni tecnico-scientifiche dei documenti. Altre fonti di stimolo sono le grandi multinazionali che richiedono la traduzione di manuali tecnici e vari documenti. La maggioranza delle applicazioni oggi disponibili è pensata, appunto, per testi specialistici (tecnico-scientifici, economici), giacché questi presentano alcune caratteristiche che permettono di garantire la qualità della traduzione. I motivi del successo sono da cercare nell'uso di un lessico ridotto e prevedibile, di polirematiche e collocazioni cristallizzate, tipiche del lessico specialistico, e di strutture sintattiche prevedibili ed esplicite.
I principali modelli adottati per programmare i sistemi di traduzione automatica si suddividono in due grandi tipi. La prima generazione di applicazioni è costituita da sistemi basati su regole, nei quali sono identificate alcune regole formali di descrizione e trasformazione delle unità. In una procedura di lavoro seriale, ogni unità testuale è analizzata e scomposta dal punto di vista morfologico fino a individuare il suo corrispondente diretto nella lingua d'arrivo. La seconda famiglia di sistemi di traduzione è costituita dai programmi d'impianto statistico (probabilistico). Questi approcci, basati su dati ricavati da larghi corpora, si sono diffusi dagli anni Novanta in poi. Utilizzando corpora paralleli, questi sistemi rilevano sulla base delle unità da tradurre strutture già tradotte, utilizzabili per la traduzione. Per scegliere il migliore candidato per la traduzione, fanno uso di parametri statistici, come la probabilità di co-occorrenza e la frequenza delle parole nei corpora di riferimento.
Chiari (2007: 122) nota che i risultati dei sistemi di impianto statistico si sono rivelati più promettenti rispetto agli approcci basati su regole, specialmente per la soluzione di determinati problemi di traduzione. I sistemi attuali di traduzione automatica sono spesso ibridi: arricchiscono le metodologie basate su regole con quelle basate su corpora e introducono moduli che permettono lo sviluppo di diverse strategie di risoluzione secondo lo specifico problema traduttivo da affrontare.[29]

4.3.3 Tecnologie della lingua parlata

Le aree oggi in più forte espansione della linguistica computazionale sono quelle legate alla tecnologia del parlato, soprattutto grazie alle loro applicazioni commerciali. Le due aree che costituiscono il fondamento delle tecnologie più avanzate, sono la sintesi del parlato e il riconoscimento automatico di testi prodotti in parlato spontaneo. La loro importanza si deve alla necessità di preparare strumenti che rendano il dialogo con la macchina il più naturale e semplice possibile, d'altro lato negli enormi benefici che derivano dall'applicazione di tali tecnologie in situazioni di disabilità, ma anche nei contesti didattici.
La sintesi del parlato (*text to speach*) si occupa della costruzione di programmi in grado di produrre frasi allo scopo di far interagire la macchina con l'uomo. La conversione di un testo in formato elettronico in una sequenza acustica, paragonabile alla voce umana, è un'attività molto complessa, in cui i dati statistici hanno un ruolo importante, soprattutto nel *parsing* sintattico e prosodico (cfr. Chiari 2007: 134-140). Grazie ai risultati raggiunti

[29] Nel settembre del 2006 è stato avviato il progetto *EuroMatrix*, il cui obiettivo è di creare nei prossimi due anni un software ibrido che sia capace di tradurre tra tutte le lingue europee.

nel settore del trattamento automatico del linguaggio, oggi questa tecnologia è avanzatissima: sono disponibili voci maschili o femminili molto difficilmente distinguibili dalla voce umana, con corretta pronuncia delle parole della frase. Le applicazioni sono innumerevoli ed aumentano ogni giorno. La sintesi del parlato è una realtà operativa nelle telecomunicazioni, ed esistono strumenti con cui utenti disabili possono produrre parlato usando una tastiera, i non vedenti possono ottenere la versione parlata di un testo scritto ecc. Un altro settore è costituito dall'educazione linguistica per la predisposizione di strumenti per l'apprendimento di una (seconda) lingua (cfr. sez. 3.2.8).

Il riconoscimento del parlato (*speech recognition*) racchiude una serie di applicazioni in grado di identificare le sequenze del parlato spontaneo, convertirle in testo elettronico e permettere l'interazione con programmi usando la voce. Questo settore nasce come derivazione degli studi di sintesi di testi, perciò è detta anche *speech to text*. Questi sistemi servono a diversi scopi: dettare testi e dati al computer, dare comandi vocali (si pensi alla funzione *voice dial* dei cellulari), sviluppare strumenti di scrittura per disabili, ecc. (cfr. Chiari 2007: 142-145). Dagli anni Ottanta in poi nascono i modelli di tipo statistico che hanno migliorato le prestazioni mediante fasi di addestramento. Questi sistemi fanno uso di corpora di parlato spontaneo, sotto forma di *training corpus*, al fine di garantire un adeguato trattamento di diversi input linguistici. In genere, durante la fase di riconoscimento sono suggeriti diversi candidati possibili, le cui caratteristiche, per la presenza di omofonie più o meno estese, sono compatibili con l'input vocale. Negli ultimi anni si sono sviluppati sistemi contenenti moduli probabilistici che assegnano ad ogni possibile candidato diverse probabilità e selezionavano la sequenza più probabile.

De Mauro (1994: 116) avverte che il successo di tali operazioni è assicurato da due condizioni: che le realizzazioni siano scandite e complete, formali (ogni forma deve avere un corrispettivo fonico non ambiguo); e che il meccanismo identificatore possa avvalersi di una memoria dotata di istruzioni sulla probabilità d'occorrenza delle sequenze. Come nel caso della traduzione automatica, la prestazione aumenta con la prevedibilità dell'input (si pensi ai linguaggi speciali), e decresce nel caso di testi che spaziano su domini molto diversi. Non sorprende, quindi, che tra le tipologie funzionali del riconoscimento del parlato il minor tasso d'errore si registra nel caso dei comandi vocali, poiché le possibilità date all'utente sono limitate a un numero ristretto di comandi.

5 Il lessico nei sistemi di certificazione

In questo capitolo si passano in rassegna le iniziative volte a definire il lessico necessario per ottenere titoli di competenza in italiano come lingua straniera. Nella prima parte del capitolo si descrivono due progetti importanti del Consiglio d'Europa, mentre nella seconda si analizzano i requisiti lessicali dei sistemi di certificazione elaborati in Italia da vari enti di certificazione.

5.1 Gli apporti del Consiglio d'Europa

Uno degli obiettivi prioritari del Consiglio d'Europa fin dalla sua fondazione nel 1949, è stato quello di incoraggiare e favorire la conoscenza reciproca tra i popoli europei. Tale conoscenza si attua anche attraverso la promozione di una più ampia diffusione delle lingue europee moderne. Come recita il *Quadro comune europeo* (p. 3):[1]

> "solo una migliore conoscenza delle lingue europee moderne riuscirà a facilitare la comunicazione e l'interazione tra cittadini che parlano madrelingue diverse e ad agevolare la mobilità, la comprensione reciproca e la cooperazione in Europa, superando pregiudizi e discriminazioni."

Negli anni Settanta il Consiglio d'Europa è stato la principale agenzia comunitaria che ha affrontato il tema della politica linguistica. Ciò era la conseguenza di una scelta ideale, ma Vedovelli (2002a: 31) avverte che accanto a questa motivazione, la necessità di elaborare una politica linguistica comune è stata sollecitata anche dai problemi posti dai movimenti migratori. Negli anni settanta i lavoratori emigrati dall'Europa del sud nei Paesi dell'Europa settentrionale si sono ormai stabilizzati, e impongono agli Stati ospiti di affrontare problemi come l'inserimento scolastico delle seconde generazioni, e lo sviluppo della competenza linguistica degli immigrati stessi. I progetti del Consiglio d'Europa degli anni Settanta mirano a gestire il contatto fra le lingue dei Paesi ospiti e quelle dei gruppi di immigrati. Nel tentativo di risolvere questi problemi, il Consiglio affronta globalmente la questione della didattica delle lingue, riportando in primo piano la riflessione su tali tematiche.

5.1.1 Livello soglia

È all'interno di quest'azione linguistica, finalizzata a creare una politica linguistica comune che prende forma il progetto *Lingue moderne*, mirante alla diffusione della conoscenza delle lingue europee tra gli adulti della Comunità europea. Il progetto ha implicato un'importante fase di riflessione teorico-metodologico sull'insegnamento delle lingue, di cui i vari "livelli soglia" sono i risultati concreti. Nelle parole di J. A. Van Ek (1979: 57), responsabile del citato progetto, il *Livello soglia* è:

[1] *Quadro comune europeo di riferimento per le lingue: apprendimento, insegnamento, valutazione* (2002) Consiglio d'Europa, La Nuova Italia – Oxford, Firenze. (Titolo originale: *Common European Framework of Reference for Languages: Learning, Teaching, Assessment 1996-2001*, Council of Europe, Strasbourg.)

> "un livello di capacità predominantemente orale al quale i discenti saranno in grado di sopravvivere, dal punto di vista linguistico, in un Paese straniero e di stabilire e mantenere contatti sociali con parlanti della lingua straniera […] Ciò significa che i nostri obiettivi minimi di apprendimento dovranno includere cose come la capacità di scambiare informazioni su se stessi e sugli altri, sul proprio lavoro, sui propri interessi, sulla propria provenienza sociale, sul proprio paese, ecc."

Si tratta di aspirazioni fondamentali nel quadro della cooperazione europea. In quanto obiettivi di insegnamento/apprendimento delle lingue, tuttavia, sono relativamente diversi da quelli tradizionali, soprattutto da quelli tipici dei curricoli scolastici. I normali corsi di lingue sono costituiti incentrando l'attenzione sulla lingua da insegnare, mentre la "sensibilità all'utente" è bassa, e tocca forse solo il livello lessicale. Il principio alla base del progetto del Consiglio d'Europa è invece proprio la massima sensibilità all'utente, così da capovolgere la gerarchia usuale nella selezione del materiale linguistico.
L'approccio di fondo del progetto è appunto il concetto di sistema di unità (*unit system*), secondo cui le quattro attività linguistiche fondamentali (parlare, ascoltare, leggere, scrivere) e l'insieme dei dati da acquisire può essere analizzato in unità che messe insieme in modo graduale, formano il corpo organico della conoscenza. Si cita Galli de' Paratesi (1981: 28).

> "Gli obiettivi di un corso di lingua sono *un insieme di forme diverse atomizzate di competenza* che permette ai discenti di fornire delle prestazioni attraverso la lingua ... nelle situazioni che si suppone possano scaturire dai loro bisogni." (corsivo nostro)

Le concrete forme linguistiche (lessicali e grammaticali) da insegnare sono quindi selezionate non in base al loro valore nel sistema della lingua, ma all'uso che determinati gruppi di parlanti possono farne in determinate situazioni. La lingua è vista dunque come uno strumento sociale, che permette di interagire con gli altri. Nel quadro teorico fortemente influenzato dalla sociolinguistica e dalla pragmatica, risulta centrale l'idea che per selezionare in modo corretto le forme e le strutture da insegnare, bisogna partire dall'individuazione di uno o più destinatari tipo, dei quali fosse necessario analizzare i bisogni comunicativi, dunque non solo le situazioni di uso reale della lingua, i ruoli sociali dei partecipanti, gli scopi e gli argomenti dell'interazione (orientamento funzionale), ma anche i concetti e le nozioni che gli utenti ipotizzati avrebbero avuto bisogno di esprimere nella L2 (orientamento nozionale).
L'attenzione è indirizzata ad un pubblico adulto, spesso già inserito nel mondo professionale, che, a differenza dei giovani in fase di formazione scolastica, sanno con notevole precisione che cosa devono imparare. Si parte quindi dall'assunto che gli individui adulti, che intraprendono lo studio di una lingua, sono spinti da motivazioni ben precise. Una volta scelto l'utente-tipo, i suoi bisogni comunicativi sono analizzati in termini funzionali (in quali situazioni, con quali ruoli, circa quali argomenti e con quali scopi si troverà ad interagire nella lingua in esame) e nozionali (quali "nozioni", ovvero quali concetti dovrà poter esprimere e/o comprendere). Questa griglia funzionale-nozionale che nasce dall'analisi dei bisogni serve da filtro di scelta dei materiali linguistici; questi ultimi, debitamente organizzati in modo sequenziale, e con l'opportuno apparato didattico, formeranno i corsi di lingua (cfr. Ciliberti 1994: 100-105).
La base per la costruzione dei sillabi nozionali-funzionali è la teoria degli atti linguistici, introdotta negli anni Sessanta da Austin e Searle. Parlando compiamo degli atti, come ad es. presentarsi, chiedere informazioni ecc. Bisogna offrire al discente una competenza comunicativa che gli permetta di usare la lingua in un determinato contesto, selezionando tra tutte le espressioni linguistiche a sua disposizione quella più adeguata a realizzare un determinato atto linguistico. Nella descrizione dettagliata del *Livello soglia,* infatti, le

concrete forme linguistiche hanno un ruolo secondario rispetto alle funzioni (atti), che il discente deve essere in grado di compiere nella lingua straniera. Ciò si rispecchia anche nella struttura pressappoco identica delle specificazioni relative alle singole lingue: la prima parte include una complessa e larga classificazione degli atti linguistici, seguita da un repertorio delle concrete forme lessicali e grammaticali. Le liste delle concrete forme linguistiche non si pongono come un insieme chiuso, bensì sono da considerare come linee guida. Pertanto il *Livello soglia* non è altro che la formulazione di obiettivi generali che servono come guida per l'organizzazione dell'attività didattica, sempre in vista dei bisogni degli studenti. Il concetto di livello soglia non si pone, infatti, come polivalente a tutti i bisogni semantici, ma come specifico dei bisogni comunicativi di determinati gruppi di destinatari (cfr. Galli de' Paratesi (1981: 15).

I vari livelli soglia sono il penultimo stadio di questo processo: repertori di forme e strutture, nelle rispettive lingue, selezionati in base a criteri funzionali e nozionali esplicitati, criteri a loro volta motivati sulla base di un gruppo target, accuratamente definito. Queste opere rappresentano il penultimo stadio perché sono concepite per utenti che operano nell'organizzazione dell'insegnamento dell'italiano come lingua straniera, quindi non sono intese per essere usate direttamente dai discenti stessi. Qui bisogna notare un'importante caratteristica di tutti i livelli soglia, ovvero la flessibilità. Le descrizioni, nel loro complesso, possono essere adattati alle esigenze di specifici gruppi di utenti, variando ad esempio il repertorio lessicale.

Osserviamo che nonostante il numero limitato delle unità lessicali (1.500 del *Treshold Level* e del *Livello Soglia* e ca. 2.200 della lista ungherese) e la gamma ridotta delle strutture grammaticali da imparare, gli obiettivi didattici sono assai ampi: dotare l'apprendente dei mezzi linguistici necessari per cavarsela in quasi tutte le situazioni quotidiane (Van Ek 1977). Le motivazioni che hanno indotto a scegliere il limite di carico non sono date, Galli de' Paratesi (1981: 15) si limita ad osservare solo che:

> "al di sotto di questa soglia mancano gli strumenti linguistici necessari per garantire ai gruppi di destinatari scelti una soddisfacente interazione sociale nelle situazioni che si troveranno ad affrontare."

400 unità lessicali del *Treshold Level* e del *Livello Soglia* e una quantità non precisata delle forme della lista ungherese sono proposte per l'apprendimento passivo. Purtroppo non si spiegano i criteri di scelta delle forme R (espressioni che gli allievi devono usare solo "ricettivamente") rispetto alle altre.

Nell'ambito del progetto, tra il 1975 e il 2005 sono stati messi a punto una serie di sillabi, a cominciare con *Treshold Level* per l'inglese, il concetto "livello soglia" è stato applicato a tutte le lingue europee, tra cui anche all'italiano, dalla sociologa Nora Galli de' Paratesi (1981) (2.2.17) e all'ungherese dai linguisti Aradi, Erdős e Sturcz (cfr. 3.2.6).[2]

Accanto al *Threshold Level* (1975 e l'ultima edizione del 1991) per l'inglese sono stati elaborati, sempre da Trim e Van Ek, il *Waystage Level* (1980) (il livello inferiore al

[2] Per l'inglese vedi: Van Ek J. A. (1975) *TheTreshold Level English,* Strasbourg, Council of Europe (2ª edizione aggiornata 1990); Van Ek J. A., Trim J. L. M. (1980) *Waystage English*, Strasbourg (2ª edizione aggiornata 1990, Cambridge, CUP.), Council of Europe; per il francese: Coste D. *et alii* (1976) *Un Niveau Seuil*, Conseil de l'Europe, Strasbourg; per lo spagnolo: Slagter P. J. (1979) *Nivel Umbral,* Consejo de Europa, Strasbourg; per il tedesco: Baldegger M., Müller M., Schneider M. (1980) *Kontaktschwelle Deutsch als Fremdsprache,* Europarat, Strasbourg; per l'italiano: Galli de' Paratesi N. (1982) *Livello soglia*, Consiglio d'Europa, Strasburgo; per le altre lingue si pubblicano i rispettivi livelli soglia nell'ordine cronologico seguente: danese (1983), olandese (1985), norvegese (1988), basco (1988), portoghese (1988), galiziano (1993), gallese (1994), russo (1996), maltese (1997), lituano (1997 e 2000), estone (1997), lettone (1997), catalano (1999), greco (1999, 2001, 2003 e 2005) ungherese (2000), ceco (2001), rumeno (2002), sloveno (2004).

Threshold con sole 500 unità lessicali) e il più recente *Vantage Level* (1996) (il livello superiore al *Threshold)* che rappresentano una continuità sia nel senso di una progressione nell'apprendimento, sia nel senso di una comune linea teorico-metodologica. I destinatari del *Vantage Level* sono adulti che soffrono di una certa frustrazione, poiché non dispongono degli strumenti linguistici (in particolare lessicali) adeguati a rispondere alle esigenze della vita quotidiana, e che desiderano acquisire un più completo controllo del processo comunicativo. In generale, si tratta di un livello, il cui raggiungimento consente una maggiore "autonomia" nell'uso linguistico. La struttura e la suddivisione in capitoli del *Vantage Level* corrisponde a quelle del *Threshold Level*.
I criteri di sociolingistica e pragmatica sono, per loro natura, criteri d'analisi in situazioni date, utili per mettere in relazione variabili linguistiche rilevate con variabili delle situazioni, in modo da individuare regolarità altrimenti nascoste. Permettono di descrivere in modo più fine ed organizzato la nostra conoscenza del mondo. Notiamo con Berretta (1983: 245) che è proprio questa maggiore finezza ed organizzazione della descrizione che distingue i livelli soglia dai manuali di conversazione per turisti, anch'essi basati sulla scelta e analisi delle situazioni-tipo in cui l'utente si troverà ad usare la lingua. Pertanto nel Livello soglia le situazioni meglio descritte sono le stesse che, nella vita quotidiana, sono altamente standardizzate anche in termini di formule linguistiche usate, mentre le interazioni personali imprevedibili restano genericamente definite. È indiscutibile il merito dell'approccio di offrire una base comune, concreta ed accessibile, su cui insegnanti, autori di manuali ed elaboratori di test possono creare dei materiali.[3]

5.1.2 Il *Quadro comune europeo per le lingue*

L'impegno del Consiglio d'Europa in questo campo non si è fermato all'elaborazione dei livelli soglia. Nel novembre 2001 per iniziativa del Governo Federale Svizzero si è tenuto a Rüschlikon un convegno internazionale dal titolo: "*Transparency and Coherence in Language Learning in Europe: Objectives, Evaluation, Certification*." Il congresso si inseriva nel contesto del progetto denominato: *Modern Languages Project. Language Learning for European Citizenship* (1989-1996). Gli esperti decisero di sviluppare un documento che potesse diventare una sorta di quadro di riferimento europeo in cui, chi dovesse sviluppare curricoli, materiali per l'insegnamento o per la verifica, potesse trovare riflessioni teoriche, indicazioni, spunti espressi in forma coerente. La prima versione del 1995 è stata sottoposta ad un'accurata revisione e nel 2001 è uscito il documento dal titolo: *Common European Framework of Reference for Languages*.[4] Questo documento rappresenta l'ultimo stadio di un processo avviato nel 1971, rivolta ad abbattere le barriere linguistiche fra i popoli europei.
Sempre nell'ambito delle iniziative del Consiglio d'Europa vale la pena di ricordare brevemente il progetto *Portfolio Europeo delle Lingue (Európai Nyelvtanulási Napló)*,

[3] Uno dei manuali didattici elaborati a partire dalla lista del *Livello soglia* per l'italiano è il libro di Chiùchiù, Minciarelli e Silvestrini (1990), intitolato *In italiano*.
[4] La versione inglese è disponibile on line: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf e in edizione cartacea: *Common European Framework of Reference for Languages: Learning, Teaching, Assessment* (2001) Council of Europe, Strasbourg. Il documento è stato tradotto in quasi tutte le lingue europee, l'altro in ungherese: *Közös Európai Referenciakeret: nyelvtanítás, nyelvtanulás, értékelés* (2002) Európa Tanács, Strasbourg (comunemente detto *KER*), e in italiano: *Quadro comune europeo di riferimento per le lingue. Apprendimento, insegnamento, valutazione* (2002) Consiglio d'Europa. Strasburgo (comunemente chiamato *Quadro comune* o *Framework*).

uno strumento che s'ispira al *Quadro comune europeo,* ma che è pensato allo scopo di descrivere la progressione di giovani e adulti nella conoscenza di una data lingua straniera. Lo studente sarà chiamato a registrare le proprie esperienze linguistiche e formative (sia quelle apprese in maniera sistematica in un ambiente istituzionale, sia quelle non sistematiche) nel *Portfolio* che a sua volta documenterà il percorso e il livello di apprendimento delle lingue da parte dello studente, fornendogli una specie di lasciapassare per il lavoro o per lo studio, riconosciuto da tutti i Paesi europei. Nel *Portfolio* si registreranno: attestati e certificati conseguiti dall'apprendente; esperienze di studio (corsi di lingua) e di soggiorno all'estero; griglie per l'auto-osservazione e l'autovalutazione continua; prestazioni personali sotto osservazione. Vedovelli (2002a: 64) ricorda, infatti, che per rendere effettivamente possibile l'attuazione del *Portfolio*, quale biografia linguistica dei cittadini europei, occorre un codice comune per la traduzione dei vari sistemi di valutazione. La riconoscibilità sul piano europeo dei titoli e delle competenze registrate nel *Portfolio* è resa possibile appunto dai parametri comuni di riferimento esposti nel *Quadro comune.*

Scopo del *Quadro comune* è di stabilire dei livelli comuni di competenza linguistica, al fine di promuovere il riconoscimento transnazionale delle certificazioni europee. Esso (p. 1) intende:

> "fornire una base comune per l'elaborazione di sillabi linguistici, di linee guida per costruire curricula, esami, libri di testo ecc. in Europa."

Per far questo, descrive in modo esaustivo ciò che gli apprendenti devono imparare a fare per usare la lingua per comunicare, e quali conoscenze e abilità devono sviluppare. La scala proposta dal *Quadro comune* vuole fornire uno strumento di lavoro basato sui criteri di stabilità e di standardizzazione.

L'opera riprende l'impianto originario, già espresso nei livelli soglia, secondo cui al centro di ogni insegnamento linguistico sta l'apprendente con i suoi bisogni e l'obiettivo primo è lo sviluppo della competenza linguistico-comunicativa. Questa va a sua volta considerata una somma di più competenze:

a) una competenza pragmatica (capacità di compiere azioni linguistico-comunicative adeguate ai bisogni, dunque di agire in modo efficace nei diversi contesti);
b) una competenza sociolinguistica (capacità di dominare le condizioni socioculturali d'uso della lingua, le convenzioni sociali, i generi testuali, le varietà di lingua più adeguate alle diverse situazioni);
c) una competenza linguistica (capacità di discriminare e scegliere gli elementi linguistici più opportuni per realizzare le diverse intenzioni comunicative). In questo gruppo si inserisce l'oggetto del nostro esame, ovvero la competenza lessicale che (insieme a quella grammaticale, semantica, fonologica, ortografica e ortoepica) costituisce un indicatore della competenza linguistica.

Tutte queste (sotto)competenze vanno sviluppate armonicamente, vanno selezionati e messi in sequenza i contenuti ritenuti più idonei al raggiungimento degli obiettivi d'apprendimento pertinenti ai vari livelli di competenza. Le attività linguistiche (ricezione, produzione, interazione e mediazione) si attuano in quattro domini specifici (pubblico, personale, occupazionale ed educativo) che coincidono per larga misura con quelli del *Livello Soglia*.[5] Tutte le situazioni che possono ricorrere nei vari domini sono descrivibili in termini di luogo in cui accadono, di istituzioni le cui norme ne determinavano l'andamento, di persone coinvolte, di oggetti animati e inanimati dell'ambiente, di eventi che si verificano, di azioni eseguite dalle persone coinvolte e di testi orali o scritti che è necessario capire o produrre. Per ogni dominio è stata elaborata

[5] Cfr. i domini del LS: gruppo dei pari; servizi pubblici; trattative commerciali; istruzione.

una tabella di nozioni, alle quali dovrebbero corrispondere i contenuti di lingua straniera di un sillabo scolastico.[6]
I livelli di competenza sono scanditi dal Consiglio d'Europa in 6 diverse fasce indicate con le sigle A1, A2, B1, B2, C1, C2 che corrispondono rispettivamente al livello di conoscenza di un utente basico, indipendente e competente. La tripartizione dei profili riflette la distinzione effettuata dal Consiglio d'Europa alla fine degli anni Settanta.

I livelli del *Quadro comune*		
A1	Utente basico	Breakthrough
A2		Waystage
B1	Utente indipendente	Threshold
B2		Vantage
C1	Utente competente	Proficiency
C2		Mastery

Livelli di competenza del *Quadro comune europeo*

In due ricerche, condotte tra il 1993 e il 1996 con la partecipazione di quasi 300 insegnanti e 2.800 studenti, sono stati elaborati i c.d. descrittori che permettono di definire, per ognuno dei sottolivelli previsti, sia le attività corrispondenti ai vari ambiti di abilità, sia ciò che ad un certo livello l'apprendente deve saper fare. I descrittori di competenze, riportati nella tabella sottostante, servono a dare un quadro globale. Un'analisi di funzioni, nozioni, grammatica e lessico necessari per portare a termine i compiti comunicativi descritti nelle scale potrebbe far parte dell'elaborazione di nuovi repertori di specificazioni linguistiche relative alle singole lingue.

A1	Riesce a comprendere e utilizzare espressioni familiari di uso quotidiano e formule molto comuni per soddisfare bisogni di tipo concreto. Sa presentare se stesso/a e altri ed è in grado di porre domande su dati personali e rispondere a domande analoghe (il luogo dove abita, le persone che conosce, le cose che possiede). È in grado di interagire in modo semplice purché l'interlocutore parli lentamente e chiaramente e sia disposto a collaborare.
A2	Riesce a comprendere frasi isolate ed espressioni di uso frequente relative ad ambiti di immediata rilevanza (ad es. informazioni di base sulla persona e sulla famiglia, acquisti, geografia locale, lavoro). Riesce a comunicare in attività semplici e di routine che richiedono solo uno scambio di informazioni semplice e diretto su argomenti familiari e abituali. Riesce a descrivere in termini semplici aspetti del proprio vissuto e del proprio ambiente ed elementi che si riferiscono a bisogni immediati.
B1	È in grado di comprendere i punti essenziali di messaggi chiari in lingua standard su argomenti familiari che affronta normalmente al lavoro, a scuola, nel tempo libero ecc. Se la cava in molte situazioni che si possono presentare viaggiando in una regione dove si parla la lingua in questione. Sa produrre testi semplici e coerenti su argomenti che gli siano familiari o siano di suo interesse. È in grado di descrivere esperienze e avvenimenti, sogni, speranze, ambizioni, di esporre brevemente ragioni e dare spiegazioni su opinioni e progetti.
B2	È in grado di comprendere le idee fondamentali di testi complessi su argomenti sia concreti sia astratti, comprese le discussioni tecniche nel proprio settore di specializzazione. È in grado di interagire con relativa scioltezza e spontaneità, tanto che l'interazione con un parlante nativo si sviluppa senza eccessiva fatica e tensione. Sa produrre testi chiari e articolati su un'ampia gamma di argomenti e esprimere un'opinione su un argomento di attualità, esponendo i pro e i contro delle diverse opzioni.
C1	È in grado di comprendere un'ampia gamma di testi complessi e piuttosto lunghi e ne sa ricavare anche il significato implicito. Si esprime in modo scorrevole e spontaneo, senza un eccessivo sforzo per cercare le parole. Usa la lingua in modo flessibile ed efficace per scopi sociali, accademici e professionali. Sa produrre testi chiari, ben strutturati e articolati su argomenti complessi, mostrando di saper controllare le strutture discorsive, i connettivi e i meccanismi di coesione.
C2	È in grado di comprendere senza sforzo praticamente tutto ciò che ascolta o legge. Sa riassumere informazioni tratte da diverse fonti, orali e scritte, ristrutturando in un testo coerente le argomentazioni e le parti informative. Si esprime spontaneamente, in modo molto scorrevole e preciso e rende distintamente sottili sfumature di significato anche in situazioni piuttosto complesse.

[6] Nel dominio pubblico, sono indicati come luoghi le strade, i negozi, le sale teatrali ecc.; come istituzioni gli organismi sanitari, i partiti politici ecc.; come persone i pubblici ufficiali, i religiosi, i camerieri, i commessi ecc.; come oggetti il denaro, i beni, i documenti ecc.; come azioni gli acquisti, i viaggi; come testi gli annunci pubblici, le etichette, i menu, ecc. (cfr. Mazzotta 2002: 12).

Il *Quadro comune europeo* si limita dunque a suggerire una cornice comune, un quadro teorico di riferimento utile nella pianificazione curricolare, cui seguiranno sillabi differenziati per lingua. L'interpretazione degli indicatori sulla competenza di un apprendente di un dato livello è lasciata agli addetti del settore, proprio perché non si tratta quindi di un programma, ma di uno strumento per sviluppare programmi (Vedovelli 2002a: 82).[7]

Presso l'Università per Stranieri di Perugia è in corso un progetto finalizzato alla descrizione dei contenuti linguistici per la lingua italiana, rapportati ai livelli A1 – B2 del Quadro Comune di Riferimento Europeo per le Lingue. Il progetto si colloca all'interno dell'azione promossa dalla Divisione per le Politiche Linguistiche del Consiglio d'Europa. Modelli di riferimento riconducibili al progetto italiano sono già stati realizzati, o sono in corso di realizzazione, per altre lingue europee.

Oltre al progetto per la definizione dei materiali referenziali, è in corso un altro progetto internazionale, che intende creare procedure comuni e condivise al fine di validare la correlazione di esami linguistici ai livelli del Quadro Comune di Riferimento Europeo. L'Università per Stranieri di Perugia è l'unica istituzione per la lingua italiana coinvolta in entrambi i progetti.

I contenuti linguistici per l'italiano, rapportati ai livelli A1 – B2 del Quadro Comune Europeo, vengono definiti in base alle seguenti categorie descrittive: generi discorsivi; funzioni; nozioni generali; nozioni specifiche; elementi grammaticali; elementi fonetici e prosodici. Gli inventari saranno raccolti in un CD ROM per facilitarne la consultazione. L'opera sarà poi corredata di un supporto cartaceo contenente saggi ed approfondimenti relativi a: metodi e procedure adottati per la realizzazione degli inventari, strategie di apprendimento, competenze culturali. La pubblicazione dei materiali è prevista per la seconda metà del 2007.

5.1.2.1 La competenza lessicale nel *Quadro comune*

Occorre chiarire innanzi tutto che la competenza lessicale fa parte del gruppo più ampio delle competenze linguistiche (cfr. sopra). La competenza lessicale stessa si suddivide in due sottogruppi: elementi lessicali e grammaticali. Li riportiamo di seguito (*Common European Framework*, 2001: 110-111, traduzione nostra).

> Gli elementi lessicali includono:
>
> a) Locuzioni fisse, consistenti di più di una parola, che vanno apprese nel loro insieme. A questo gruppo appartengono:
> - formule sentenziali, come saluti, proverbi, arcaismi relitti che realizzano delle funzioni linguistiche
> - idiomi, come metafore semanticamente opache, infensificatori il cui uso è ristretto dal punto di vista contestuale e stilistico
> - strutture fisse, imparati e usati come insiemi non analizzabili in parti
> - collocazioni fisse, consistenti di parole usate spesso insieme
>
> b) Parole singole. Una certa parola può avere una serie di significati diversi. Qui appartengono le voci delle classi di parole aperte: sostantivi, verbi, aggettivi, avverbi, a volte anche serie chiuse (nomi dei giorni della settimana, nomi dei mesi dell'anno, unità di misura ecc.)

[7] Per verificare il livello di conoscenze linguistiche secondo la scala del *Quadro comune è* stato messo a punto un sistema computerizzato (www.dialang.org) che permette la valutazione della competenza linguistica in una L2 tra 14 lingue europee, tra cui l'italiano. Il sistema non solo indica che cosa è stato sbagliato dal soggetto, ma suggerisce quali sono i suoi punti di forza e di debolezza e consiglia strategie per migliorare le proprie abilità.

Gli elementi grammaticali fanno parte delle serie chiuse (almeno in inglese), qui appartengono: articoli, quantificatori, dimostrativi, pronomi personali, interrogativi e relativi.

È interessante notare che all'interno della competenza pragmatica, il *Quadro comune* richiama alla conoscenza di proverbi, detti celebri e popolari, la cui forza risiede nel loro significato traslato che rimanda a un'eredità culturale condivisa da una comunità linguistica. Pur essendo strumenti preziosi per la formazione della competenza pragmatica, in realtà spesso gli idiomi e i modi di dire popolari sono tralasciati dagli insegnanti di lingue.
La padronanza del vocabolario si misura sotto due aspetti: l'estensione e il controllo del vocabolario. Riportiamo di seguito, in inglese, per garantire una maggiore precisione, i descrittori relativi ai singoli livelli. Tralasciamo i descrittori relativi al controllo e riportiamo solo quelli dell'estensione del vocabolario (*Common European Framework*, 2001: 112):

A1	Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations. Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations.
A2	Has a sufficient vocabulary for the expression of basic communicative needs. Has a sufficient vocabulary for coping with simple survival needs.
B1	Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics.
B2	Has a good range of vocabulary for matters connected to his/her field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution.
C1	Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.
C2	Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.

Descrittori di competenze proposti dal *Quadro comune europeo di riferimento* (Tab. 6)

Sono dunque questi i criteri, formulati in modo olistico, in base ai quali gli addetti ai lavori dovranno compiere le proprie scelte. Affinché non si arrivino a risultati del tutto soggettivi, nella guida per lo sviluppo delle descrizioni referenziali (*Reference Level Descriptions for National and Regional Languages*, 2005: 6) si raccomanda, prima di altri criteri, il ricorso a indici di frequenza, con particolare riguardo ai dati ricavati da testi orali; corpora costituiti da testi prodotti da apprendenti in situazioni di valutazione. In conformità a questi criteri, per elaborare gli esponenti lessicali e grammaticali per la lingua inglese, sono presi in considerazione i dati del *Cambridge Learner Corpus*, comprensivo di 26 milioni di occorrenze. Il team dell'Università per Stranieri di Perugia, ente cui è stata affidata l'elaborazione dei materiali referenziali per l'italiano, analizzano dati autentici raccolti dalle esecuzioni di candidati agli esami CELI.
È evidente quindi, che l'elaborazione delle specificazioni, ovvero la selezione delle concrete forme lessicali e grammaticali, deve basarsi su dati oggettivi ricavati da materiali di riferimento come liste di frequenza e corpora. Non è un caso, quindi, che la competenza linguistica personale dei valutatori e docenti, come parlanti nativi o quasi nativi, è raccomandato in ultimo luogo tra i criteri da seguire.

5.2 Certificati nazionali di italiano L2

Lo scopo di questo capitolo è di esaminare dal punto di vista dei requisiti lessicali le certificazioni ufficiali di competenza attualmente esistenti per l'italiano L2. La diffusione di questi certificati serve a fissare delle soglie, anche per l'ampiezza del lessico, sotto le quali non scendere; d'altra parte, conoscere quali materiali di riferimento sono stati utilizzati per la selezione del lessico richiesto può avere risvolti concreti per la didattica.
In Italia le certificazioni di italiano L2 ufficialmente riconosciute sono attualmente quattro. La *Certificazione di Italiano come Lingua Straniera* (CILS) dell'Università degli Stranieri di Siena, il *Certificato di Conoscenza della Lingua Italiana* (CELI) dell'Università per Stranieri di Perugia e il *Certificato di Competenza della Lingua* (PLIDA) rilasciato dalla Società Dante Alighieri. Ciascuno di questi esami è articolato in 6 livelli (A1-C2), in conformità alle raccomandazioni del *Quadro comune europeo.* Il *Certificato per la Conoscenza dell'Italiano come L2* (IT) dell'Università Roma tre ha un unico livello, decisamente alto o avanzato (C1-C2). Il livello più basso è rappresentato dal *Certificato di competenza elementare in italiano come lingua straniera* (ele-IT), che misura la competenza in italiano al livello B1.
In questo capitolo ci concentriamo sui requisiti lessicali definiti per le certificazioni CILS, CELI e PLIDA.[8]

5.2.1 Certificazione di Italiano come Lingua Straniera (CILS)

La certificazione CILS, nata nel 1993, comprendeva originariamente quattro livelli, cui si sono aggiunti i due nuovi livelli CILS A1 e CILS A2. La creazione di questi nuovi livelli, che testano una competenza inferiore a quella prevista dal livello Uno, permettono che il sistema copra tutti i livelli di competenza previsti dal *Quadro comune*.

CILS A1	A1
CILS A2	A2
CILS UNO	B1
CILS DUE	B2
CILS TRE	C1
CILS QUATTRO	C2

Corrispondenza tra i livelli CILS e stadi di apprendimento del *Quadro comune*.

Per quanto riguarda la struttura e l'organizzazione linguistica dei materiali delle prove, è necessario sottolineare che i test e le prove partono sempre da un testo che costituisce il centro della comunicazione. I testi dei vari livelli hanno un grado diverso di leggibilità, controllata mediane la formula Gulpease (v. sez. 4.2) e apposite formule statistiche, diversi livelli lessicali e morfosintattici. Per il controllo lessicale si utilizzano il *Vocabolario di base della lingua italiana* (VdB) (De Mauro *et al.* 1980) e il *Lessico di frequenza dell'italiano parlato* (LIP) (De Mauro *et al.* 1993), due strumenti che permettono di "valutare realmente quale lessico usano gli italiani, sia nella lingua scritta che in quella parlata" (Barki *et al.* 2003: 121).[9]

[8] Per una storia delle certificazioni di italiano come L2 si rimanda a Vedovelli (2002b. 189-208).
[9] Si cita ancora Barki *et al.* (2003: 121): "Il LIP è la lista di frequenza più completa e rappresentativa dell'uso della lingua parlata dagli italiani e la sua importanza è da atribuirsi al fatto che 'fotografa' la realtà del parlato contemporaneo, sempre più staccata dalla lingua scritta codificata dalle grammatiche."

I risultati delle ricerche condotte presso l'Università per Stranieri di Siena relative alla "analisi delle caratteristiche delle fasi iniziali dell'apprendimento dell'italiano da parte degli stranieri" sono presentati in Barki et al. (2003: 23). Tale analisi è accompagnata sia da una proposta classificatoria per i primi due livelli di competenza (A1 e A2) fatta secondo i criteri del *Quadro comune*, sia dall'individuazione di diversi pubblici di apprendenti e da un tentativo di definizione dei sillabi relativi.
La Pre-CILS (nome riassuntivo per le certificazioni di livello A1 e A2) si articola in moduli certificatori che hanno una struttura ed un formato simili ma sono diversi per contenuti, tipologie e generi testuali, contesti comunicativi. Per ciascuno dei moduli sono stati creati sillabi e test di verifica e valutazione seguendo le medesime procedure e adottando gli stessi criteri riassunti dal *Quadro comune*. Per il momento esistono sei moduli certificatori per un totale di nove pubblici (Strambi 2004). In conformità all'approccio centrato fondato sull'apprendente ed orientato all'agire comunicativo del *Quadro comune,* gli atti comunicativi sono suddivisi in base ai quattro domini principali di comunicazione (personale, pubblico, occupazionale, educativo) ed in base ai contesti che definiscono le situazioni comunicative (luoghi dove avviene l'interazione, istituzioni, persone e oggetti coinvolti, eventi che intercorrono, azioni che si svolgono). Per ogni modulo è stata creata una tabella che individua i domini ed i contesti della comunicazione che quella specifica categoria di apprendenti, cui il modulo è destinato, riesce a gestire. Facendo riferimento alle tabelle di domini e contesti d'uso della lingua e alla descrizione o sillabo delle abilità di ciascun pubblico sono stati individuati i tipi di testi da proporre nei test del modulo ed il lessico di riferimento.

A1	La selezione dei testi d'esame si basa su un repertorio lessicale di circa 850 parole, di cui ca. 600 rientrano nel *Vocabolario di base della lingua italiana* (per l'82% nel Vocabolario fondamentale) e 250 sono raccolte in campi semantici relativi ai domini e contesti d'uso dello specifico pubblico cui il modulo è destinato. I testi potranno contenere un ulteriore 10% di parole che non fanno parte del *Vocabolario di base*, ma il cui significato risulterà desumibile dal cotesto.
A2	La selezione dei testi d'esame si basa su un repertorio lessicale di circa 1200 parole, di cui ca. 750 rientrano nel *Vocabolario di base della lingua italiana* (per l'80% nel Vocabolario fondamentale) e 450 sono raccolte in campi semantici relativi ai domini e contesti d'uso dello specifico pubblico cui il modulo è destinato. I testi potranno contenere un ulteriore 10% di parole che non fanno parte del *Vocabolario di base*, ma il cui significato risulterà desumibile dal cotesto.

Requisiti lessicali dell'esame CILS nei livelli A1-A2, ricavato da: Strambi (2001).

Nel caso di questi primi livelli, trattandosi di apprendenti non autonomi, la scelta del lessico "attivo" risulta particolarmente problematica, poiché il lessico deve essere ridotto e al contempo rappresentativo dei contesti d'uso in cui gli apprendenti potrebbero trovarsi a dover comunicare.
La tabella seguente rappresenta la comparazione del lessico generale del livello CILS A1 con il VdB:

VdB	**Numero di lemmi**	**Percentuale**
Vocabolario fondamentale	500	82,78%
Vocabolario di alto uso	65	10,76%
Vocabolario di alta disponibilità	26	4,3%
Non VdB	13	2,15%
Totale	604	100%

Lessico generale del livello CILS A1 comparato con il VdB (tratto da Barki *et al.* 2002: 122).

Il lessico che non è compreso nel VdB riguarda lemmi che recentemente sono diventati molto frequenti come *cd, cellulare, centesimo, euro, supermercato, video* ecc. pertanto

non potevano rimanere fuori dalla lista. Quanto al livello A2, il lessico "attivo" sale a 800 lemmi.
La tabella presenta la comparazione del lessico generale del livello CILS A1 con il VdB:

VdB	Numero di lemmi	Percentuale
Vocabolario fondamentale	651	81,38%
Vocabolario di alto uso	97	12,31%
Vocabolario di alta disponibilità	38	4,75%
Non VdB	14	1,75%
Totale	800	100%

Lessico generale del livello CILS A2 comparato con il VdB tratto da Barki *et al.* (2002: 122).

Le due liste dei lemmi concreti sono presentati in Barki *et al.* (2003: 123-126, 127-131). Osserviamo che i lemmi nuovi rispetto al livello A1 appartengono soprattutto alla fascia fondamentale e, in misura decrescente al vocabolario d'alto uso e di alta disponibilità.
Vediamo ora quali sono i requisiti lessicali relativi ai livelli più alti. Le indicazioni sono contenute nelle *Linee guida CILS*, pubblicate a cura di Vedovelli (2006), disponibile in formato elettronico sul sito dell'università:

B1 Il candidato deve capire il senso generale di testi contenenti parole del *Vocabolario di base della lingua italiana*, più una eventuale parte di lessico comune fino al 5%. Nella produzione orale e scritta il candidato dovrà saper usare parole del lessico fondamentale e parte del lessico di alta disponibilità.
B2 Il candidato deve sapersi orientare in testi contenenti parole del *Vocabolario di base della lingua italiana*, ma anche altri livelli di lessico (fino al 7% max.). Nella produzione orale e scritta dovrà saper usare parole del lessico fondamentale e anche parte del lessico di alta disponibilità.
C1 Il candidato deve sapersi orientare in testi contenenti parole del *Vocabolario di base della lingua italiana*, ma anche altri livelli di lessico (fino al 15% max.). Nella produzione orale e scritta dovrà dimostrare di possedere un bagaglio lessicale contenente le parole del lessico fondamentale e di alta disponibilità, e saper usare anche parole appartenenti al lessico di alto uso.
C2 Il candidato deve possedere un bagaglio lessicale molto ampio e deve dimostrare di saper utilizzare le parole in modo adeguato, pertinente e preciso. Deve essere capace di produrre sinonimi in registri linguistici differenti e identificare la collocazione sociale, culturale e storica delle parole al fine della loro adeguata utilizzazione.

Requisiti lessicali dell'esame CILS nei livelli B1-C2, ricavato da: Vedovelli (2006).

Nell'ambito dei quattro livelli (CILS UNO – QUATTRO) il repertorio lessicale presente nei testi è costituito dalle parole del VdB, più una percentuale di lessico comune che varia dal 5% del livello Uno (B1) al 15% del livello Tre (C1). Per il livello successivo non si precisa la percentuale di vocaboli del lessico comune, giacché questo livello avvicina le competenze lessicali di un parlante nativo.[10]
Quanto alle parole di cui è richiesta la conoscenza attiva, si parte da un insieme di 604 parole che nascono dall'intersezione delle tre fasce del VdB (CILS A1), per arrivare a possedere un "bagaglio lessicale molto ampio" che comprende tutto il VdB (CILS QUATTRO).

5.2.2 Certificato di Conoscenza della Lingua Italiana (CELI)

[10] Barki *et al.* (2003: 52) tengono a precisare che se il lessico presente in un testo non rientra nel sillabo previsto per quel determinato livello CILS, il suo valore deve comunque essere deducibile chiaramente dal contesto.

Gli esami CELI prevedono 5 livelli progressivi: dall'elementare all'avanzato, cui si aggiunge, in vista dei criteri del *Quadro comune*, il primissimo livello, CELI Impatto, volto a verificare una competenza inferiore a quella prevista da CELI 1. Riportiamo le corrispondenze tra i livelli CELI e quelli del *Quadro comune*.

CELI Impatto	A1
CELI 1	A2
CELI 2	B1
CELI 3	B2
CELI 4	C1
CELI 5	C2

Corrispondenza tra i livelli CELI e stadi di apprendimento del *Quadro comune*.

I materiali dell'esame sono selezionati da un gruppo di insegnanti esperti, coordinato da un team leader che cura la selezione del materiale per ogni livello d'esame. Sul sito dell'associazione ALTE (*Association of Language Testers in Europe*) (www.alte.org), di cui l'Università per Stranieri di Perugia è l'unico membro italiano, si tiene a precisare che nella selezione dei testi provenienti dalle fonti più svariate, si fa continuo riferimento alle liste di frequenza dell'italiano, tuttavia non sono date indicazioni precise.
I riferimenti relativi al livello di vocabolario previsto nei vari livelli non compaiono per niente sul sito dell'Università per Stranieri (www.unistrapg.it), il che sicuramente non agevola la preparazione agli obiettivi di apprendimento richiesti dal livello del certificato. Tuttavia in una pubblicazione apposita (Grego Bolli e Spiti 2000) si trovano indicazioni interessanti riguardo il livello di lessico previsto. Purtroppo l'opera non contempla, fra i livelli di certificazione, il livello CELI Impatto (A1), introdotto recentemente, affinché il sistema copra tutti i livelli previsti dal *Quadro comune*. Ecco i requisiti lessicali relativi ai cinque livelli dell'esame CELI:

A2	Il candidato deve essere in grado di capire testi contenenti parole del vocabolario fondamentale e di alto uso, oltre che alcune parole di maggior uso all'interno del vocabolario di alta disponibilità. Deve dimostrare inoltre di conoscere le parole di uso più comune, contenuti nel vocabolario fondamentale.
B1	Il candidato deve essere in grado di capire testi contenenti parole del vocabolario di base, più una minima percentuale di parole del lessico comune. Il candidato deve avere una sufficiente conoscenza del vocabolario fondamentale.
B2	Il candidato deve essere in grado di capire testi contenenti parole del vocabolario di base, più una parte di lessico comune fino al 10%. Il bagaglio lessicale del candidato deve comprendere, oltre al vocabolario fondamentale, anche alcune (intorno al 10%) delle parole comprese nel vocabolario di alto uso.
C1	Il candidato deve essere in grado di capire testi contenenti oltre a parole del vocabolario di base, alcune parole non tecniche, ma appartenenti alle aree di base dei lessici specialistici più comuni e più una parte di lessico comune 15%. Il repertorio lessicale del candidato deve corrispondere a tutto il vocabolario fondamentale più il 60% circa delle parole comprese nel vocabolario di alto uso.
C2	Il candidato deve essere in grado di capire, testi contenenti oltre a parole contenute nel vocabolario di base, più una parte di lessico comune fino 25%. Il candidato dovrà saper usare in maniera adeguata il vocabolario fondamentale e il vocabolario di alto uso, producendo sinonimi su registri differenti ed usando i meccanismi di formazione del lessico.

Requisiti lessicali dell'esame CELI nei vari livelli, ricavato da: Grego Bolli e Spiti (2000).

Nell'ambito dei cinque livelli (CELI 1 – 4) il repertorio lessicale presente nei testi è costituito dalle parole del VdB, più una percentuale di lessico comune che varia da "una minima percentuale" del livello 1 (B1) al 25% del livello 4 (C2). Quanto alle parole di cui è richiesta la conoscenza attiva, si parte dalle "parole di uso più comune" del vocabolario fondamentale (CILS 1, livello A2), per arrivare al vocabolario fondamentale e quello d'alto uso a livello 4 (C2).

Consultando le tabelle sui requisiti lessicali degli esami CILS e CELI, diventa evidente che ci sono ampie zone di sovrapposizione, ma anche alcune differenze sostanziali. Un essenziale punto comune è che entrambi i lavori utilizzano, come strumento di riferimento, il *Vocabolario di base della lingua italiana* (De Mauro *et al.* 1980) e in genere impiegano la terminologia lessicografica elaborata da De Mauro ed il suo team. Si pensi in particolare alla stratificazione del lessico cui le marche d'uso adottate nei vari dizionari, come il GRADIT o il *Dizionario per il terzo millennio* fanno riferimento.
In ciascuna delle due specificazioni il lessico passivo è descritto in termini di parole appartenenti al VdB e di una certa percentuale di voci appartenenti al vocabolario comune (da un minimo del 5% delle parole dei testi a livello B1 ad un massimo del 25% delle parole dei testi a livello C1). Pare quindi che non ci siano differenze sostanziali in quest'ambito tra le due certificazioni. È interessante, tuttavia, che a livello C1 il CELI prevede la conoscenza di alcune parole tecniche, sempre a patto che appartengano "alle aree di base dei lessici specialistici più comuni."
Il punto cruciale è piuttosto l'uso diverso che i due istituti fanno del VdB per determinare il lessico ad uso attivo.

	CILS	**CELI**
A2	800 lemmi, di cui 81% VF, 12% VAU, 5% VAD e 2% non VdB	le parole più comuni del VF
B1	VF + parte del VAD	VF
B2	VF + parte del VAD	VF + 10% del VAU
C1	VF + VAD + parte del VAU	VF + 60% del VAU
C2	bagaglio lessicale ampio	VF + VAU

Lessico attivo previsto dalle certificazioni CILS e CELI.

Le difformità si manifestano fin dal livello A2: mentre per sostenere il CILS è richiesta la conoscenza di un vocabolario di 800 lemmi che nascono dall'intersezione del VF, VAU e VAD e alcuni termini non compresi nel VdB, il CELI richiede l'acquisizione di un numero non specificato di parole fondamentali. Questo disinteresse nei confronti delle altre fasce del VdB si riscontra anche al livello B1, livello dell'autonomia linguistica, in cui, per il CELI, è sufficiente la conoscenza del VF, contrariamente ai requisiti del CILS, secondo cui il candidato deve saper usare anche una parte del VAD. Si osserva che nei vari livelli CILS l'importanza del vocabolario di alta disponibilità aumenta sistematicamente. Barki *et al.* (2003: 122) tengono a sottolineare che la conoscenza di questi termini, in quanto riferiti ad oggetti che un parlante nativo non nomina, ma usa quotidianamente, è di vitale importanza per chi apprende l'italiano L2. La scelta di proporre un vocabolario di alta disponibilità non è dunque casuale, ma tiene conto della sua rilevanza come parte del sapere linguistico di un parlante (Vedovelli 1995). È in quest'ottica che la conoscenza attiva del vocabolario di alta disponibilità precede l'apprendimento del vocabolario di alto uso, che entra tra le richieste per la prima volta a livello C1, mentre una conoscenza dell'intero vocabolario di alto uso è prevista solo per il livello più alto, C2.[11]
A questo punto è legittimo domandarsi se il disinteresse dimostrato dai valutatori del CELI nei confronti di questa fascia del lessico sia una scelta fondata su ragioni scientifiche o meno. È davvero realistico aggiudicare il certificato di livello C2, livello che attesta di avere competenze di un parlante quasi nativo, ad un candidato che conosca le ca. 5.000 parole del VF e VAU, ma non comprenda parole del VAD come, ad esempio, *mestolo*?

[11] Sull'importanza dei vocaboli di alta disponibilità nell'apprendimento di una L2 cfr. anche sez. 6.4.

5.2.3 Certificato di Competenza della Lingua (PLIDA)

I requisiti lessicali di sotto sono stati elaborati per sostenere l'esame PLIDA della Società Dante Alighieri di Firenze. Nel 2004 l'ente ha provveduto a creare un sillabo, a cura di Patota e Pizzoli (2004) che comprenda tutto ciò che occorre conoscere e saper fare per superare le prove d'esame dalla fase iniziale (A1, A2), passando per l'italiano fondamentale (B1, B2), fino ad arrivare ai più alti gradi di competenza previsti per l'italiano professionale (C1, C2). Il sillabo è disponibile sul sito della Società (www.ladante.it). Nel documento oltre alla struttura dell'esame, sono descritti in modo dettagliato i requisiti grammaticali e lessicali relativi ai sei livelli del PLIDA. Riportiamo di seguito le specificazioni relative al lessico:

A1	Parti del corpo – Colori – Principali capi d'abbigliamento – La famiglia – Momenti della giornata, ore, giorni della settimana, stagioni, mesi – La casa – La città (via, piazza, ecc.) – Nomi di negozi e di alcuni luoghi pubblici – Direzioni – Mezzi di trasporto – Cibi – Telefono e comunicazione telefonica.
A2	Nomi di paesi e aggettivi di nazionalità – Tempo atmosferico – Viaggio (alloggi, trasporto, ecc.) – Pesi e misure – Negozi: costo degli articoli, acquisti – Nomi di oggetti di uso quotidiano – Nomi di mestieri comuni – Abbigliamento.
B1	Nomi di animali – Attività legate alle vacanze – Attività legate al proprio ambito di studio, di lavoro o di interesse.
B2	Lessico riguardante anche argomenti non quotidiani – Lessico specialistico relativo al proprio ambito di studio, lavorativo o di interesse.
C1	Qualche conoscenza del lessico specialistico di discipline affini alla propria.
C2	Padronanza lessicale ampia e variata anche in ambiti di interesse lontani dal proprio – Qualche conoscenza dei lessici specialistici di tipo più comune (lingua del commercio e della finanza, della medicina, del diritto e della burocrazia).

Requisiti lessicali dell'esame PLIDA nei vari livelli, ricavato da: Patota e Pizzoli (2004).

Come emerge dalla tabella, la selezione dei concreti elementi lessicali avviene sulla base di temi definiti in modo generico. È opinabile se queste indicazioni sommarie siano sufficienti per garantire una selezione oggettiva dei materiali di certificazione da un lato, e di valutare le produzioni scritte e orali del candidato dall'altro. Per fare un esempio banale, si può pretendere dal candidato la comprensione della voce *filobus* a livello A1, siccome per questo grado è prevista la conoscenza dei mezzi di trasporto, oppure deve conoscere solo quei mezzi che egli stesso utilizza? Quanto al tempo atmosferico, deve conoscere parole come *precipitazione, rovescio* o *diluvio*? O ci accontentiamo della ben più generica *pioggia*? E si potrebbe continuare ancora. Pare quindi, che in assenza di un repertorio delle concrete forme lessicali, gli elenchi tematici del genere abbiano scarsa utilità pratica.

È ovvio che un buon sistema di certificazione non è tale per le sue risorse organizzative o gestionali, ma soprattutto è valido per la sua capacità di fondare la certificazione su una solida base scientifica. Ciò implica una sistematica definizione dei criteri e metodi adottati, anche in vista di un campo così difficilmente afferrabile, come il lessico.
Tra le tre certificazioni esaminate, solo il CILS e il CELI ricorrono nella definizione dei requisiti lessicali a strumenti quali i dizionari di base e di frequenza. Il PLIDA, di contro, fornisce una descrizione molto generica di quelle tematiche che il discente deve affrontare, senza scendere in dettagli riguardo il tipo e l'ampiezza del lessico da utilizzare. Tuttavia è stato visto che nemmeno l'uso degli stessi materiali di riferimento garantisce in ogni caso l'omogeneità della selezione. Per una standardizzazione delle

richieste lessicali ci vorrebbero fondamenta teoriche condivise da tutti gli enti valutatori, anche in merito alla misurazione delle competenze lessicali.
Nella nuova ottica proposta dal Consiglio d'Europa, occorre quindi rivedere non solo curricoli didattici, ma anche i sistemi di valutazione allo scopo di garantire la riconoscibilità delle conoscenze linguistiche dei cittadini europei.[12] Al momento la difformità dei criteri d'accertamento adottati dagli Stati nazionali e perfino all'interno di uno stesso Paese, rendono difficile il riconoscimento reciproco delle conoscenze. Per essere riconosciuti sul piano europeo, i diplomi e i certificati d'esame elencati nel *Portfolio* di ogni alunno, infatti, devono riportare una qualifica conforme alla scala globale data dal *Quadro comune*.

[12] In Ungheria tale processo di revisione è stato avviato nel 2005: entro il 1 gennaio 2007 ogni ente istituzionalmente autorizzato a rilasciare certificati di lingua riconosciuti dallo Stato ungherese, ha dovuto verificare la compatibilità del proprio sistema di valutazione con quello del *Quadro comune*.

6 Il lessico del manuale *Progetto italiano 1*

Il presente capitolo è dedicato all'analisi del lessico contenuto nel manuale *Progetto italiano 1* (Marin e Magnelli 2006), libro didattico per l'insegnamento dell'italiano come lingua straniera a livello elementare. La scelta del manuale si può motivare con tre fattori di natura diversa. In primo luogo, si tratta di uno dei manuali attualmente più usati nell'insegnamento dell'italiano come lingua straniera in Ungheria: è utilizzato non solo nei corsi organizzati presso varie scuole di lingua, tra cui Studio Italia, il distributore del libro per l'Ungheria, ma anche in molti licei ed istituzioni universitarie.[1] Secondo le stime del distributore stesso, G. Salusinszky, ca. due terzi degli istituti di istruzione pubblica basano i propri corsi di lingua italiana per principianti su *Progetto italiano 1.* In secondo luogo, il libro fa riferimento ai criteri stabiliti dal *Quadro comune europeo* e dalle certificazioni d'italiano: si dichiara, infatti, di rappresentare i livelli di conoscenza A1-A2, il che rende legittimo un confronto diretto tra il lessico proposto e quello previsto dagli esami di lingua. In terzo luogo, il manuale dispone di un glossario proprio, tradotto, tra l'altro, in ungherese, grazie al quale diventa possibile la comparazione dei traducenti ungheresi con uno dei lessici di frequenza della lingua ungherese. Questi motivi hanno reso *Progetto italiano 1* il manuale adatto per le nostre ricerche.

Scopo del presente capitolo è quindi di verificare in quale misura il lessico proposto nel manuale *Progetto italiano 1* corrisponde al lessico previsto come fondamentale dai dizionari di frequenza e di base. Si è preferititi restringere la ricerca sul vocabolario minimo, in altre parole sul lessico base che garantisce una comunicazione di livello elementare. Ciò in vista del fatto che questo primo livello comporta un insieme lessicale relativamente stabile e comune, mentre il lessico dei livelli più avanzati varierà necessariamente secondo le esigenze, la cultura, l'esperienza e le circostanze scolastiche e/o professionali degli allievi. Impostando la ricerca sulla prima fascia del lessico, si prevede un'ampia corrispondenza tra le parole più usuali secondo i vocabolari di frequenza e di base da un lato, e tra il lessico proposto dal manuale dall'altro.

L'altro obiettivo è di testare se l'apprendente, conoscendo il lessico del manuale, sia effettivamente in grado di superare un esame d'italiano di tipo A2, livello che il manuale dichiara di rappresentare (cfr. i livelli del *Quadro comune europeo* nella sez. 5.1.2).

Trattandosi di un manuale dedicato ad apprendenti di madrelingue varie, destinato cioè ad un pubblico internazionale, è impossibile pretendere che esso tenga conto delle difficoltà e delle necessità particolari, propri degli allievi ungheresi. E' noto però che i bisogni linguistici, e in particolare lessicali, degli allievi variano non solo secondo la loro età, professione, grado di scolarizzazione, ecc. ma anche secondo la loro nazionalità. A livello lessicale ciò significa che parole altamente disponibili per motivi socio-culturali nel Paese di provenienza dello studente non possono rimanere fuori dal suo patrimonio lessicale. Per verificare se un manuale destinato ad un pubblico plurilingue soddisfa i bisogni lessicali degli allievi ungheresi, si procede al confronto tra il lessico proposto nel

[1] Scuole di lingua che basano i propri corsi su *Progetto italiano 1* sono, a titolo d'esempio: Zilon, Bonus, Polibella, Hausschule, Peregrinus, Hajtómű, Barcelona, Métisz, M-Prospect. Tra le scuole medie superiori si menzionano il liceo Dugonics András degli Scolopi di Szeged, il liceo Árpád di Pécs, il liceo e scuola professionale Szabó Ervin, numerosi licei di Budapest ed università come la facoltà di Giurisprudenza dell'università ELTE, l'Università Tecnica, l'Istituto Superiore Heller Farkas, le università di Győr, Szeged e Veszprém ecc.

manuale e quello incluso in un vocabolario di base della lingua ungherese. In questo modo si ottengono indicazioni utili riguardo all'eventuale necessità d'integrazione del lessico proposto, in vista dei bisogni lessicali specifici degli allievi ungheresi.

6.1 Il corpus di riferimento

Progetto italiano 1 è il primo livello di un corso multimediale d'italiano, realizzato dall'editore Edilingua. Dopo la prima pubblicazione nel 2000, la nuova edizione del 2006 è frutto di una revisione. Il nuovo *Progetto italiano 1* si rivolge a studenti adolescenti e adulti fornendo circa 90-100 ore di lezione in classe, e si compone di un *Libro dello studente* e un *Quaderno degli esercizi*, articolati entrambi in 12 unità didattiche, e un cd-rom interattivo. Il manuale è integrato, fra l'altro, da supplementi in varie lingue con la traduzione del lessico e brevi note grammaticali.
Consultando la premessa del *Libro dello studente*, si nota con dispiacere che manca un rimando ai criteri utilizzati per la selezione del lessico. L'unica precisazione si riferisce al fatto che nell'edizione aggiornata c'è una maggiore coerenza tra il lessico del *Libro dello studente* e quello del *Quaderno degli esercizi.* Fortunatamente però, durante un workshop organizzato dall'editore a Budapest, abbiamo avuto l'occasione di incontrare di persona uno degli autori, Telis Marin. Siamo venuti a sapere che durante la (ri)scrittura del manuale non si è fatto uso sistematico di dizionari di frequenza o di base, ci si è affidati, invece, sul senso linguistico degli autori. Tutte le volte però, quando sono emersi dei dubbi riguardo l'inclusione di una determinata parola nel libro, si è consultato il *Lessico di frequenza dell'italiano parlato* (De Mauro *et al.* 1993, cfr. 2.2.23). A questo punto emerge una domanda: perché gli autori non hanno ritenuto questo fatto degno di essere menzionato nella premessa? La scarsa attenzione dedicata alla precisazione dei criteri di selezione, e soprattutto il mancato cenno all'uso di uno strumento di controllo così rinomato è inopportuno almeno per due motivi. Da un lato perché toglie all'opera pregi come la scientificità e l'oggettività della selezione del materiale lessicale incluso; d'altro lato perché lo strumento di verifica utilizzato, il LIP, quale lista di frequenza più recente e più rappresentativa dell'italiano parlato, può essere vista da molti utenti come una garanzia per l'attendibilità della selezione effettuata.

6.2 La realizzazione

Particolarmente felice, dal punto di vista del nostro esame, è la scelta dell'editore di pubblicare in un supplemento apposito la lista delle parole ed espressioni utilizzate nel *Libro dello studente* e nel *Quaderno degli esercizi.* La versione elettronica del glossario, disponibile on-line sul sito dell'editore (www.edilingua.it), ha costituito la base della ricerca. Dal momento però che questo file, destinato ad un pubblico internazionale, non contiene i traducenti delle espressioni, è stato necessario integrare la lista con la versione italiano-ungherese del glossario (*Olasz-magyar szószedet*), pubblicato nel 2006 a cura di Studio Italia. Questo supplemento, il cui autore stranamente non è specificato, include le parole ed espressioni che figurano nei due volumi del manuale, insieme con un breve riassunto grammaticale al termine d'ogni unità.

Come rilevato, base del corpus di riferimento è stato il file scaricabile dal sito dell'editore. Questo documento si compone di due liste: nella prima parte sono elencati i vocaboli, suddivisi per unità, volume e sezione; nella seconda parte i termini sono presentati in ordine alfabetico, con rimando all'unità, al volume e alla sezione in cui è stata usata, "in modo da poter calare il termine nel giusto contesto." La seconda è una pura lista alfabetica con rinvii alla fonte, mentre la prima ha alcune caratteristiche di un dizionario: la vocale tonica è indicata con un trattino quando la sillaba accentata non è la penultima e in tutti i casi dubbi (ad es. *farmac<u>i</u>a*); per alcuni nomi si precisa il genere e/o l'articolo; nel caso dei verbi si danno i participi passati irregolari e la prima persona singolare presente dei verbi in *-isc*; s'indicano gli avverbi; infine per alcuni vocaboli si precisano le reggenze. Le voci precedute da un asterisco, appartengono ai brani audio non stampati.
La nostra analisi si basa sulla prima lista, in cui i vocaboli sono suddivisi per unità. I motivi di scegliere questa lista come base, sono stati tre, ciascuno di carattere pratico. Il primo è che il supplemento ungherese fornisce i corrispondenti solo per la prima lista, organizzata per unità, quindi, per aggiungere a ciascun vocabolo il suo traducente è stato inevitabile partire da questa lista. Il secondo è che solo la lista per unità contiene informazioni lessicografiche (genere, articolo, parte del discorso) che permettono di disambiguare gli omonimi grammaticali. Il terzo motivo è che certe espressioni presenti nella lista per unità, non sono incluse nella lista globale. Si è rilevato che mentre la lista organizzata per unità è ricca di 3.851 termini ed espressioni, quella complessiva comprende solo 2.966 voci. Ne deriva che 885 termini figurano solo nella lista per unità; si tratta soprattutto di nomi propri, parole di origine straniera ed espressioni formate con parole elencate come entrate autonome. Ad es. mentre nella prima lista sotto unità 4 troviamo il nesso *colloquio di lavoro*, nella lista globale troviamo *colloquio* e *lavoro*, ma non l'espressione. Per i suddetti motivi abbiamo preferito prendere come punto di partenza la lista disposta per unità.

6.2.1 Reperimento dei dati

Una volta scaricato il glossario elettronico dal sito dell'editore, e dopo aver rivelato che la lista organizzata per unità è quella più adatta per la nostra ricerca, abbiamo proceduto all'immissione manuale dei traducenti ungheresi. Inizialmente pensavamo che si trattasse di un'operazione ripetitiva, ma effettuabile senza particolari problemi. Nel corso del lavoro invece abbiamo notato che non c'è una corrispondenza perfetta tra la lista disponibile on-line e quella contenuta nel supplemento ungherese. Le discrepanze riguardano da un lato l'esclusione e l'immissione d'alcuni lemmi da parte degli autori ungheresi, come la mancanza dei sostantivi *dolce, mano,* ecc. nel supplemento; d'altro lato alcune voci lemmatizzate separatamente nel glossario originale, sono state riunite in sintagmi. Il fatto che gli autori hanno evitato di lemmatizzare le preposizioni, riunendoli ogni volta al sostantivo o al verbo con cui appaiono nel contesto, è una scelta spiegabile con le peculiarità della lingua ungherese, che, come noto, non dispone di preposizioni. Ad esempio, invece di riportare autonomamente le preposizioni *a* ed *in*, come avvenuto nel glossario originale, per facilitare la traduzione, si è preferito riportarle all'interno dell'espressione in cui compaiono: *in coppia, a queste parole*.
Non tutte le soluzioni adottate sono però riconducibili a ragioni linguistiche. Oltre a una serie di errori tipografici (come l'assenza del carattere ungherese *ő* ed altri, caratteri strani nelle parole ungheresi, cfr. pp. 28, 35, 41, 42, 44) possiamo rilevare non pochi

errori di ortografia ungherese (*cappuccino* p. 5*, nyitvatartás* p. 11*, mobil telefon* p. 23*, Dél Olaszország* p. 24*, quíz* p. 27*, juice* p. 31*, Budha* p. 47*, Pó Síkság* p. 51, *Földközi tenger* p. 59 ecc.) e perfino di traduzione. Quanto ai termini grammaticali, *partitivo* in ungherese non si traduce *névmás* (p. 24), né *l'articolo partitivo* è *részelő névmás* (p. 24), così come *pronome partitivo* non è *részes névmás* (p. 49). In più, *analisi* non è *esszé* e a *tesi* non corrisponde la parola *elemzés*, bensì viceversa (p. 9); *professore* nella maggioranza dei contesti non significa *professzor* (p. 10), ma semplicemente *tanár* (del resto a *professoressa* si fa corrispondere sia *tanárnő* che *professzor asszony,* p. 26); *infanzia* significa *gyermekkor*, ma non *gyermekkorúak* (p. 26); *spuntino* non è *villásreggeli* (p. 42), *sottocultura* non è *alacsony színvonalú* (p. 58) ecc. Altre volte il significato è corretto, e la classe grammaticale è sbagliata: al sostantivo *personale* non corrisponde l'aggettivo *személyi* (p. 38), al sostantivo *comportamento* non corrisponde il verbo *viselkedni* (p. 40), all'aggettivo *psicologico* non risponde il sostantivo *pszichológus* (p. 45) ecc.

Preoccupante è anche la mancata congruenza nella rappresentazione delle informazioni grammaticali: mentre all'inizio si tende ad indicare l'articolo dei sostantivi, sia in casi dubbi (*cinema, il)* sia in quelli regolari (*alfabeto, l*), verso la fine della lista l'articolo non compare nemmeno quando a ragione lo potremmo aspettare (*condizionale, serie, sede)*. Poco sistematica, e per lo più assente è l'indicazione del genere grammaticale: per *ospedale, arte* si precisa il genere, per *amore, ascensore* no. Analogamente nell'unità 4 per i verbi si riportano, accanto all'infinito anche i participi passati irregolari (a volte nella prima, altre volte nella terza persona singolare), nelle altre unità questa pratica viene meno. Nell'unità 7, ad esempio, pur trattandosi di una forma irregolare, non si specifica il participio passato del verbo *porre*.

Dal punto di vista didattico, invece, la soluzione più particolare è quella di lemmatizzare le voci nella stessa forma in cui occorrono nel testo. Gran parte dei sostantivi e degli aggettivi è lemmatizzata al plurale e/o al femminile, senza qualsiasi contesto, e solo casualmente compare, tra parentesi, la loro forma singolare. In più, a volte la traduzione ungherese dei termini riportati al plurale è al singolare: a *rose* si fa corrispondere *rózsa, prodotti* si traduce *termék* (p. 44), *oneste* diventa *őszinte* (p. 64); altre volte le voci al plurale sono tradotte al plurale: *ricordi: emlékek* (p. 45), *esigenti: igényesek* (p. 60). La mancata coerenza nella rappresentazione può far pensare all'apprendente che la forma riportata nel glossario sia il singolare del termine (quindi si avrebbe *rose* al singolare e al plurale *rosi*, *oneste* al singolare e al plurale *onesti*). Questo procedimento è fuorviante non solo perché l'apprendente può ricavarne conclusioni errate, ma anche perché, riportando le voci al plurale, lo studente non incontra, quindi non impara la forma base delle voci, di conseguenza non sarà capace di usarle in modo corretto.

I verbi compaiono per lo più all'infinito, quelli coniugati figurano all'interno di espressioni che apparentemente hanno la funzione di illustrare il contesto d'uso (*non è che non voglio*, p. 18). Non si è riusciti a decidere se dare come corrispondente ungherese l'infinito oppure la terza persona singolare dei verbi. Accanto ai verbi riflessivi troviamo tendenzialmente quest'ultima soluzione (*coprirsi: betakarózik*, p. 64), in altri casi figura l'infinito (*sperare: remélni*, p. 19). Infine, anche l'indicazione delle reggenze verbali è affidata al caso: si ha *abituarsi (a)* ed *innamorarsi (di)*, ma non si precisano le reggenze di *arrabbiarsi, fidarsi, preoccuparsi* e di tanti altri verbi, anzi, sembra che la maggior parte delle volte la reggenza sia del tutto assente. Lo stesso discorso vale per le reggenze degli aggettivi.

Giacché l'obiettivo principale di questa ricerca non è di valutare il supplemento ungherese, terminano qui le nostre osservazioni di riguardo, notiamo, tuttavia, che la

correzione degli errori di sopra ha richiesto un impegno non trascurabile da parte di chi scrive.

6.2.2 Lemmatizzazione ed organizzazione dei dati

Successivamente all'immissione dei corrispondenti ungheresi, abbiamo proceduto alla lemmatizzazione. Per lemmatizzazione intendiamo il processo con cui abbiamo ordinato le parole ed espressioni della lista organizzata per unità in un elenco alfabetico globale. Obiettivo di questo processo è stato quello di eliminare e riunire in una lista apposita tutti i nomi propri e fonosimboli; cancellare le espressioni i cui costituenti figurano autonomamente nella lista e mantenere solo quelle che contengono elementi nuovi rispetto all'elenco; ricondurre le diverse forme flesse alla loro base; individuare e lemmatizzare come entrate autonome le espressioni polirematiche; assegnare ad ogni lemma una marca grammaticale; infine riportare accanto a ciascun lemma il suo rango d'uso nel LIP.
Prima di iniziare la lemmatizzazione, abbiamo dovuto cancellare le ripartizioni interne della lista disposta per unità. Ogni vocabolo è riportato sotto l'indicazione della sezione in cui è stata usata: per facilitare l'orientamento, infatti, gli autori hanno diviso ciascun'unità del *Libro dello studente* in sezioni, segnate con le lettere maiuscole, da A ad H, e una parte dedicata alla civiltà italiana, dal titolo *Conosciamo l'Italia*. Le sezioni di civiltà contengono anche un piccolo *Glossario*, volto a facilitare la comprensione dei testi. Dopo le parole della sezione civiltà e il glossario relativo, seguono i termini degli esercizi di autovalutazione che chiudono ogni unità, infine le parole dell'appendice grammaticale, in cui sono riportate le forme irregolari più rare. Le parole del *Quaderno degli esercizi* si suddividono in tre sezioni: voci utilizzate negli esercizi; parole dei test finali che chiudono ogni unità e termini dei test di ricapitolazione che si ripetono ogni tre unità. Le parole dei test di ricapitolazione sono elencate sempre sotto quell'unità, al termine della quale il test va eseguito. All'interno di ciascuna sezione le parole sono disposte nell'ordine della loro apparizione.

Struttura delle unità			
Libro dei testi		***Quaderno degli esercizi***	
Sezioni A-H	**L**	Esercizi relativi all'unità	**Q**
Conosciamo l'Italia		Test finale	**TF**
Glossario		Test di ricapitolazione	**TR**
Autovalutazione	**AV**	(ogni 3 unità)	
Autovalutazione generale	**AVG**	Test finale generale	**TFG**
Appendice grammaticale	**AP**		

Struttura delle unità del manuale *Progetto italiano 1* e le rispettive marche.

Per facilitare il reperimento delle parole, abbiamo conservato alcuni riferimenti: a ciascun vocabolo è stato assegnato il numero dell'unità, e le marche L o Q a seconda se è usato nel *Libro* o nel *Quaderno*; in più sono state indicate le parole che compaiono nei test di autovalutazione (AV e AVG) e nelle appendici grammaticali (AP) del *Libro* oppure nei test finale (TF e TFG) o di ricapitolazione (TR) del *Quaderno*. All'interno d'ogni unità un gruppo di parole, contrassegnate da un asterisco, è usato esclusivamente nei brani audio, queste voci hanno la marca A. Ad es. una voce usata in un brano audio dell'unità 6 è contrassegnata da 6LA.
Una volta assegnate le marche, sono stati estratti e riuniti in una lista apposita tutti i nomi propri. Sono stati cancellati i nomi propri di persona, qualora indicavano personaggi

fittizi dei dialoghi (*Maria, Jennifer, Luca, Saverio* ecc.), di contro sono stati conservati i nomi di personaggi famosi, i nomi geografici, nomi di marca, ecc. In una lista apposita abbiamo riunito anche un sottogruppo di interiezioni, i fonosimboli, privi di un corrispondente ungherese.
Il passo successivo è stato di assegnare ad ogni lemma la categoria grammaticale corrispondente. Per attribuire le categorie grammaticali, abbiamo esaminato il contesto i cui la parola è stata usata, e qualora era necessario, abbiamo fatto ricorso al *Dizionario italiano per il terzo millennio* di De Mauro (2000). Anche le espressioni polirematiche sono state classificate in base alla classe grammaticale d'appartenenza: avverbiale (*d'accordo*), congiunzionale (*come mai*), preposizionale (*rispetto a*), aggettivale (*in gamba*), interiettiva (*va bene*), pronominale (*che cosa*), sostantivale (*posta elettronica*) e verbale (*prendere il sole*). Le classi grammaticali sono state codificate con lettere secondo la tabella seguente:

Ag	aggettivo	**AgPr**	polirematica avverbiale
Art	articolo		
Avv	avverbio	**AvPr**	polirematica aggettivale
C	congiunzione	**CPr**	polirematica congiunzionale
I	interiezione	**IPr**	polirematica interiettiva
Nu	numerale		
P	pronome	**PPr**	polirematica pronominale
Pz	preposizione	**PzPr**	polirematica preposizionale
S	sostantivo	**SPr**	polirematica sostantivale
V	verbo	**VPr**	polirematica verbale

Codici delle classi grammaticali nella lista alfabetica.

Successivamente abbiamo ricondotto ogni forma alla sua base, giacché molte volte i sostantivi, verbi ed aggettivi sono stati lemmatizzati nella loro forma flessa. I sostantivi elencati nel glossario al plurale sono stati lemmatizzati al singolare, solo per i *pluralia tantum* compare a livello di lemma la forma plurale (es. *terme*); in tutti i casi dubbi è indicato l'articolo. I sostantivi alterati non sono stati ricondotti alla loro forma base. Gli aggettivi sono stati lemmatizzati sotto la forma del singolare maschile; i superlativi non sono stati ricondotti alla base. I verbi coniugati sono stati ricondotti all'infinto, i verbi ungheresi, in accordo con la prassi lessicografica, sono riportati alla terza persona singolare. Quanto agli omografi, sono stati lemmatizzati separatamente non solo quelli grammaticali, ma anche i casi d'omografia lessicale, proprio al fine di conferire una maggiore precisione all'analisi. In questo modo è diventato possibile assegnare a ciascun significato traducenti e fonti separati. Nel frattempo abbiamo notato che molte espressioni figurano più di una volta, evidentemente perché sono state lemmatizzate non solo sotto l'unità in cui sono introdotte, ma anche altre volte. Queste parole sono state riunite in modo sistematico in un'unica entrata.
Un problema speciale è quello delle espressioni polirematiche. Sulla scia di De Mauro *et al.* (1992) abbiamo adottato una definizione larga di polirematiche, considerando tali: le locuzioni il cui significato non può essere scomposto in quello dei suoi membri (*acqua in bocca*); le locuzioni il cui uso è cristallizzato (*mettere da parte*); tutte le espressioni la cui funzione grammaticale non è deducibile dalla classe grammaticale cui appartengono le parole che la formano, ma è data all'insieme dei suoi membri (*per forza*). Tendenzialmente le espressioni non idiomatiche sono state scomposte e lemmatizzate ogni volta sotto gli elementi che le compongono, mentre le polirematiche sono state conservate e riportate come entrate autonome. Nonostante che all'inizio del lavoro avessimo chiarito i criteri per identificare le polirematiche (cfr. sopra), non sempre è stato facile decidere se procedere ad una lemmatizzazione autonoma, e quindi considerare l'espressione una polirematica, oppure di dividerla in costituenti. Nei casi

dubbi ci siamo serviti del *Dizionario italiano per il terzo millennio* di De Mauro (2000) che registra e marca sistematicamente le locuzioni polirematiche. Per questo, le nostre scelte non sempre corrispondono a quelle proposte dagli autori di *Progetto italiano 1*, soprattutto perché nell'elenco proposto sono state registrate anche espressioni ad hoc, come *in coppia, a Roma, al binario, cioccolata in tazza* che, in effetti, non sono polirematiche. In base a quali criteri queste espressioni sono state lemmatizzate come entrate autonome, non è chiarito dagli autori, si può solo pensare che, trattandosi per lo più di sintagmi preposizionali, essi possano causare particolari difficoltà per lo studente, perciò gli si propone di apprendere il sostantivo insieme alla preposizione con cui nella maggioranza dei casi co-occorre.
Un'operazione assai gravosa è stata di verificare la presenza di ciascun lemma nel LIP, e di riportare per ogni entrata il rango d'uso nella lista di frequenza decrescente. Nel LIP i sostantivi ed aggettivi alterati, i superlativi e i verbi riflessivi sono registrati come forma, ma non come entrata. Per tutte queste categorie abbiamo assegnato il rango del lemma cui è stato ricondotto nel LIP.
Per la verifica della frequenza delle locuzioni abbiamo utilizzato la lista n. 5 del LIP che registra 1.933 polirematiche (pp. 532-540). Delle espressioni polirematiche il LIP non precisa il rango d'uso, fornisce solo la frequenza assoluta, quindi per le locuzioni abbiamo dato questo indice. Se il LIP è stato il primo a registrare sistematicamente tali unità, oggi disponiamo di vari altri strumenti per verificare l'effettivo uso delle espressioni polirematiche, si pensi in particolare alla serie di dizionari legati al nome di De Mauro (1999, 2000, 2002, 2003). Un ulteriore strumento di particolare utilità per l'analisi delle polirematiche è quello della società Èulogos (www.eulogos.net). Le espressioni, organizzate alfabeticamente e per rango d'uso, sono state raccolte nel corso di analisi di vari testi e corpora.[2] Abbiamo preferito ricorrere, oltre al LIP, alla lista Èulogos non solo perché l'accessibilità on line ha veramente facilitato il lavoro, ma soprattutto perché l'elenco registra un numero consistente di polirematiche non presenti nel LIP. Queste locuzioni sono state contrassegnate nell'elenco da una x.

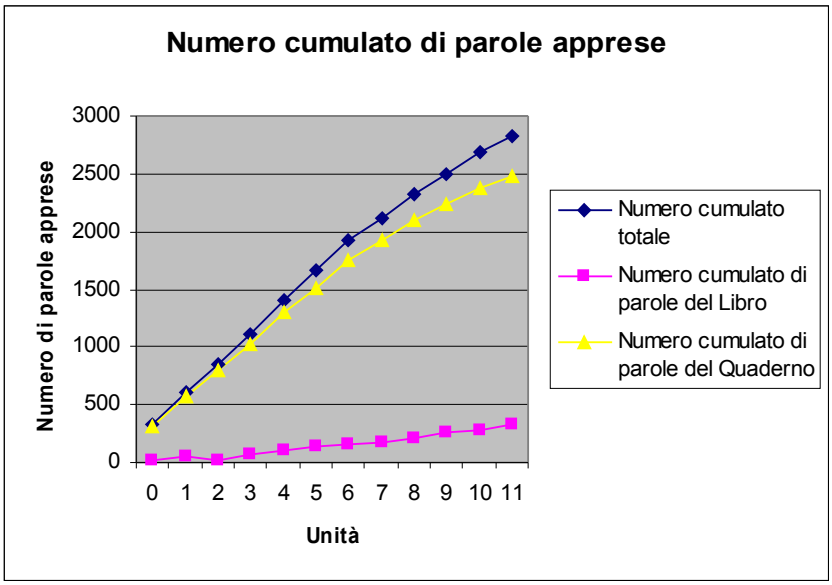
6.3 I risultati

L'ampiezza totale del lessico è di 2.819 lemmi, esclusi i nomi propri ed i fonosimboli. Notiamo che in questo numero sono inclusi anche gli omografi grammaticali e lessicali, che compaiono in due o più record con diversa categoria grammaticale e/o significato, inoltre si presentano come lemma autonomo anche le espressioni polirematiche. Lo spoglio dei due volumi del manuale ha dato origine alla seguente ripartizione: 2.457 lemmi ricavati dal *Libro dei testi*; 362 tratti dal *Quaderno degli esercizi*. Il corpus contiene anche 86 nomi propri (73 tratti dal *Libro* e 13 dal *Quaderno*), riportati in appendice, scorporati rispetto al lessico del manuale, insieme ai fonosimboli.
In questo capitolo cerchiamo di approfondire, per quanto possibile, i due aspetti legati alla quantità e alla tipologia di parole che fanno parte del lessico proposto nel manuale *Progetto italiano 1*. Proviamo ad analizzare, da un lato la dimensione e la ripartizione del lessico all'interno del manuale, dall'altro le caratteristiche tipologiche dei termini proposti.

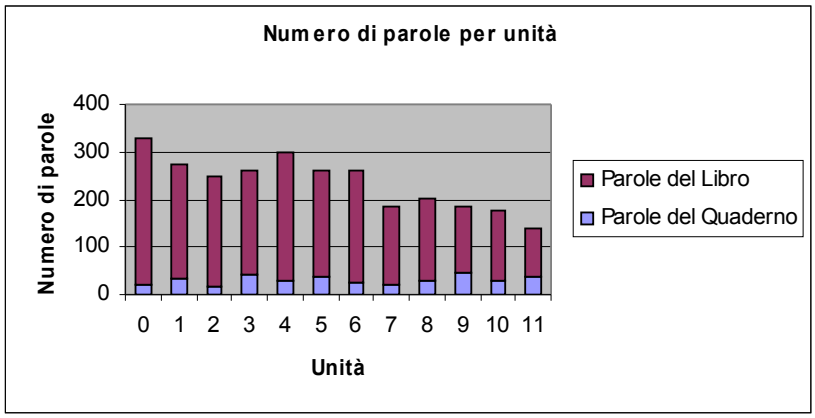
[2] Nelle tabelle è riportato anche il confronto con le liste delle polirematiche del LIP e con le liste delle polirematiche del *Lessico di Due Parole* (LDP), pubblicato in appendice a Piemontese (1998) (cfr. 4.2.2).

6.3.1 Caratteri quantitativi del lessico

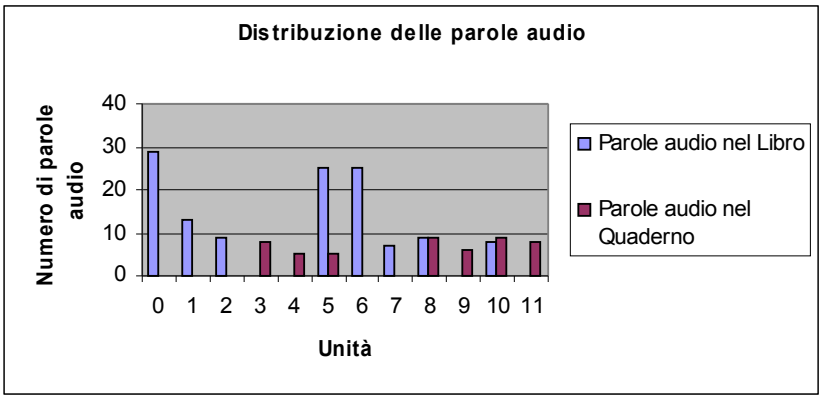
La nostra lista alfabetica globale comprende 2.585 parole e 234 espressioni. Il grafico sottostante rappresenta l'andamento dell'acquisizione del lessico proposto. Partendo dai 329 termini acquisiti nell'unità introduttiva, alla fine dell'unità 11 l'apprendente dovrebbe conoscere 2.819 parole ed espressioni. Il lessico risulta, come si legge nel grafico, distribuito relativamente bene: dopo aver superato le prime tre unità (0-2), l'apprendente conosce 850 termini (30% del totale); dopo le prime sei lezioni (0-5), cioè dopo la metà del libro, ne conosce 1669 (59%); superate le prime nove unità (0-8), sa 2.319 parole (82% del totale); arrivando alla fine del manuale a 2.819 voci (100%).



A meglio guardare, ma anche solo considerando che superata la metà del libro, l'allievo conosce più del 50% del lessico proposto (esattamente il 59%), si osserva una lieve flessione nel ritmo dell'introduzione del lessico nella seconda parte del volume, fatto provato anche dalla figura sottostante. Il grafico raffigura il numero di lemmi nuovi in ciascun'unità, indicando anche la ripartizione del lessico tra il *Libro* e il *Quaderno*.
Tenuto alto il numero di termini nuovi nelle prime sei unità, nell'unità 7 si nota un calo abbastanza notevole: il lessico nuovo qui si riduce del 30% rispetto al repertorio dell'unità 6 (187 termini proposti, contro i 262 dell'unità 6), e nelle unità successive si stabilisce attorno a 180 termini. In sintesi, mentre le unità 0-6 contengono il 68% del lessico totale, le lezioni 7-11 ne includono solo il 32%. Sembra quindi, che la maggior parte del carico lessicale sia concentrata nella prima metà del manuale. La differenza quantitativa tra il lessico dell'unità più ricca di vocaboli nuovi, quella introduttiva, contrassegnata da zero, e la lezione contenente il numero minimo di parole sconosciute, l'unità 11, è più del doppio, (329 vs. 141). Questa soluzione è contestabile in vista del fatto che nelle prime unità, e soprattutto in quella introduttiva, l'acquisizione della pronuncia, delle regole ortografiche e delle strutture grammaticali richiede uno sforzo notevole da parte dell'apprendente, perciò sembra poco ragionevole proporgli una quantità eccessiva di elementi lessicali. Una volta acquisite le strutture grammaticali fondamentali, invece, occorrerebbe porre maggior accento sull'arricchimento lessicale.



Quanto consistenti sono le differenze quantitative fra il lessico proposto nel *Libro* e nel *Quaderno*? Dai 2.819 lemmi in totale 2.457 (87,15%) sono contenuti nel *Libro* e 362 (12,85%) nel *Quaderno*. Il lessico del *Libro* risulta, come si legge nel grafico, molto più ricco di quello del *Quaderno*. Rispetto all'edizione precedente del manuale, gli autori hanno ridotto il numero di lemmi introdotti ex-novo nel *Quaderno*, proprio al fine di garantire "una maggiore coerenza tra il lessico del *Libro dello studente* e quello contenuto nel *Quaderno degli esercizi*", come recita la prefazione. In ciascuna lezione il numero di parole proposte nel *Quaderno* oscilla attorno ai 30 lemmi, in particolare esso varia dai 15 lemmi dell'unità 2 ai 45 lemmi dell'unità 9. Da un'analisi più approfondita rispetto ai dati riportati nel grafico, emerge però che la distribuzione è molto meno omogenea di quanto sembra in base ai soli numeri assoluti. Rapportando il numero di parole contenute nel *Quaderno* alla quantità di voci nuove incluse in una determinata unità, la percentuale di parole proposte nel quaderno varia dal 6% (unità 0 e 2) al 26,2% (unità 11). È chiaro che non sempre si è riusciti ad osservare il criterio di coerenza lessicale tra il *Libro* e il *Quaderno*, spiccano, sotto quest'aspetto, l'unità 9 (24,5%) e l'unità 11 (26,2), in cui il numero di parole introdotte ex-novo nel *Quaderno* raggiunge un quarto del carico lessicale totale.[3]
Vediamo ora quante parole sono contenute nei brani audio non scritti.



Si rileva che la maggior parte dei vocaboli proposti negli esercizi di comprensione auditiva si trova nel *Libro* (125 termini), mentre meno della metà è compresa nel *Quaderno* (50 voci). La distribuzione di queste parole, consigliate per l'apprendimento passivo, non è molto equilibrata, esse si concentrano per lo più nelle unità 0, 5 e 6. Anche

[3] Dal punto di vista didattico è anche discutibile quanto è opportuno introdurre vocaboli nuovi nei test finale e di ricapitolazione, volti a verificare le conoscenze acquisite nelle unità precedenti, e che, in teoria, non dovrebbero contenere materiale nuovo.

in questo caso, tra il numero maggiore di parole audio nell'unità 0 (29 termini) e il numero minimo (5 parole) nell'unità 4 vi è una differenza notevole.
Una ripartizione più omogenea dal punto di vista qualitativo potrebbe favorire l'ampliamento costante del vocabolario passivo.

Dopo le considerazioni dedicate alla questione del ritmo con cui s'introduce il nuovo vocabolario, vediamo ora come sono ripartite le categorie grammaticali. Per calcolare la distribuzione delle classi grammaticali abbiamo conteggiato anche le espressioni polirematiche, sempre sotto la categoria cui appartengono.

Categoria grammaticale	**Percentuale di lemmi nel glossario**
sostantivo	51%
aggettivo	19%
verbo	18%
avverbio	7%
altre categorie	5%

Distribuzione delle categorie grammaticali nel lessico del *Progetto italiano 1*.

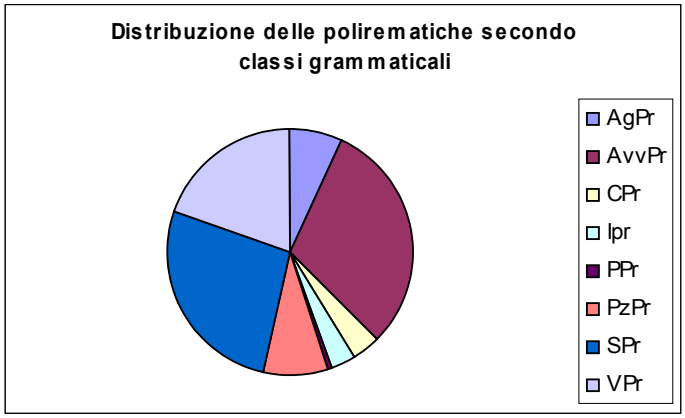
I sostantivi sono indubbiamente la categoria grammaticale più rilevante, infatti, coprono il 51% dell'intero lessico, mentre la percentuale di verbi ed aggettivi è molto simile tra loro e notevolmente superiore a quella degli avverbi. Tutte le altre classi grammaticali (congiunzioni, interiezioni, pronomi, numerali, articoli, preposizioni) costituiscono insieme circa il 5% del lessico.
La forte presenza di nomi non deve meravigliare, considerato il carattere nozionale e denotativo di questa categoria, e l'ampio uso che se ne fa nei principali costrutti scientifici in ogni ambito d'uso della lingua. Più sorprendente è l'alta presenza degli avverbi, dovuta per gran parte alla scelta di lemmatizzare come entrate autonome le espressioni polirematiche, appartenenti, per lo più, al tipo avverbiale.
È possibile confrontare la distribuzione dei lemmi nelle diverse categorie grammaticali con quella di un più ampio campione del lessico italiano, basandosi sui dati relativi al *Vocabolario di base* (De Mauro 1980) riportati in Thornton, Iacobini e Burani (1997: 68). La percentuale di lemmi appartenenti alle diverse categorie grammaticali può essere valutata sia considerando il lemmario VdB come un insieme, sia valutando la distribuzione delle categorie nelle tre fasce di frequenza in cui è suddiviso il VdB.

Categoria grammaticale	**Vocabolario fondamentale**	**Vocabolario di alto uso**	**Vocabolario di alta disponibilità**	**Vocabolario di Base**	**Lessico *Progetto italiano 1***
sostantivo	46,7%	58,6%	75,4%	60,7%	50,73%
verbo	25,8%	20,7%	12,8%	19,6%	17,75%
aggettivo	14,9%	18,2%	11,3%	14,9%	19,02%
avverbio	5,4%	1%	0,2%	2%	7,03%
pronome	2,5%	0,03%	0,04%	0,8%	1,34%
congiunzione	1,5%	0,3%	0,04%	0,6%	1,5%
preposizione	1,3%	0,2%	0,1%	0,5%	1,1%
interiezione	0,5%	0,6%	0,04%	0,4%	1,03%
numerale	1,1%	0,3%	0,2%	0,5%	0,43%
articolo	0,1%	-	-	0,02%	0,07%

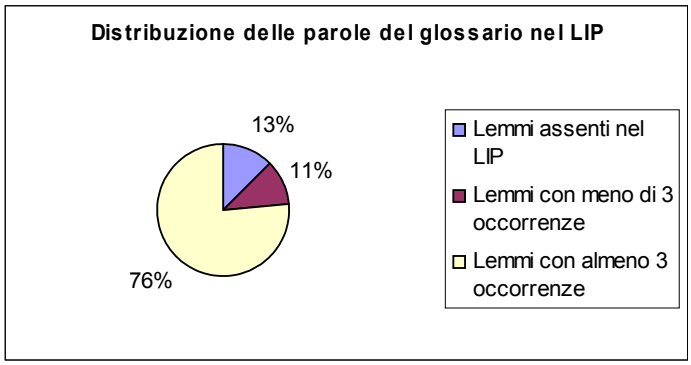
Distribuzione percentuale dei lemmi del VdB e del glossario del manuale *Progetto italiano 1*.

Come si può rilevare dalla tabella, il sostantivo è la categoria la cui presenza percentuale è nettamente superiore nei due lemmari, ma i nomi sono rappresentati in termini percentuali più nel VdB (60,7%) che nel nostro elenco (50,73%). Considerando la ripartizione interna del VdB, si osserva il tendenziale incremento dei sostantivi, con

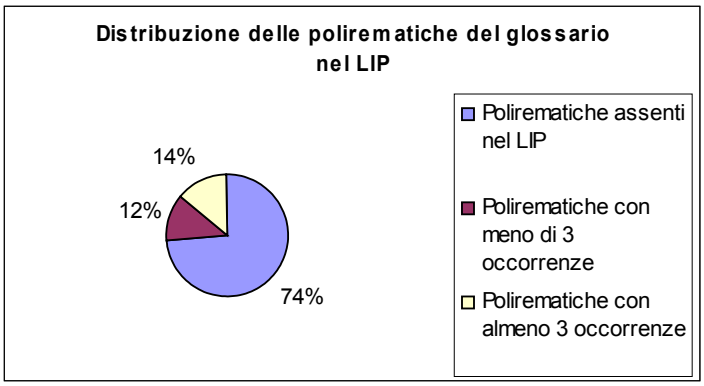
particolare riguardo alla fascia d'alta disponibilità. In base alla percentuale di sostantivi, trovata nel nostro elenco, il lessico in analisi si colloca tra la fascia del vocabolario fondamentale e quello d'alto uso.
La percentuale di verbi decresce chiaramente nelle tre fasce: nella fascia d'alta disponibilità (12,8%) è pari circa alla metà della percentuale di verbi del vocabolario fondamentale (25,8%). Nel nostro glossario la presenza di verbi si aggira su 18,2%, il che corrisponde pressappoco alle caratteristiche di distribuzione della fascia d'alto uso.
Un'importante differenza si nota fra la percentuale di aggettivi e verbi. Nel VdB i verbi (19,6%) sono più presenti degli aggettivi (14,9%), mentre nel nostro elenco la presenza degli aggettivi (19,02%) è pari circa alla percentuale di verbi (17,75%). Anche sotto quest'aspetto la nostra lista presenta caratteristiche simili a quelle del vocabolario d'alto uso, in cui gli aggettivi costituiscono il 18,2% del lemmario, contro le percentuali notevolmente più ridotte delle altre due fasce.
Quanto alla presenza di avverbi, la bassa percentuale nel VdB si spiega con il fatto che esso di regola non registra gli avverbi in *-mente* (sono presenti solo *talmente* e *solamente*), benché nel LIF siano presenti molti altri avverbi in *-mente* entro la fascia d'uso considerata per costituire il VdB. L'incidenza degli avverbi nel lessico di base italiano è quindi sottostimata nel VdB. Occorre anche ricordare che l'alta percentuale di avverbi e di interiezioni nella nostra lista è dovuta al fatto che sono stati registrati come lemmi autonomi tutte le espressioni polirematiche, di cui una parte notevole è di tipo avverbiale, come vedremo di seguito.
Come era prevedibile, la presenza percentuale delle categorie minori (pronome, congiunzione, preposizione, interiezione, numerale, articolo), già bassa nel nostro elenco, si riduce ulteriormente nel VdB: si tratta, infatti, di categorie costituite da un insieme quasi chiuso di elementi di alta frequenza, concentrate nel vocabolario fondamentale, la cui presenza percentuale diminuisce con l'incremento del numero di lemmi presi in considerazione. Sorprende l'alta incidenza delle interiezioni nel glossario. Come nel caso degli avverbi, la parte del leone è costituita dalle interiezioni polirematiche, registrate come lemma autonomo. Valutando la presenza delle categorie minori nel nostro elenco, si conferma quanto detto, cioè che in base ai dati di distribuzione, oltre che sulla base dell'estensione del lemmario, il lessico del manuale si colloca tra il vocabolario fondamentale e quello d'alto uso.
Il glossario registra 234 polirematiche in tutto, quindi circa l'8% dei lemmi è costituito da locuzioni. Le più numerose sono le polirematiche avverbiali con 72 lemmi, cui seguono nell'ordine le locuzioni sostantivali con 63 lemmi, le polirematiche verbali con 46 lemmi, le locuzioni preposizionali con 20 lemmi, le locuzioni aggettivali con 16 lemmi, le locuzioni interiettive e congiunzionali entrambe con 8 lemmi, ed una polirematica pronominale.



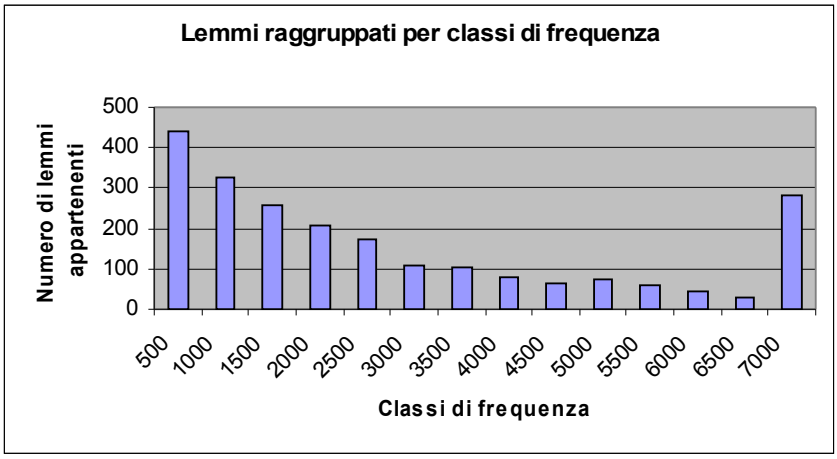
È un dato interessante che la categoria più rappresentata tra le polirematiche sia quella degli avverbi. La massiccia presenza degli avverbi è, in parte, una conseguenza delle caratteristiche morfologiche dell'italiano: per formare parti del discorso con valore avverbiale, solo in determinati casi è possibile ricorrere alla derivazione in *-mente,* e in tutti gli altri si procedono alla creazione di sintagmi avverbiali, che noi chiamiamo, appunto polirematiche. In genere, il parlato preferisce le costruzioni avverbiali analitiche alla derivazione: *a mano* invece di *manualmente*, *a lungo* e non *lungamente*, *all'inizio* invece di *inizialmente* ecc. Inoltre, ricordiamo con Vedovelli (1993: 131) che il numero elevato di avverbi è una specificità della lingua parlata, in cui questi elementi tendono ad assumere una funzione che contribuisce alla strutturazione specifica del testo parlato. Nei testi parlati gli avverbi spesso prendono una funzione discorsiva, connettiva o riempitiva, si pensi alle espressioni come: *ad un certo punto, di solito, in ogni caso, in particolare* ecc.
Seguono le tabelle che informano sulle percentuali di lessico coperte dal LIP. Il primo grafico rappresenta la percentuale di parole-lemma contenute nel LIP.



Dal grafico emerge che il 76% delle parole non polirematiche che figurano nel glossario, è compresa tra i 7.213 lemmi del LIP con almeno tre occorrenze, mentre l'11% registra una frequenza assoluta inferiore a tre e il 13% è completamente assente nel LIP. In conformità a questi dati si prevede, che circa un quarto del glossario è costituito da vocaboli la cui appartenenza alla fascia fondamentale è almeno discutibile.
Quanto alle polirematiche, la situazione si capovolge, come si legge nel grafico. Soltanto il 26% di tutte le locuzioni compare nel LIP, mentre il 74% ne è assente. Le espressioni con una frequenza assoluta superiore a tre costituiscono il 14% dell'elenco. Questo dato mostra da un lato le lacune che si riscontrano nella registrazione delle polirematiche nel LIP che, ricordiamo, è stata la prima lista di frequenza a riportare in un elenco apposito le locuzioni, d'altro lato dimostra che gli autori non hanno consultato i relativi dati del LIP.



L'ultimo grafico rappresenta la suddivisione del lessico in classi di frequenza, per dare un'idea migliore della diversa rilevanza dei singoli lemmi all'interno del glossario. Se si suddividono, adottando la metodologia seguita da Juilland (1973), Sciarone (1977) e dal LIF (1971) i lemmi più frequenti in classi di 500, in base alla loro frequenza d'uso, si vede chiaramente come il primo gruppo di 500 sia di gran lunga il più rilevante.



Nel grafico si legge appunto qual è il numero di voci del glossario, coperte dai lemmi del LIP, suddivisi in classi di 500 unità, in base alla loro frequenza d'uso. Da un'analisi approfondita risulta che i lemmi che compaiono tra i primi 500 del LIP, costituiscono il 17% dei lemmi non polirematici, d'altro canto, i lemmi che appaiono agli ultimi posti, tra 6500 e 7000, costituiscono l'11%. A quest'ultimi vanno aggiunti quei lemmi (il 13% del totale) che non compaiono affatto nel LIP. I lemmi assenti nel LIP, insieme a quelli aventi un rango più elevato, in particolare da 5000 in su, saranno oggetto della nostra analisi nella sezione successiva.

6.3.2 Caratteri qualitativi del lessico

A questo punto viene da chiedersi, quali parole compongono, concretamente, il lessico proposto nel manuale? Si tratta delle parole più note, le più utilizzate, insomma le più indispensabili? Per verificare l'effettiva rilevanza delle parole nella lingua d'uso, abbiamo condotto un confronto sistematico tra il glossario e il lemmario del LIP (De Mauro *et al.* 1993), lista di frequenza che gli autori hanno dichiarato di aver consultato in tutti i "casi dubbi."

Tra le diverse configurazioni ci sembra utile concentrare l'attenzione sulla lista che definisce il rango d'uso, vale a dire, la posizione nella lista definita dal prodotto fra la frequenza normalizzata nelle diverse liste e la dispersione nei sottocorpora. Rispetto alla lista di frequenza assoluta, l'elenco di rango d'uso permette di fare una valutazione più accurata, poiché il valore d'uso evidenzia meglio le tendenze generali del corpus complessivo (cfr. De Mauro *et al*. 1992)
Tralasciamo, nella nostra analisi, quelli che possono essere gli elementi primi, imprescindibili nell'apprendimento di una lingua. Gli articoli, le congiunzioni, i pronomi e le preposizioni sono parole funzionali che rappresentano i legami necessari all'interno di una frase, appartengono a delle classi chiuse e registrano una frequenza elevata e una distribuzione normale in quasi tutti i tipi di testi. Non abbiamo analizzato inoltre la classe, sempre chiusa, dei numerali. La nostra attenzione si è soffermata prevalentemente su sostantivi, verbi, aggettivi, avverbi, esotismi, nomi propri, interiezioni e fonosimboli e in particolare sulle espressioni poliremmatiche.

6.3.2.1 Sostantivi

La categoria grammaticale più consistente in assoluto comprende 1.366 sostantivi. Come si poteva prevedere in base alla sola numerosità di questa classe, è qui che troviamo il maggior numero di lemmi assenti nel LIP. Li elenchiamo di seguito:

> *abbonato, abitante, aceto, aggiunta, alimento, alluminio, antenna, aperitivo, archeologia, archeologo, aroma, arrosto, balcone, banana, basilica, biglietteria, birreria, bruschetta, bufala, bugia, caffelatte, caimano, calzatura, camino, camomilla, carbonara, cavatappi, cenone, cereale, ciclone, cioccolata, colapasta, coltello, combinazione, concorso, condimento, consumazione, contenitore, contorno, contro, cosmetico, costiera, cubetto, decodificatore, detersivo, dicembre, dietologo, disaccordo, disappunto, divorzio, dormita, Epifania, etto, euro, fettuccina, finezza, focaccia, furto, fusillo, gambero, giostra, giubbotto, giungla, gnomo, grattugia, guanto, idolo, incendio, insegna, involtino, ipermercato, irregolarità, lancetta, lasagna, legionario, linguina, liquore, locandina, macchinetta, mappa, marinara, mazzo, mensa, mestolo, meteo, metro, miele, miopia, modella, momento, montatura, mulino, musicista, nazista, nebbia, neorealismo, nipotino, notiziario, nuvola, nuvolosità, occhietto, ombelico, pancetta, pandoro, panetteria, panettone, paninoteca, pannacotta, passante, pescivendolo, pianista, pillola, pirateria, portacenere, postino, primo, prosciutto, pugno, radiotelevisione, raffinatezza, ragù, rammarico, rapido, regata, ricotta, rigatone, ripostiglio, ristorazione, ristretto, salame, salatura, salita, schiuma, scompartimento, scontrino, sottocultura, spumante, spuntino, stecchino, storiella, stracciatella, studentessa, supplemento, tabaccheria, tabacco, tacchino, tagliatella, tagliere, tango, tassista, tastiera, tatuaggio, telefonia, televendita, tiramisù, tortellino, tramezzino, traversa, tubetto, varietà, vasetto, viabilità, videogioco, volume.*

A questi 164 lemmi vanno aggiunti i 19 termini grammaticali o legati in ogni caso all'insegnamento che pur non essendo indispensabili per la comunicazione, di certo non possono mancare in un manuale del genere. Si tratta di:

> *abbreviazione, ausiliare, autovalutazione, congiuntivo, congiunzione, consonante, desinenza, enfasi, glossario, imperativo, imperfetto, indicativo, perifrasi, plurale, preposizione, pronome, ricapitolazione, sillaba, singolare.*

Il primo gruppo include soprattutto sostantivi concreti, legati a determinati campi lessicali classici: animali (*bufala, caimano, tacchino*), fenomeni meteorologici (*nuvola, nebbia, temporale*), vestiti (*guanto, giubbotto*), unità di misura (*etto*), professioni (*archeologo, dietologo, musicista, pescivendolo, postino, tassista*), utensili e oggetti della vita quotidiana (*cavatappi, contenitore, cosmetico, grattugia, mestolo, pillola, stecchino* ecc.). Il gruppo più folto è quello dei cibi e delle bevande (*bruschetta, caffellatte, carbonara, cioccolata, contorno, camomilla* e tanti altri).

Ci sono poi alcuni nomi, che sono entrati a far parte della nostra quotidianità proprio in quei quindici anni che sono passati dalla pubblicazione del LIP (1993). La parola *euro* ed i termini legati alle telecomunicazioni come *decodificatore, telefonia, videogioco* erano sicuramente meno usati in quell'epoca. Osserviamo tuttavia che se gli autori avessero consultato il LIP, avrebbero dovuto includere *decoder*, invece di *decodificatore*, che pur essendo un esotismo, è attestato nel LIP. Agli inizi degli anni Novanta non si parlava di *ipermercati,* nel LIP troviamo, infatti, solo *supermercato*. Nemmeno *televendita* era una delle parole più essenziali; se lo è adesso, è discutibile. In De Mauro (2000) è marcato, infatti, TS (tecnico-specialistico).
È opportuno citare alcuni casi bizzarri, tra i lemmi assenti nel LIP: di *legionario* il minimo che si può dire di rilevarne l'anacronisticità; anche la scelta di includere *lancetta* 'óramutató' è singolare; *sottocultura* è un termine tecnico dell'antropologia; *viabilità* e *traversa* sono parole di alta specificità, sostituibili con dei termini più generici e comuni, come *circolazione* e *via secondaria*. Non si capisce il motivo per cui si è deciso di usare una parola come *sottocultura*, termine tecnico dell'antropologia. Nell'unità dedicata al cinema italiano (7) si trovano termini come *caimano, ciclone, neorealismo, idolo, tango*; nell'unità in cui si parla d'abbigliamento (8) troviamo *miopia* e *montatura*; infine nell'unità in cui il tema principale è la *ristorazione*, appunto, è introdotta un'eccessiva quantità di termini legati all'alimentazione e alla cucina. Questi andrebbero rivisti anche dal punto di vista del valore d'uso. *Salatura*, ad esempio, è un termine prettamente tecnico. In alcuni casi abbiamo due termini sinonimici, di cui uno ben più usato dell'altro: *consumo* e *consumazione*, *dispiacere* e *rammarico*, *gusto* e *aroma*, *concorso* e *gara*, *proposta* e *suggerimento, succo e spremuta, marca* e *marchio*, *procedura* e *processo*, *spot* e *pubblicità, maglietta* e *t-shirt* ecc. Nel segno dell'economicità, in ogni caso uno dei due termini andrebbe cancellato, specialmente quando uno ha una frequenza notevolmente inferiore a quella dell'altra.
Quanto agli alterati, nella lista si trovano quasi solo diminutivi, in particolare tra i nomi di parentela: *fratellino, nipotino, sorellina,* ma non solo: *bacione, cubetto, cenone, cioccolatino, macchinetta, occhietto, oretta, scatoletta, storiella, tazzina, tubetto, vasetto.* Alcuni di questi li ritroviamo sulla lista dei lemmi assenti nel LIP, in quanto non riportate come forma sotto il sostantivo-lemma corrispondente.
I sostantivi del glossario sono concentrati per lo più nella fascia fino al rango 2000: la presenza è fitta in particolare fino al rango 1000, per il resto, le voci del glossario sono sparse in ogni fascia di frequenza, infatti, non sono pochi i lemmi con ranghi pari a 6505:

> *aerobica, agevolazione, aglio, alfabeto, altare, antipasto, appartamento, arancia, aranciata, arredamento, ascensore, ascesa, aspirina, attrice, avverbio, bancomat, barba, barista, bevanda, bibita, biro, biscotto, cabina, caffettiera, calamità, calciatore, cameriere, campanile, campionato, cantautore, Capodanno, cappotto, cappuccino, cappuccio, carbone, Carnevale, carriera, carrozza, cesto, chirurgo, chitarra, cioccolatino, cioccolato, clientela, commessa, completo, condizionale, conduttore, congratulazione, convalida, cornetto, cottura, crociera, cruciverba, dentifricio, dizionario, documentario, edicola, espresso, facilità, facilitazione, fantascienza, farmacia, fermata, Ferragosto, fioraio, forchetta, forno, francobollo, frigorifero, fruttivendolo, gallina, gel, genio, giornalaio, gnocco, gonna, illustrazione, incrocio, inserto, lattina, lattuga, lavanderia, leggenda, lente, libreria, lotto, macellaio, margherita, maturazione, medicinale, mittente, monaco, mortadella, mozzarella, nuoto, oceano, osteria, palestra, palio, paradiso, parentela, parmigiano, parodia, parrucchiere, passatempo, passeggero, pasticceria, patria, pecorino, penisola, pepe, pesto, pizzaiolo, pneumatico, poliziotto, pollice, portineria, presentatrice, presepe, pronuncia, proverbio, pulsante, puntata, rabbia, Rinascimento, riquadro, ritrovo, sacchetto, saldo, salmone, scaloppina, sciarpa, scrivania, serenata, settimanale, sfilata, simpatia, sinonimo, sito, soccorso, somiglianza, sostantivo, spaghetti, spremuta, stagionatura, stilista, stivale, stomaco, tacco, talento, tè, tegame, telecomunicazione, telefilm, temporale, tenore, terme, tigre, tonnellata, totocalcio, tram, trio, turco, utensile, uva, valigia, vecchietta, vignetta, virtù, viscosa, vitamina, volto, yogurt, zabaione.*

Spiccano, anche in questa categoria, i termini legati all'alimentazione: *antipasto, arancia, aranciata, gnocco, mortadella, parmigiano* ecc. Vi troviamo alcuni nomi di professione: *barista, calciatore, chirurgo, commessa, giornalaio, parrucchiere, pizzaiolo, poliziotto, presentatrice, stilista*; oggetti della vita quotidiana: *aspirina, ascensore, caffettiera, dizionario,* ecc. Particolarmente strana è la presenza di *ritrovo*, accanto ad *incontro*, *volto* invece della ben più frequente *faccia*, per non parlare di *maturazione, stagionatura*, termini tecnici dell'industria alimentare. Sorprendono poi i termini sinonimici *agevolazione* e *facilitazione*, entrambi assenti nel LIP. Non è chiaro perché il discente debba apprendere due termini marcati entrambi TS in De Mauro (2000), che peraltro in ambito non tecnico possono essere sostituiti con il ben più frequente *sconto*, del resto attestato nel manuale.
Sono questi i termini, insieme a quelli assenti nel LIP, sull'uso dei quali sarebbe utile riflettere. Rapportando la somma dei termini mancanti nel LIP e di quelli aventi una frequenza inferiore a tre, al numero totale di sostantivi del glossario, si osserva che circa il 25% dei nomi ha una frequenza bassa, vale a dire che l'inclusione di un sostantivo su quattro è discutibile.
Tornando al confronto con il LIP, della prima fascia dal rango inferiore a 500 mancano pochi lemmi, tutti d'importanza fondamentale però. Si tratta di:

> *legge, discorso, questione, politica, rispetto, dio, diritto, fondo, effetto, titolo, iniziativa, roba, maniera, valore, confronto, iniziativa.*

Della fascia successiva da 500 a 1000 ricordiamo solo le lacune più importanti:

> *lavoratore, comune, condizione, cultura, fase, ricerca, struttura, compito, materia, lettura, dibattito, associazione, capacità, istituzione, direzione, volontà, fenomeno, segnale, comunità, opinione,* ecc.

Sono tutte parole assai usate che avrebbero il diritto di comparire in un manuale per principianti.

6.3.2.2 Verbi

Confrontando sistematicamente i 453 verbi del glossario con quelli del LIP, emerge che alcuni verbi pur essendo assenti nel LIP, strumento di verifica utilizzato dagli autori, sono stati inclusi nel manuale. Si tratta di:

> *abbellire, addobbare, affettare, assaporare, digitare, drammatizzare, mascherarsi, pettinarsi, pranzare, raddoppiare, scolare, tardare, timbrare, trionfare.*

Mentre nel caso di *affettare, pranzare, pettinarsi* e forse *digitare,* nonostante la bassa frequenza d'uso, la disponibilità lessicale è abbastanza alta, perciò può essere ragionevole includerli nel manuale, per gli altri lemmi non si può dire altrettanto. I verbi *assaporare, tardare, trionfare* potrebbero essere sostituiti da altri ben più frequenti, e del resto, compresi nel glossario. Scegliendo sinonimi meno raffinati e più usuali, come *provare, fare tardi* e *vincere*, l'apprendente avrebbe la possibilità di imparare termini con un ampio ambito semantico. È chiaro come tale ricchezza di vocaboli disorienti il discente e non rinforzi l'apprendimento delle espressioni di più vasta applicabilità semantica.
Altri termini sono stati inclusi nonostante abbiano una frequenza assoluta inferiore a tre:

> *ammirare, annoiarsi, appendere, baciare, celebrare, cenare, certificare, coccolare, convalidare, differire, esportare, grattugiare, imbucare, mescolare, passeggiare, rivivere, rosolare, sbrigarsi, sospettare, stimare.*

Abbiamo elencato solamente i lemmi con un rango d'uso pari a 6505, ma se avessimo preso come limite il rango d'uso di 3000, il numero di lemmi inclusi nel glossario, la lista sarebbe stata molto più lunga. Qui appartiene anche una serie di verbi che pur non avendo una frequenza molto elevata, sono usati nei testi d'istruzione. Tra parentesi il rango d'uso: *abbinare* (3309), *commentare* (3607), *formulare* (3277), *pronunciare* (4172), *localizzare* (4081) e anche *drammatizzare* che nel LIP non figura per niente. Si citano ancora *tradire* (5623), *distrarre* (5505), *premiare* (5357), *indossare* (5008), *attrarre* (5285) *soddisfare* (4837), *trascorrere* (4728), *appassionare* (3460). Viene da chiedersi dunque, se questi verbi appartenevano ai casi "dubbi" in cui gli autori hanno consultato la lista LIP, e se la risposta è affermativa, sarebbe interessante sapere se è stato fissato un rango d'uso oltre al quale respingere le parole.
Confrontando i verbi del glossario con la fascia di più alta frequenza del LIP, in particolare con i verbi aventi un rango d'uso inferiore a 500, notiamo che, fortunatamente, mancano solo *intendere, valere* (presente però nell'espressione *vale la pena*) e *pigliare*, che invece rappresenta una varietà familiare di *prendere* e per questa sua caratteristica stilistica è ragionevole che rimanga esclusa. Nella fascia successiva, da 500 a 1000 i verbi mancanti sono:

> *intervenire, riportare, buttare, rifare, dimostrare, richiamare, definire, capitare, levare, attaccare, stabilire, comprendere, costruire, occorrere, pregare, staccare, tentare*.

Naturalmente, aumentando i ranghi d'uso, il numero di verbi assenti nel glossario cresce in maniera esponenziale, ciononostante tra i verbi appartenenti alla fascia da 1000 a 1500 a ragione potremo aspettare di trovare nel glossario:

> *approvare, consentire, derivare, fregare, dichiarare, sviluppare, fissare, basare, curare, bloccare, sommare, compiere, abbandonare, determinare, combattere, rispettare, scappare, raggiungere, domandare* ecc.

In sintesi va rilevato che oltre ad un piccolo nucleo di verbi, assenti o aventi una bassa frequenza d'uso nel LIP, il rimanente è caratterizzato da una selezione abbastanza accurata, per gran parte con verbi di alta frequenza.

6.3.2.3 Aggettivi

Gli aggettivi costituiscono, con 520 elementi la seconda classe grammaticale più numerosa nel nostro glossario. Riportiamo gli aggettivi assenti nel LIP, ma presenti nel glossario:

> *abbottonato, abituale, accogliente, adattato, affumicato, africano, altrui, amalfitano, amaro, amichevole, analfabeta, antichissimo, apprezzato, assegnato, autobiografico, cappuccino, caprese, caratterizzato, castano, cileno, concluso, confermato, coordinato, cremoso, decaffeinato, determinativo, digitale, egizio, elencato, energetico, extravergine, farcito, fenomenale, finanziato, fondato, geniale, gestito, graduale, grattugiato, gustoso, idratante, impaziente, impensabile, impersonale, importato, indeterminativo, informato, intransitivo, irlandese, irrinunciabile, italo-americano, lavorato, leggendario, londinese, luminoso, lussuoso, melodico, mite, modale, mosso, multicolore, nutritivo, nutrizionale, nuvoloso, parabolico, partitivo, pendente, post-laurea, pubblicizzato, raggiungibile, realizzabile, reclamizzato, restante, richiesto, riconosciuto, riflessivo, rinomato, ripetuto, ritmico, rivestito, rossastro, rumoroso, saporito, saracino, satellitare, singolo, sottostante, spericolato, spiritoso, stressato, terzultimo, transitivo, trascorso, tricolore, usato, vedente, veneziano.*

La lista contiene ben 97 aggettivi, da cui possiamo concludere che più del 18% degli aggettivi è assente nel LIP! Ciò significa in pratica che un aggettivo su cinque non compare nella comunicazione orale.

Alcune scelte lessicali degli autori, nonostante non siano presenti nel LIP, sono da difendere. La mancanza di *determinativo, impersonale, indeterminativo, intransitivo, modale, partitivo, riflessivo, transitivo* si deve al loro carattere nettamente grammaticale. Sono parole che pur non avendo una frequenza elevata, a ragione sono incluse in un libro didattico.
Tuttavia non mancano le scelte lessicali bizzarre. Ad esempio *satellitare* e *digitale* sono termini appartenenti al mondo delle telecomunicazioni. Altamente tecnici sono *nutrizionale, extravergine, post-laurea* che ben difficilmente uno straniero avrà occasione di usare. Ci sono poi alcuni aggettivi che mancano nel LIP per la loro ricercatezza. Uno straniero non ha un bisogno pressante di imparare, almeno nei primi stadi dell'apprendimento, parole come *irrinunciabile*, *rinomato, saporito, fenomenale* e *geniale,* anche perché li può facilmente sostituire con termini più comuni che peraltro fanno parte del glossario (*necessario, famoso, gustoso, fantastico*). Lo stesso vale per *rossastro*: ad un livello elementare il discente non ha bisogno di esprimere sfumature del genere. La mancata economicità si nota nell'inclusione di due aggettivi del tutto sinonimici, *reclamizzato* e *pubblicizzato*, di cui né l'uno né l'altro figura nel LIP. Emerge poi una serie di aggettivi denominali, derivati da nomi propri: *cileno, amalfitano, londinese, veneziano,* sulla cui inclusione in un manuale di primo apprendimento occorrerebbe ripensare, soprattutto perché lo stesso contenuto semantico può essere espresso con un sintagma preposizionale trasparente (es. *del Cile, di Londra*), nei quali il nome ha un'evidente priorità semantica rispetto all'aggettivo e, solitamente, più alta frequenza.
Altrettanto parecchi sono gli aggettivi presenti nel glossario, ed aventi nel LIP un rango d'uso pari a 6505:

> *accompagnato, affollato, agitato, agroalimentare, ammirato, argentino, australiano, austriaco, avvenuto, banale, bolognese, canadese, controllato, cotto, creato, deserto, desiderato, entusiasta, esigente, esistente, estivo, etrusco, ferroviario, forestale, geloso, grammaticale, ignoto, illuminato, indeciso, indimenticabile, infelice, influenzato, interregionale, interrotto, interurbano, irregolare, lirico, marocchino, marrone, mini, moderato, morbido, numerato, offerto, offeso, ospitale, parlato, penultimo, possessivo, preferito, proposto, protetto, raffinato, ripieno, riposato, rispettivo, roseo, scortese, situato, sorpreso, spettacolare, stressante, tardo, terrestre, ungherese, utilizzato, verticale.*

Molte assenze e frequenze basse nel LIP si trovano tra i participi passati, usati e lemmatizzati nel manuale con valore aggettivale. Anche in questo gruppo non mancano gli aggettivi denominali: *argentino, australiano, austriaco, bolognese, canadese, marocchino, ungherese*. Ripetiamo le osservazioni di sopra: è inutile caricare la memoria dell'apprendente di aggettivi che egli non ha un bisogno pressante di imparare dal momento che facilmente li può sostituire con dei sintagmi preposizionali. Le stesse considerazioni valgono per aggettivi come *estivo, forestale, terrestre.* Straordinario è l'uso del termine *agroalimentare*, marcato TS in De Mauro (2000); risentono di troppa ricercatezza stilistica i termini *ignoto, roseo* (quest'ultimo compare accanto a *rosa* e *rossastro*). È anche strano l'uso aggettivale del prefissoide *mini*.
Nonostante una parte notevole di aggettivi sia assente o abbia una bassa frequenza nella lista LIP, per gli aggettivi aventi un rango inferiore a 1000, la consistenza tra il LIP e il nostro glossario è impressionante. Mancano solo alcuni aggettivi:

> *proprio, pubblico, tale, attuale, determinato, tecnico, fisico, valido, superiore, specifico, convinto.*

La fascia successiva, dal rango 1000 a 1500, invece, avrebbe potuto contribuire notevolmente all'arricchimento della classe degli avverbi. Invece dei 90 lemmi assenti nel LIP, sarebbe stato opportuno includere i seguenti aggettivi, tutti con rango tra 1000 e 1500:

cristiano, fondamentale, unito matrimoniale, puro, concreto, santo, lontano, qualunque, corrente, pratico, teorico, zitto, definitivo, disponibile, immediato, continuo, ulteriore, complessivo, morale ecc.

È questa la categoria grammaticale i cui esponenti richiederebbero la revisione più profonda, da un lato tramite l'espunzione di gran parte dei lemmi assenti nel LIP o attestati con un basso valore d'uso, d'altro lato immettendo gli aggettivi con un rango d'uso inferiore a 2000.

6.3.2.4 Avverbi

La categoria degli avverbi è di solito una delle più problematiche, perché vi sono assegnate molte parole la cui funzione non è facilmente riconducibile ad una delle altre parti del discorso (cfr. De Mauro *et al.* 1993: 89). Nel nostro caso gli autori hanno provveduto per risolvere questo problema: l'unica classe grammaticale, indicata abbastanza sistematicamente nel glossario originale è quella degli avverbi. La relativa numerosità di questa categoria si deve al fatto che accanto ai 126 avverbi troviamo ben 72 locuzioni avverbiali, che però saranno esaminati più avanti. Per adesso si vedano gli avverbi non polirematici.
Soltanto due sono i lemmi assenti nel LIP: *dettagliatamente* e *distrattamente*, mentre gli avverbi *dove* e *quando* sono lemmatizzati come congiunzioni, quindi mancano, nella loro funzione avverbiale, nel LIP. Tra gli avverbi più usati nel corpus LIP, in particolare tra i primi 500 mancano solo *completamente, chiaramente, mo', neanche*. Di *mo'* occorre rilevare che fa parte della varietà centro-meridionale dell'italiano, e quindi è comprensibile la scelta degli autori di escluderla come voce marcata dal punto di vista geografico. Avverbi presenti nel glossario ma aventi un rango pari a 6505 nel LIP sono: *elegantemente, raramente*.
Della fascia da 500 a 1000 andrebbero inseriti *esatto, giustamente, evidentemente, effettivamente, eventualmente, nulla, normalmente*; continuando con la fascia tra 1000 e 1500 avremmo anche *immediatamente, personalmente, giusto, estremamente, perfettamente, sinceramente* ecc. Si tratta per lo più di aggettivi derivati in *-mente*. Essendo questa regola di formazione di parola completamente produttiva in italiano, non sarebbe problematica l'assenza nel manuale di numerosi aggettivi in *-mente*, se gli autori avessero richiamato l'attenzione dell'allievo sul fatto che in pratica da ogni aggettivo è possibile derivare avverbi tramite questo suffisso. Una nota del genere sarebbe auspicabile soprattutto in vista del fatto che la base aggettivale di quasi tutti gli avverbi assenti nel glossario, ma frequenti nell'uso, è inclusa nel manuale.
Considerati i risultati di sopra, si conclude che gli avverbi impiegati hanno una loro legittimità riscontrabile nel LIP e così la classe degli avverbi risulta quella più equilibrata.

6.3.2.5 Esotismi

Gli esotismi sono termini d'origine straniera, ovvero prestiti da altre lingue, marcati ES nel dizionario di riferimento usato (De Mauro 2000). Nel glossario si trovano 43 esotismi, (di cui 7 espressioni polirematiche), pari all'1,5% del totale dei lemmi. Questa percentuale sembra elevata soprattutto in vista del fatto che nel LIP gli esotismi costituiscono solo 0,3% del totale dei lemmi (De Mauro *et al.* 1993: 150).
I sostantivi sono la categoria grammaticale cui la quasi totalità degli esotismi appartiene, ad eccezione dell'aggettivo *extralarge*. La stragrande maggioranza degli esotismi è

anglicismo, mentre sporadica è la presenza di francesismi. Accanto a parole ormai ben radicate e attestate con ranghi più o meno elevati anche nel LIP, come:

> *computer* (2127), *spot* (2418), *test* (2895), *standard* (2915), *mass-media* (4825), *show* (5328), *jeans* (4024), *box* (4326), *depliant* (3039), *record* (4395), *soap opera* (con la frequenza assoluta 2),

troviamo altrettanti lemmi con frequenza assoluta inferiore a tre (rango 6505):

> *derby, fan, menù, manager, picnic, scooter, slogan, tailleur, zapping, extralarge.*

Riportiamo i termini e le locuzioni assenti nel LIP:

> *bacon, brioche, chef, comfort, designer, discount, e-mail, festival, franchising, pullover, reality, shampoo, showman, souvenir, sponsor, thriller, tournée, t-shirt, call center, internet point, self service, spaghetti western, talk show.*

È da notare che la metà delle parole ed espressioni straniere non è presente nel LIP. Tra i termini assenti troviamo da un lato voci presenti nell'italiano da un certo periodo, anche se con bassi indici d'uso (*chef, festival, shampoo*); d'altro lato ci sono termini entrati nella lingua negli ultimi quindici anni, passati dalla pubblicazione del LIP. Ad esempio, assente nel LIP, tuttavia ormai parte della vita quotidiana sono le voci *e-mail* ed *internet point,* il cui uso si è consolidato proprio in questi ultimi anni, pertanto da un lato è comprensibile che non faccia parte del LIP, d'altro lato è giusto che sia insegnato anche a livello elementare. Per il resto si tratta quasi esclusivamente di termini, la cui appartenenza al lessico italiano di base è almeno dubitabile. Ci si chiede quando mai l'allievo avrà il bisogno di dire o di capire le parole *franchising* e *discount,* marcati entrambi TS (tecnico-specialistico) in De Mauro (2000). Pur trattandosi di parole ormai diffuse a livello internazionale, la loro sfera d'uso rimane quello dell'economia e del commercio, quindi a maggior ragione questi termini dovrebbero figurare in un manuale di linguaggio economico. Considerazioni identiche valgono per la parola *sponsor*. Siccome il rispettivo verbo *sponsorizzare* è attestato nel LIP con il rango d'uso 4407, è lecito chiedersi perché il sostantivo ha avuto la precedenza. L'introduzione di altre parole (*spot, show, fan, brioche, derby, scooter, showman, souvenir, bacon, t-shirt, pullover*) è altrettanto incomprensibile in vista del fatto il loro corrispondente italiano, oltre ad avere un valore d'uso ben più elevato, per la maggior parte, è incluso anche nel glossario.
Pare, quindi, che il linguaggio utilizzato dagli autori sia modellato su quello dei mass media, in cui, sappiamo, la diffusione degli anglicismi è un fenomeno massiccio. Temiamo però che il fascino che questi anglicismi hanno esercitato sugli autori sia dovuto alla loro relativa diffusione nel parlato di alcune persone, di solito poco esperte di lingue straniere, e nello scritto di giornalisti desiderosi di stupire il lettore con l'uso di termini poco consueti. A questo punto invece emerge una domanda: non sarebbe più opportuno evitare di far imparare agli allievi termini stranieri il cui successo nella lingua italiana è assolutamente imprevedibile? Si pensi in particolare a *designer* e *reality*. Sicuramente sarebbe più utile introdurre solo quegli esotismi che sono entrati stabilmente nella lingua italiana, e quindi hanno una certa frequenza nell'uso, sempre a patto che non abbiano un loro corrispondente italiano.

6.3.2.6 Nomi propri

Fanno parte di questo gruppo i nomi di ditte e marche, i nomi geografici, le denominazioni di monumenti, personaggi famosi ecc. Lo spoglio ha dato origine a 86 nomi propri, di cui 73 tratti dal *Libro* e 13 dal *Quaderno*. Tra i 75 nomi propri del *Libro*, solo 4 sono proposti nei brani audio. La categoria più numerosa è quella dei nomi

geografici, in particolare i nomi di Paesi (18) e città (26). Sono presenti i seguenti Paesi con le rispettive capitali: l'*Italia* (anche col nome *Belpaese*), *la Francia, l'Inghilterra, la Spagna, il Portogallo, la Svizzera,* mentre senza le loro capitali sono nominate la *Germania, l'Olanda, il Belgio, l'Ungheria*, *la Russia, il Marocco*, e tra i Paesi d'oltremare *gli Stati Uniti, la Cina, l'Australia, il Brasile e l'Argentina*. Sorprende la presenza dell'*Ungheria*, che compare come unica tra i Paesi dell'Europa centro-orientale, soppiantando perfino la *Slovenia*, che pur essendo confinante con l'Italia, rimane assente. Non c'è da meravigliarsi dell'inclusione del *Marocco*, giacché la sua presenza è giustificata da evidenti ragioni socio-politiche della realtà italiana, né della *Cina*, scelta probabilmente motivata dalle enormi potenzialità economiche di questo paese. Quanto all'*Argentina* e al *Brasile*, è cosa nota che in entrambi i Paesi vive una numerosa comunità italiana, in cui l'italiano è parlato ed appreso come lingua seconda.
Considerando i nomi delle città, troviamo 8 nomi di città non italiane, tra cui oltre a quelle europee anche *Hollywood* e *Cartagine*, e 18 nomi di città italiane, di cui sono menzionati nel glossario più di una volta *Roma, Milano, Napoli, Firenze, Pisa, Perugia*. Tra i nomi geografici troviamo i mari che circondano l'Italia e *le Alpi*, ma non l'*Appennino*.
Tredici sono i nomi di personaggi di cui solo due, *Verdi* e *Boccaccio*, sono legati alla cultura italiana, del resto si tratta per lo più di personaggi storici, come *Cesare, Cleopatra, Romolo, Remo, Annibale ecc*. Abbiamo inoltre il nome di *Cristo* e della *Vergine Maria*, e in segno di un qualche universalismo religioso, anche di *Budda*.
Sono relativamente numerosi i marchionimi: accanto a *Ferrari, Aprilia, Lancia,* abbiamo *Parmigiano Reggiano, Grana Padano* e *Moka*, mentre i nomi di monumenti italiani si restringono al *Colosseo,* al *Maschio Angioino* e alla *Scala*. Tra i nomi d'istituzioni spiccano due squadre di calcio, la *Juve* e il *Milan* e certamente non può mancare l'*Unione Europea*.

6.3.2.7 Interiezioni e fonosimboli

Come rilevato Voghera nel LIP (1993: 91) le interiezioni non corrispondono ad una classe grammaticale definita morfologicamente o funzionalmente: si tratta piuttosto di segni che realizzano un atto linguistico intero, senza integrarsi in un'unità di rango maggiore. Sono, infatti, classificate come interiezioni parole o espressioni molto diverse tra loro per significato e per forma. Possono essere usate come interiezioni parole appartenenti a categorie grammaticali diverse: *prego* (verbo), *bene* (avverbio), *peccato* (sostantivo), *forte* (aggettivo). L'attribuzione della marca "interiezione" è avvenuta in ogni caso in base al valore pragmatico della parola nel contesto in cui è stata usata. La nostra lista contiene 29 interiezioni che si suddividono in 22 interiezioni semplici e 7 polirematiche interiettive, in più 3 fonosimboli. Tra le interiezioni è particolarmente numeroso il gruppo di parole ed espressioni di saluto, la cui funzione è di marcare l'inizio o la fine di un'interazione verbale: *buongiorno, ciao, salve, a stasera*. Del resto le interiezioni esprimono disappunto, dispiacere: *uffa, mannaggia, accidenti, fa schifo*; piacere: *che bello, d'accordo, va be';* sorpresa: *macché; ma dai, però*; oppure ringraziamento: *grazie, di niente, prego*.
Abbiamo inoltre isolato un piccolo sottoinsieme dei fonosimboli. Sono sequenze foniche che non hanno un significato lessicale, ma possono avere diverse letture olofrastiche. Tre sono i fonosimboli lemmatizzati nella nostra lista, trascritti come *beh, mah, ehi*. È difficile dare un traducente per queste sequenze, infatti, non sono incluse nel glossario italiano-ungherese, perciò si è deciso di individuare la loro funzione discorsiva prevalente. In mancanza di un inventario esauriente dei fonosimboli usati in italiano, si è

fatto uso della lista di frequenza dei fonosimboli, contenuta nel LIP. Si è rivelato che ciascuna delle tre sequenze è contenuta tra i 36 fonosimboli del LIP: la più frequente è *beh*, trascritta però *be'*, come segnalatore dell'inizio di un turno di conversazione, che registra 266 occorrenze nel corpus LIP; seguita da *mah*, che secondo il LIP esprime incredulità e figura 71 volte nel corpus; infine segue *ehi*, voce di richiamo o di saluto, con 9 occorrenze (cfr. LIP: 92-93, 531). Considerata l'alta incidenza dei fonosimboli nel linguaggio parlato, è lodevole il tentativo degli autori di far conoscere all'apprendente queste sequenze tanto importanti dal punto di vista pragmatico.

6.3.2.8 Polirematiche

Si è deciso di riservare una sezione a parte all'analisi delle espressioni polirematiche, giacché esse costituiscono una parte non trascurabile del vocabolario proposto. Sull'importanza delle espressioni polirematiche nell'insegnamento delle lingue hanno richiamato l'attenzione vari autori (cfr. Fóris 2002: 90, Vedovelli 1993: 119), ma l'idea più interessante tra tutti è quella di De Renzo che nel suo contributo propone di revisionare la fascia d'alta disponibilità considerando le espressioni polirematiche, che "fanno parte di quella quotidianità linguistica, che costituisce il principio su cui si fonda il vocabolario di base" (De Renzo 2005: 232).[4]
In base ad un confronto sistematico con la lista Èulogos e con il LIP (cfr. 6.3.2), si rileva che più della metà, esattamente il 57% delle locuzioni del glossario non compare in nessuna delle due liste. Ne riportiamo alcune:

> *al forno, all'arrabbiata, all'italiana, di provincia, in gamba, vero e proprio, a lungo, a pezzi, al massimo, all'italiana, alla grande, alla spina, d'altra parte, di conseguenza, da parte, da solo, in altri termini, in continuazione, in particolare, l'altro ieri, per forza, per niente, sì e no, tempo fa, a stasera, frutto di stagione, albero genealogico, cassetta per le lettere, call center, buca delle lettere, chilometro orario, codice d'avviamento fiscale, prezzo fisso, centro commerciale, chilometro orario, lavoro pubblico, protezione civile, ufficio postale, mettersi in testa, vacanza studio, treno ad alta velocità, avere intenzione, essere in giro, avere torto, mettersi insieme, self service, spaghetti western, fare il filo, stare insieme, tenere compagnia,* ecc.

Elenchiamo di seguito le locuzioni assenti nel LIP, ma presenti nella lista Èulogos:

> *alla moda, alla rinfusa, in pratica, in punto, a letto, all'interno, in base, frutto di stagione, albero genealogico, abito da sera, alta società, caffè corretto, fetta biscottata, giochi olimpici, grande schermo, occhiali da sole, occhiali da vista, pentola a pressione, participio passato, posto di lavoro, vigile del fuoco, avere paura, fare quattro passi, mandare a quel paese, prendere il sole.*

Le espressioni di quest'ultimo gruppo, pur essendo assenti nel LIP, sono state ricavate da vasti corpora, e quindi sono ben attestate nell'uso corrente, il che può legittimare il loro uso nel manuale. Lo stesso vale per una serie di termini strettamente legati alla didattica dell'italiano: *passato prossimo, periodo ipotetico, numero ordinale* ecc.
Non si può dire altrettanto per le locuzioni esotiche, assenti sia nel LIP sia nella lista Èulogos, come *talk show, internet point, call center, self service, spaghetti western*; per i termini sinonimici, di cui, in conformità con il principio dell'economicità, uno dei due andrebbe cancellato: *buca delle lettere* e *cassetta per le lettere, punto d'incontro* e *punto di ritrovo.* Singolare ci è parso il bisogno che gli autori hanno sentito di impiegare le espressioni *lavoro pubblico*, *protezione civile*, *codice di avviamento postale,* ciascuno di

[4] De Renzo (2005: 231) ha rilevato che in un campione di venti libri di lettura della scuola elementare, pubblicati tra il 1993 e il 2001, si contano 1.332 polirematiche, la cui stragrande maggioranza è indicata nel GRADIT come facente parte della fascia del vocabolario comune (CO), mentre solo due sono le polirematiche registrate come VdB. Si argomenta che queste espressioni, insieme agli esotismi diffusi negli ultimi decenni e ben note alla maggioranza degli allievi, potrebbero contribuire ad una revisione approfondita del *Vocabolario di Base*.

carattere tecnico-specialistico. Anche l'introduzione di *albero genealogico* andrebbe evitato in un libro per principianti. Risentono di ricercatezza stilistica le espressioni come *in altri termini, di conseguenza, in particolare,* che in un manuale per principianti dovrebbero essere proposti al massimo per l'apprendimento passivo, ma sicuramente non come espressioni utili per scrivere una lettera, come avviene nel manuale alla fine dell'unità 3.
Soffermiamoci sul caso di *alla rinfusa*. È da rilevare che oltre a non apparire nel LIP, tale espressione è usata raramente in italiano. Se gli autori avessero voluto suggerire un'espressione più corrente, avrebbero potuto sfruttare il più frequente *senza ordine*. Questa locuzione, se è meno caratteristica, offre però il vantaggio dell'uso di due termini aventi una ricca frequenza, mentre la voce *rinfusa* è usata, oltre alla locuzione in oggetto, solo in ambito tecnico-specialistico. Lo studente ha così memorizzato una voce che non avrà (quasi) mai occasione di usare e di sentire.
Confrontando ancora il glossario con le voci di frequenza superiore a 12 della lista LIP, si nota che la maggior parte di essi figura nel nostro glossario, mentre si reclama l'assenza di espressioni come:

> *senz'altro, punto di vista, rendersi conto, in effetti, ogni tanto, in modo che, senz'altro, a parte che, tutto sommato, d'altra parte, a volte, essere in grado, carta d'identità, dopo di che, a proposito.*

Benché l'assenza di questi termini sia un difetto importante, è lodevole la scelta degli autori di includere alcuni termini legati strettamente alla realtà quotidiana, come *vigile del fuoco, vigile urbano, centro commerciale, settimana bianca, caffè corretto.*
In sintesi, mentre per circa la metà delle espressioni non si può che approvare la scelta degli autori, l'altra metà della lista andrebbe rivista sotto il profilo della frequenza d'uso, con particolare riguardo alle locuzioni avverbiali e verbali. Uno strumento prezioso per il controllo potrebbe essere appunto la lista Èulogos, ma soprattutto il GRADIT o il *Dizionario per il terzo millennio* di De Mauro (2000) che registrano entrambi sistematicamente le locuzioni insieme alla loro marca d'uso.

6.3.2.9 Conclusioni

Riportiamo di seguito le percentuali di lemmi assenti nel LIP o aventi una frequenza d'occorrenza inferiore a tre, distribuiti secondo categorie grammaticali. Come si legge nella tabella, tra le categorie maggiori sono i sostantivi e gli aggettivi che richiederebbero la revisione più profonda.

	Sostantivi	**Aggettivi**	**Verbi**	**Avverbi**	**Esotismi**	**Polirematiche**
Percentuale di lemmi assenti nel LIP	13,4%	18,7%	3,1%	3,2%	50%	73,5% (54,7%)*
Percentuale di lemmi con frequenza < 3	12,7%	12,9%	4,4%	1,6%	26%	12,4%

*Tra parentesi è riportata la percentuale di lemmi assenti sia nel LIP che nella lista Èulogos.

Percentuale di lemmi assenti o aventi una frequenza inferiore a tre

Spiccano le percentuali di aggettivi: un aggettivo su cinque è completamente assente nel LIP, e sommandovi la percentuale di aggettivi di bassa frequenza, si ottiene che quasi ogni secondo aggettivo andrebbe riconsiderato in base al LIP. Seguendo l'identico ragionamento, ogni quarto sostantivo richiederebbe una verifica. Come era prevedibile, la maggiore corrispondenza tra i dati del LIP e i lemmi del glossario si riscontra nel caso degli avverbi e verbi, in entrambi di queste categorie la percentuale di lemmi poco frequenti è esigua. In base al LIP andrebbero eliminati tre quarti degli esotismi, mentre nel caso delle polirematiche questo valore è ancora più elevato, anche se, come è stato

osservato, esistono ormai altri strumenti di verifica che rappresentano meglio del LIP il valore d'uso delle locuzioni.
Ci pare di poter affermare che il manuale è stato realizzato più sulla base della competenza linguistica personale degli autori, e non tanto in base ai puri dati di frequenza. Sembra che l'uso che gli autori hanno fatto del LIP è stato assai sporadico, per non dire casuale.

6.4 Comparazione del glossario con il VdB

In questa sezione si leggono i risultati del confronto del nostro glossario con il *Vocabolario di base della lingua italiana* (VdB) di De Mauro *et al.* (1980). Il VdB rappresenta il più potente punto di riferimento per il lessico di base della lingua italiana, essendo tuttora l'unico dizionario di base dell'italiano che oltre al criterio della pura frequenza fa uso anche di criteri soggettivi per reperire le parole d'alta disponibilità. Come esplicitato nella sez. 2.2.15, la scelta dei lemmi del VdB è stata verificata in varie sedi, sia da parte di allievi delle elementari, che da parte di persone adulte con un ritardo mentale lieve oppure difficoltà di lettura o di scrittura. I controlli sistematici hanno confermato che il VdB costituisce un fondamento saldo del lessico dei parlanti nativi, e, nello stesso tempo, è punto di riferimento per la valutazione delle competenze lessicali in italiano L2 in sistemi di certificazione quali il CILS e il CELI (v. sez. 5.2.1, 5.2.2).
Si può ipotizzare che le unità lessicali proposte del glossario possano essere misurate in rapporto alle fasce del VdB. Se la fascia del vocabolario fondamentale (VF) è quella che copre la maggior parte dei contesti, allora le unità lessicali proposte dovranno appartenere innanzitutto a questa fascia, per fornire all'apprendente un input corretto, senza distorsioni. Si può ipotizzare, poi, che la fascia del vocabolario d'alto uso (VAU), che copre un'area di usi inferiore, dovrà essere presente in misura molto minore negli stadi iniziali del processo di apprendimento. La presenza o l'assenza di parole del VAU, infatti, dovrebbe essere il vero elemento discriminante tra i livelli di competenza lessicale. Il problema diventa complesso nel caso del vocabolario d'alta disponibilità (VAD) che, lo ricordiamo, ha una bassa frequenza d'occorrenza ed è legato ad esperienze molto familiari. Trattandosi di un contesto d'apprendimento formale (guidato), l'apprendente può elaborare il VAD se questo è presente nell'input, grazie alle scelte operate nei materiali didattici, mentre nei contesti spontanei il rischio di non occorrenza del lessico AD è maggiore. Si prevede quindi che le parole AD abbiano una presenza massiccia, paragonabile a quella del VAU.
Per il confronto abbiamo utilizzato il servizio *Censor* della società Èulogos. Il programma, disponibile sul sito http://www.eulogos.net/it/censor/, oltre ad analizzare la leggibilità dei testi (cfr. 4.2), confronta le parole del testo con i vocaboli del VdB. Per usufruire del servizio, è sufficiente mandare per e-mail il testo da analizzare in formato testo semplice, e in un paio di giorni si ricevono i dati richiesti.
Dall'analisi si ricava che il glossario comprensivo di 2.819 lemmi (parole ed espressioni) contiene 3.226 occorrenze in tutto (compresi gli eventuali articoli e le indicazioni del genere grammaticale), di cui 2.517 voci diverse. Delle 3.226 occorrenze 424, pari al 13,14% del totale sono assenti nel VdB. Rapportando il numero di parole diverse non VdB (397) al totale di voci diverse del glossario (2.517), la percentuale di parole assenti nel VdB raggiunge il 15,8%.

La tabella sottostante raffigura la ripartizione delle voci comprese nel VdB. I dati riportati si riferiscono sempre al numero di occorrenze totali. La terza colonna rappresenta la distribuzione dei lemmi appartenenti al VdB nelle tre fasce, mentre l'ultima colonna raffigura la ripartizione interna di quelle parole del glossario che fanno parte del VdB.

Livello del VdB	**Numero di parole**	**% parole**	**% parole tra le parole VdB**
Fondamentale	2040	63,24	72,81
Alto uso	563	17,45	20,09
Alta disponibilità	199	6,17	7,10
Totale parole VdB	2802	86,86	100

Distribuzione delle parole riconducibili al VdB.

Questi risultati corrispondono alle nostre attese solo in parte. Come previsto, la stragrande maggioranza dei vocaboli appartiene alla fascia del VF, tuttavia sorprende l'alta percentuale di parole AU e l'esigua presenza di parole AD. Trattandosi del primissimo livello d'apprendimento, è singolare la presenza massiccia dei vocaboli AU (attorno al 20% delle occorrenze), giacché, come detto, proprio la misura di parole AU è il fattore differenziante tra i vari livelli di competenza lessicale. Ricordiamo, infatti, che secondo i requisiti lessicali relativi al lessico attivo necessario per sostenere l'esame CILS di livello A2, la percentuale di parole AU è del 12%. I due dati si riferiscono però a due realtà diverse: i nostri dati descrivono la percentuale di termini AU di tutte le occorrenze del glossario, mentre quelli relativi all'esame CILS rappresentano la percentuale di termini AU nel lessico attivo che l'apprendente deve saper usare a livello A2. Sembra tuttavia, che la presenza del VAU nel glossario sia in una certa misura sovradimensionata, a scapito delle percentuali di parole del VF e soprattutto del VAD. Che i vocaboli del VF vadano appresi prima degli altri, è indiscutibile. Ciò che richiede una discussione più approfondita è la bassa percentuale di parole AD rispetto alle altre due fasce.

Prima di valutare questo dato, occorre fare alcune annotazioni. Le tre fasce del VdB sono spesso raffigurate come tre cerchi concentrici: il cerchio più esteriore contiene le parole fondamentali; quello intermedio include il VAU; il cerchio più piccolo comprende il VAD. In un suo contributo De Renzo (2005: 216) ha additato che, in realtà, la metafora dei cerchi concentrici (cfr. Vedovelli 1995: 62), non ritrae efficacemente il VdB. Le tre fasce, infatti, non si presentano sequenzialmente: l'acquisizione del VAD non giunge necessariamente dopo le parole fondamentali o d'alto uso. Lo stesso pensiero è stato avanzato da Tullio De Mauro in un dialogo condotto con chi scrive all'Università di Roma "La Sapienza", in occasione di un viaggio di studio: a forza di mantenere basso il numero di parole presenti nell'input didattico, bisogna sacrificare il VAU piuttosto che il VAD. Per costruire un lessico di circa cinquemila parole, De Mauro ha consigliato di utilizzare l'intera la fascia del VF e del VAD, in più le parole più usuali del VAU, selezionate in base ai dati di frequenza del LIP. Le parole disponibili fanno parte delle strutture abituali di pensiero e di comportamento, rappresentano, appunto, la parte più viva della lingua. È evidente dunque che, per chiunque voglia comunicare con i parlanti di un dato popolo, emerge la necessità di assimilare precisamente questo vocabolario, che ne costituisce il vero strumento d'espressione. Queste parole sono di vitale importanza per gli apprendenti d'italiano L2, poiché sono riferite ad oggetti che un parlante nativo non nomina, ma usa quotidianamente. L'esperienza rivela che anche persone che sarebbero tenute a conoscere un certo vocabolario, suscettibile di apparire nella lingua parlata d'ogni giorno, ignorano elementi della L2, addirittura banali nella L1. Ci si

riferisce alle situazioni quando uno non trova l'equivalente in lingua straniera di espressioni banali che nella L1 non fanno nessun problema. Per avvalermi dell'esempio di De Mauro, "uno straniero non può non conoscere la parola *peperoncino*."[5] Qualora questi termini non compaiano sulle pagine del libro didattico, l'apprendente non ha la minima possibilità di imparare i vocaboli AD.
Paragonando la percentuale di voci AD nel glossario (7,1%) alla percentuale di termini AD di cui è richiesta la conoscenza attiva per sostenere l'esame CILS (4,75%), si costata che la prima supera la seconda. Ciò tuttavia non deve ingannare sul fatto che i vocaboli AD siano troppo presenti nelle pagine del manuale, ma al contrario: per garantire che una certa parte di questa fascia entri a far parte del vocabolario attivo del discente, è indispensabile che la percentuale dei termini che vi appartengono, nell'input didattico superi la soglia prevista per l'uso attivo.
La presenza di termini AD è di particolare importanza nell'input di livello elementare. Trattandosi di fatti, esperienze e oggetti della vita di tutti i giorni, questi termini occorrono quasi esclusivamente nei livelli iniziali, mentre nei livelli superiori tendono a scomparire, man mano che i temi di vita quotidiana vengono abbandonati. Bini (1997: 18) osserva che la probabilità d'occorrenza nell'input di termini come *lavatrice, pentola, postino* diminuisce notevolmente, poiché si tratta di parole molto legate al contesto e di basso rendimento. Nello stesso tempo, le situazioni routine, in cui queste parole compaiono, sono stranote agli studenti, perciò le parole AD non colpiscono molto la loro immaginazione. Le parole AD, infatti, entrano difficilmente nel vocabolario degli studenti, indi è opportuno garantire che esse siano presentate in modo sistematico nell'input fin dai primi stadi di apprendimento. In vista dei suddetti, a livello elementare lo scopo centrale dovrebbe essere l'apprendimento della maggioranza dei vocaboli della fascia fondamentale, più un'apposita intersezione tra VAD e VAU, attribuendo particolare peso ai termini del VAD.[6]
Mentre per la selezione dei vocaboli AU il LIP può fornire un aiuto effettivo, non si può dire altrettanto per le voci AD. De Palo (1997: 396) ha reso evidente che il 73,7% del VAU è compreso nel LIP, di cui 340 lemmi compaiono addirittura tra i primi 2000 del LIP. Nel caso del VAD, invece, l'aiuto che il LIP può fornire è meno immediato: il 57,5% del VAD è assente nel LIP, del restante 42,5% che è presente, il 2,6% compare tra i primi 2000 lemmi del LIP; il 10,7% compare nella fascia da 200 a 5000; 6,3% nella fascia 5001-7000; e 22,9% nella fascia con rango superiore a 7000.[7] Gli stessi dati spiegano la ridotta presenza di termini AD nel nostro manuale in cui la selezione lessicale, almeno teoricamente, è basata sui dati di frequenza del LIP.
Consideriamo ora le parole del glossario che sono assenti nel VdB. È stato additato che il 12,09% di tutte le occorrenze e 15,1% dei termini analizzati non fa parte del VdB. Prima di procedere all'analisi di queste voci, occorre chiarire che nel VdB non sono riportati: i sostantivi alterati; la forma alterata e il superlativo di aggettivi e avverbi; gli avverbi derivati in *-mente*; gli esotismi; le espressioni polirematiche; i nomi propri e le sigle; le abbreviazioni e le forme con apocope. Mentre le abbreviazioni mancano anche nel nostro

[5] Su questo argomento v. Vedovelli e Gensini (1983: 179-196) e Vedovelli (1993: 140-147). Si vedano anche le considerazioni nostre esposte alla fine della sez. 5.2.2.
[6] Va anche rilevato che il VAD, appunto perché legato a situazioni, esperienze, ambienti e oggetti della vita quotidiana che sono suscettibili di sensibili mutamenti nel corso degli anni, è più esposto al variare della cultura materiale e richiede aggiornamenti relativamente frequenti (De Mauro 1999: XX). Dall'iniziale elaborazione del VdB, risalente al 1980, il VAD è stato sottoposto a diverse verifiche empiriche, le quali hanno portato alla configurazione attuale del 1997, contenente 2.337 parole AD.
[7] Questi dati confermano quanto additato da Domenico Russo (2005: 246), ovvero che la natura del VAD è tale da attraversare tutti gli intervalli di valore d'uso dei lemmi delle liste di frequenza, il VAD si può immaginare come "una sorta di basso continuo lessicale sotteso all'intero lemmario."

glossario, vi troviamo una quantità non trascurabile di alterati, superlativi, avverbi in *-mente*, esotismi e qualche forma con apocope, pertanto i risultati ricavati dal confronto con il VdB vanno interpretati tenendo conto di queste premesse.
Riportiamo, innanzi tutto, i sostantivi (sottolineati gli esotismi):

> *abbinamento, abbonamento, abbreviazione, accessorio, aerobica, agevolazione, aggettivo, alternativa, antipasto, appendice, archeologia, archeologo, aroma, ascesa, aspirina, ausiliare, automobilismo, autovalutazione, avverbio, bacon, bancomat, basilica, biro, birreria, box, brano, brioche, bruschetta, bufala, caffelatte, caffettiera, caimano, calamità, call, canone, carbonara, cellulare, center, centesimo, cereale, chef, ciclone, clientela, colapasta, collezione, comfort, compositore, condimento, condizionale, conduttore, congiuntivo, congratulazione, coniugazione, consumatore, contenitore, convalida, cosmetico, costiera, crociera, cruciverba, cucchiaino, decodificatore, denominazione, depliant, derby, descrizione, designer, desinenza, destinatario, destinazione, detersivo, dietologo, diminuzione, disaccordo, disappunto, discoteca, discount, docente, documentario, dollaro, duomo, e-mail, enfasi, euro, Eurostar, evento, fantascienza, fascicolo, ferragosto, fettuccina, filosofo, finezza, focaccia, foto, franchising, fruttivendolo, fusillo, gel, giornalaio, giungla, glossario, gnocco, gnomo, golf, granchè, grattugia, imperativo, imperfetto, impero, indicazione, ingrediente, inserto, Intercity, internet, involtino, ipermercato, irregolarità, jeans, lancetta, leggenda, legionario, locandina, maglietta, manager, mappa, marchio, marinara, mass, medicinale, medioevo, mensile, menù, meteo, metrò, miglioramento, miopia, modalità, montatura, mortadella, mozzarella, nazista, neorealismo, notiziario, nuvolosità, occhiale, opportunità, ordinazione, ortopedico, palio, pandoro, panettone, paninoteca, pannacotta, pantalone, parmigiano, parodia, participio, particolarità, pecorino, penisola, percentuale, perifrasi, pescivendolo, pianista, picnic, pirateria, pizzaiolo, point, possessivo, prefisso, preposizione, presentatrice, presepe, pro, pronome, pulsante, quiz, radiotelevisione, raffinatezza, rammarico, reality, regata, regia, ricapitolazione, rigatone, rinascimento, rinfusa, rinnovamento, ripostiglio, riquadro, risotto, ristorazione, salatura, scaloppina, schema, scooter, self, service, show, showman, sinonimo, sintesi, slogan, soap, sostantivo, sottocultura, souvenir, specialità, spettatore, sponsor, spot, stagionatura, statistica, stecchino, stereo, stilista, stracciatella, supermercato, supplemento, tabaccheria, tagliere, tailleur, talk, tango, tassì, tassista, tatuaggio, tecnologia, telecomando, telecomunicazione, telefonia, telegiornale, telespettatore, televendita, tematica, terme, test, testata, thriller, tiramisù, tortellino, totocalcio, tournée, tramezzino, trapassato, trattoria, trio, t-shirt, utensile, verbo, viabilità, videogioco, viscosa, western, yogurt, zabaione, zaino, zapping,.*

Come era intuibile, la maggioranza delle voci assenti nel VdB appartiene alla categoria dei sostantivi. All'interno di questo gruppo spiccano gli esotismi. Nella sezione precedente abbiamo verificato che il glossario comprende 43 esotismi in tutto, qui ne ritroviamo esattamente 37 (sottolineati). Soltanto sei sono i termini stranieri entrati a far parte del VdB: *fan* (AU), *record* (AD), *shampoo* (AD), *festival* (AD), *pullover* (AD) e *computer* (AD). Questo risultato costituisce un'ulteriore prova della necessità di rivedere e di ridurre gli esotismi contenuti nel manuale.
I sostantivi assenti nel VdB possono essere raggruppati attorno ad alcuni temi specifici: termini grammaticali (*ausiliare, enfasi, participio, trapassato*) e didattici (*ricapitolazione, autovalutazione, sintesi*); relativi all'alimentazione (*bruschetta, panettone, risotto, ristorazione, salatura, stracciatella, stagionatura*); alle telecomunicazioni (*cellulare, decodificatore, prefisso, radiotelevisione, pulsante, spot, show, telecomando, telefonia*); al commercio (*agevolazione, clientela, centesimo, franchising, ipermercato, televendita*); alle professioni (*tassista, archeologo, ortopedico, dietetico, pizzaiolo, pianista* ecc.); agli utensili (*biro, caffettiera, grattugia, tagliere*). Sono in minoranza i sostantivi astratti come *abbinamento, calamità, evento, irregolarità, leggenda, schema, sottocultura, tematica* ecc. Mentre in rapporto al LIP il 13,4% dei sostantivi del glossario è assente, confrontando il glossario con il lemmario del VdB, questa percentuale sale a 18%.
Considerate le caratteristiche di distribuzione dei due lemmari, ovvero la più alta percentuale di aggettivi nel glossario rispetto al VdB (cfr. sez. 6.3.1), è lecito attendersi che per quanto riguarda gli aggettivi, vi sia una certa discrepanza fra il VdB e il nostro glossario. Ecco gli aggettivi assenti nel VdB:

agroalimentare, amalfitano, amatissimo, amichevole, autobiografico, balsamico, benvenuto, biologico, biscottato, caprese, caratterizzato, castano, condizionato, coordinato, cremoso, cronologico, decaffeinato, determinativo, digitale, egizio, energetico, equilibrato, etnico, extralarge, extravergine, famigliare, farcito, fenomenale, finanziato, forestale, genealogico, gestito, giovanile, globale, graduale, grammaticale, grattugiato, gratuito, gustoso, idratante, igienica, impensabile, impersonale, indeterminativo, indicativo, influenzato, informale, interregionale, interurbano, intransitivo, introduttivo, ipotetico, irrinunciabile, italo-americano, leggendario, lirico, lussuoso, melodico, mini, modale, nutritivo, nutrizionale, olimpici, orale, ordinale, ospitale, parabolico, partitivo, penultimo, pessimista, post-laurea, preferibile, prestigioso, provocatorio, pubblicizzato, raggiungibile, realizzabile, reclamizzato, riflessivo, rispettivo, ritmico, saporito, saracino, satellitare, sonora, sottostante, spericolato, spettacolare, stabile, stressante, stressato, televisivo, terzultimo, transitivo, variabile.

Gli aggettivi sono di carattere grammaticale (*determinativo, indicativo, partitivo*); relativi all'alimentazione (*balsamico, biscottato, decaffeinato, farcito, extravergine, nutritivo, nutrizionale, saporito*); vi sono aggettivi denominali (*amalfitano, caprese, egizio*); infine alcuni aggettivi non specifici (*ritmico, stressante, cronologico, amichevole, ospitale, raggiungibile* ecc.). Sembra che contrariamente al VdB, i cui lemmi sono caratterizzati da un ampio spettro semantico, il glossario tende a registrare aggettivi i quali esprimono lessicalmente significati ricavabili dai vocaboli compresi nel glossario stesso. Come rilevato in precedenza, la funzione qualificativa degli aggettivi può essere espressa, oltre che da distinti lessemi, anche tramite l'impiego di sintagmi (*lussuoso – di lusso, forestale – relativo alla foresta, decaffeinato – senza caffeina*), in cui il sostantivo ha una più vasta applicabilità semantica rispetto all'aggettivo. I dati numerici danno una precisa conferma di quanto rilevato nella sezione precedente. Prendendo come punto di riferimento il LIP, il 18% degli aggettivi andrebbe eliminato; partendo dal VdB questa percentuale rimane quasi inalterata al 17%.
Sono in netta minoranza i verbi assenti nel VdB:

abbellire, abbinare, addobbare, coccolare, convalidare, digitare, drammatizzare, formulare, grattugiare, imbucare, individuare, intervistare, localizzare, motivare, pranzare, rivivere, timbrare, trionfare.

Questi verbi figurano tipicamente nei testi d'istruzione (*abbinare, drammatizzare, formulare, individuare; localizzare, motivare*); altri sono verbi generici (*addobbare, pranzare, rivivere*). Anche nel caso dei verbi trovano conferma le rilevazioni della sezione precedente: la percentuale di verbi assenti nel LIP si aggira sul 3%, nel caso del VdB è del 3,6%.
Seguono infine i pochi avverbi assenti nel VdB:

dettagliatamente, distrattamente, gratuitamente, oralmente.

La percentuale di avverbi assenti nel VdB corrisponde grosso modo a quella di voci non comprese nel LIP, fornendo un'ulteriore prova della fondatezza della selezione operata nel campo degli avverbi.
È sporadica l'assenza di altre categorie nel VdB. Ricordiamo le interiezioni *arrivederla, mannaggia*; i numerali *miglialio, ventina*; il pronome *altrui*; la preposizione *tramite*.
Riportiamo in una tabella apposita la percentuale di lemmi assenti nel LIP e nel VdB, ripartite per categoria grammaticale (del conteggio sono escluse le polirematiche e le categorie minori):

	Percentuale di lemmi assenti			
	Sostantivi	Aggettivi	Verbi	Avverbi
LIP	13,4%	18,7%	3,1%	3,2%
VdB	17,9%	17,7%	3,6%	2%

Percentuale di lemmi assenti nel LIP e nel VdB, ripartite secondo categoria grammaticale

Tutto questo conferma non solo a livello delle singole voci, ma anche sotto una prospettiva più ampia quanto osservato nella sezione precedente, ossia che le due categorie che più necessitano di un controllo approfondito, sono i sostantivi e gli aggettivi, mentre per le altre classi grammaticali le voci proposte sono effettivamente quelle più fondamentali.
Basare le scelte lessicali di un manuale sulla sensibilità linguistica degli autori ha il pregio di permettere l'inclusione di una serie di parole che pur non avendo un'altissima frequenza nei testi scritti e parlati, sono ben note a tutti gli italiani, appartengono cioè alla classe delle parole d'alta disponibilità, come vedremo di seguito. D'altra parte però, a questa sensibilità per le conoscenze lessicali dei parlanti nativi non si è associato il tentativo di seguire, almeno in questa primissima fase dell'apprendimento, il criterio del "minimo sforzo", che impone di selezionare i termini d'impiego più frequente e di più vasta applicabilità semantica che meglio si combinano con altre parole. Non si è riusciti, infatti, ad escludere del tutto i tecnicismi, gli esotismi ed i termini di bassa frequenza. Accanto ad una notevole quantità di termini generici nel glossario si trovano voci d'alta preziosità linguistica, ben sostituibili con perifrasi o altre parole più semplici e più usati. Sotto quest'aspetto andrebbero rivisti in particolare gli aggettivi e i sostantivi, con particolare riguardo a quelli di origine straniera.

6.5 Grado di copertura del lessico

Per verificare l'effettiva utilità del manuale *Progetto italiano 1* per la preparazione di un esame di lingua di livello A2, abbiamo controllato il grado di copertura dei lemmi del glossario rispetto ad una prova di lettura calibrato a livello A2.[8] Ricordiamo, che questo è il livello che gli studenti, secondo l'intenzione degli autori di *Progetto italiano 1*, dovrebbero padroneggiare dopo aver imparato le strutture contenute nel libro. Scopo del nostro conteggio è di vedere, quante fra le parole utilizzate nella prova di lettura sono contenute nel glossario, e di cercare di verificare se l'allievo in possesso del lessico offertogli come input sia effettivamente in grado di affrontare l'esame. Nella sez. 1.1.4 abbiamo richiamato l'attenzione sul fatto che il grado di copertura di un lessico non va identificato con la sua utilità comunicativa. In quest'ottica precisiamo che l'obiettivo del conteggio è di verificare il grado di coerenza tra l'input didattico e i requisiti lessicali in sede di certificazione, senza voler misurare l'efficienza comunicativa del lessico proposto.
Per la verifica abbiamo scelto l'esame CELI 1 dell'Università per Stranieri di Perugia, livello che corrisponde ad una conoscenza di grado A2 della lingua italiana. Le capacità richieste ad un candidato del CELI 1 si basano sulle capacità tipiche del livello A2 del *Quadro comune di riferimento europeo*: "il candidato deve aver sviluppato le abilità linguistiche necessarie per la 'sopravvivenza' cioè per poter iniziare ad interagire nella vita sociale e lavorativa in un contesto dove ci si esprima in italiano."[9] Così come gli altri esami CELI, anche il CELI 1 copre le quattro principali abilità linguistiche: leggere, scrivere, ascoltare, parlare. Per i nostri obiettivi abbiamo ritenuto opportuno prendere

[8] Per copertura s'intende la percentuale d'occorrenze che sono contenute nel glossario.
[9] Ricavato dal sito del Centro per la Valutazione e la Certificazione Linguistica dell'Università per Stranieri di Perugia (www.cvcl.it).

come testo di riferimento gli esercizi di comprensione scritta. Per i requisiti lessicali relativi al livello A2 si rinvia alla sez. 5.2.2.
I testi utilizzati per la prova di lettura sono in genere brevi, ad es. avvisi pubblici, istruzioni, messaggi pubblicitari di interesse generale. I candidati devono "dimostrare di saper comprendere il significato generale e alcuni dettagli di tipo pratico." Il vocabolario di questi testi comprende "alcuni internazionalismi e le parole di più alta frequenza." La prova include cinque brevi testi, ciascuno accompagnato da un test di scelta multipla. I testi sui quali abbiamo calcolato il grado di copertura del glossario sono state tratte da un cd pubblicato dal Consiglio d'Europa (Sheils 2005), che registra prove d'esame calibrate secondo i livelli del *Quadro comune di riferimento europeo.*
Per motivi di spazio abbiamo sottoposto alla verifica ogni secondo esercizio, cioè gli item n. 1, 3 e 5. Proponiamo i testi in due versioni. Nella prima versione abbiamo tolto tutte le parole tranne quelle comprese nel glossario del manuale *Progetto italiano 1*; nella seconda è riportato il testo integrale. Accanto ai testi presentiamo le rispettive domande.

Esercizio n. 1

> Elio: "Ricordo la prima vacanza a 16 anni, in giro per le isole Eolie: per me, che sono di Messina, non è stato un grande viaggio. Ma allora mi è * di essere arrivato in America. Con tre amici (anche una *) abbiamo girato, abbiamo anche dormito in *, abbiamo vissuto in un'atmosfera *! Tanti anni dopo sono tornato là a cercare i sapori e i colori di quella vacanza, ma non ho più * quell'atmosfera."
>
> Elio:
> A) descrive una vacanza in *
> B) ricorda le * di una vacanza
> C) parla di un viaggio tra Messina e l'America

La soluzione del primo esercizio richiede la conoscenza di pochi vocaboli non ancora presentati all'apprendente: il grado di copertura del testo e delle affermazioni è infatti del 93%. La questione principale però è se conoscendo il 93% dei termini lo studente è in grado di capire l'argomento, e di selezionare la risposta giusta. Il testo ha un tema molto generale: il ricordo di una vacanza, quindi per la sua comprensione basta conoscere alcune parole d'alta frequenza, come *viaggio, campeggio, tenda, emozione, atmosfera*. Anche se di questi elementi solo il primo e l'ultimo figurano nel glossario, sembra che lo studente abbia buone probabilità di trovare la risposta giusta (B). Una volta capite le parole *vacanza, viaggio, girare* da un lato e *atmosfera, sapore, colore* dall'altro, si intuisce che qui si tratta appunto delle emozioni che quella vacanza ha suscitato. Ecco il testo integrale.

> Elio: "Ricordo la prima vacanza a 16 anni, in giro per le isole Eolie: per me, che sono di Messina, non è stato un grande viaggio. Ma allora mi è sembrato di essere arrivato in America. Con tre amici (anche una fidanzatina) abbiamo girato, abbiamo anche dormito in tenda, abbiamo vissuto in un'atmosfera elettrizzante! Tanti anni dopo sono tornato là a cercare i sapori e i colori di quella vacanza, ma non ho più respirato quell'atmosfera".
>
> Elio:
> A) descrive una vacanza in campeggio
> B) ricorda le emozioni di una vacanza
> C) parla di un viaggio tra Messina e l'America

Esercizio n. 3

> Sono *: le tenete sotto la pioggia poi le * e sono buone come prima. Si aprono e si * con facilità. Stanno bene in *. Le * turistiche della De Agostini segnano un bel passo avanti nella * di queste * eppure grandi compagne dei nostri *.
>
> Il testo descrive:
> A) un abito *
> B) una cartina turistica
> C) un libro *

La copertura è dell'87%, ma è evidente che le parole sconosciute sono essenziali per capire il messaggio del testo. Per risolvere l'esercizio con successo, l'allievo deve attivare non solo le proprie conoscenze lessicali, bensì anche quelle culturali. Deve essere in grado di dedurre il significato dal contesto senza capire tutte le parole chiave, portatrici del messaggio del testo: *impermeabile, tasca, pianta, compagna, tascabile*. Il nome dell'editore De Agostini può essere d'aiuto, sempre a patto però, che lo studente riconosca che si tratta di un editore. E' tuttavia dubitabile che un'apprendente di livello elementare che studia l'italiano nel proprio Paese, abbia molte possibilità di far conoscenza dell'editoria italiana. Mettiamo pure che l'allievo riconosce il nome. A questo punto gli rimangono sempre due scelte, B e C, giacché questa casa editrice pubblica sia libri che cartine. Per selezionare la risposta giusta (B) è, infatti, indispensabile conoscere oltre a *cartina*, anche il suo sinonimo che figura nel testo, *pianta*. Senza capire *pianta*, ed è questo il nostro caso, l'unica chance dell'allievo è di intuire, in base a parole come *turistico* e *pioggia*, che qui non si tratta di libri, ma di mappe. Segue il testo integrale.

> Sono impermeabili: le tenete sotto la pioggia poi le scrollate e sono buone come prima. Si aprono e si ripiegano con facilità. Stanno bene in tasca. Le piante turistiche della De Agostini segnano un bel passo avanti nella concezione di queste umili eppure grandi compagne dei nostri spostamenti.
>
> Il testo descrive:
> A) un abito impermeabile
> B) una cartina turistica
> C) un libro tascabile

Esercizio n. 5

> Per decisione dell'* europea, i nuovi * di * in euro sono diversi da quelli in * per colore e disegno. I vecchi * in * devono essere consegnati in banca o distrutti. E' * anche correggere i vecchi * in * e trasformarli in euro.
>
> Il testo contiene spiegazioni su come:
> A) scrivere un * in euro
> B) comportarsi con i * di * in *
> C) trasformare i vecchi * in *

In questo testo il grado di copertura del glossario è pari al 79%. Il compito presuppone la conoscenza di vocaboli come *autorità, assegno, carnet, lira, vietato*, nessuno dei quali è presente nel glossario. In teoria si potrebbe ipotizzare che la parola *lira* sia stranota a tutti gli adulti da una certa età in su, ma non bisogna dimenticare che esiste una generazione giovane che non è mai venuta in contatto con la vecchia moneta, perché sono ormai cinque anni che il mezzo di pagamento dell'Italia è l'euro. Per questo gruppo di discenti la *lira* rimarrà sconosciuta, se non proposta a livello dell'input didattico. L'altra parola chiave è *assegno*, che pur essendo marcato AU nel VdB, non figura nel glossario. Anche in questo caso pare che le probabilità del discente di selezionare l'affermazione giusta, non dipende tanto dal grado con cui ha assorbito il lessico contenuto nel manuale, invece

da altri fattori, come le conoscenze sul mondo, la logica, la capacità deduttiva, ecc. Riproponiamo la versione integrale del testo.

> Per decisione dell'autorità europea, i nuovi carnet di assegni in euro sono diversi da quelli in lire per colore e disegno. I vecchi carnet in lire devono essere consegnati in banca o distrutti. E' vietato anche correggere i vecchi assegni in lire e trasformarli in euro.
>
> Il testo contiene spiegazioni su come:
> A) scrivere un * in euro
> B) comportarsi con i * di * in lire
> C) trasformare i vecchi * in lire

È interessante osservare come diminuisce il grado di copertura del glossario, man mano che il tema dei testi diventa più caratteristico. Da una copertura massima del 93% del primo esercizio si scende a 79% nell'ultimo compito. Nonostante questa flessione, il grado di copertura rimane relativamente elevato, e in teoria dovrebbe garantire la comprensione di tutti i testi. Sappiamo che i testi sono costituiti da un elevato numero di parole di alta frequenza e un numero ridotto di voci meno frequenti. Il problema è che gran parte del messaggio dei testi è affidata proprio a queste parole di bassa frequenza.
In generale possiamo affermare che più l'argomento è specifico, maggior parte del messaggio è portato dalle parole meno frequenti. La comprensione di queste voci crea problemi agli studenti, poiché esse non figurano nell'input didattico di livello elementare (A2). Pare quindi che lo studente che abbia imparato il lessico del manuale *Progetto italiano 1* ha buone probabilità di capire tutti i testi dall'argomento generale, mentre si trova in difficoltà e può contare solo sulle proprie conoscenze non strettamente linguistiche qualora incontri un testo dall'argomento specifico.

6.6 Apporti di un dizionario di base ungherese

Occorre chiarire prima di tutto, quali apporti un dizionario di base ungherese può dare alla selezione del lessico in un manuale didattico della lingua italiana. Rinviando a quanto esposto agli inizi del cap. 3, si osserva che la risposta è da cercare nelle differenze socio-culturali che possono sussistere perfino tra due Paesi non troppo distanti né geograficamente, né culturalmente. È stato anche evidenziato che l'apprendimento di una lingua straniera ha due risvolti importanti: oltre che l'apprendente viene a conoscenza con la cultura del Paese straniero, egli deve essere capace di descrivere nella lingua straniera gli aspetti specifici della cultura del proprio Paese. In termini più generici, egli deve essere in grado di esprimere i concetti e denominare gli oggetti che nel proprio Paese d'origine fanno parte della cultura e/o della quotidianità. Queste parole, che peraltro non necessariamente coincidono con le voci più usuali della L2, non dovrebbero rimanere fuori del suo vocabolario di L2, anche perché tra i temi degli esami orali di italiano L2, organizzati presso enti di certificazione ungheresi, figurano vari temi legati alla cultura e alla quotidianità ungherese (piatti tipici, monumenti e curiosità, organizzazione politica, sistema scolastico, sistema pensionistico ecc. dell'Ungheria).
È legittimo chiedersi come si può garantire che questo lessico entri a far parte delle conoscenze degli allievi? La soluzione migliore, sotto questo aspetto, è di adottare un libro didattico concepito specificatamente per utenti di madrelingua ungherese. Tra questi si ricordano i manuali di Móricz e Szabó (1993, 1994) e di Angelini e Móricz

(1983) che oltre a contenere il lessico italiano di base, selezionato in conformità ai dati del LIF, comprendono una serie di vocaboli importanti per l'utente ungherese.[10] Si pensi in particolare ai termini legati alla civiltà ungherese: nomi di monumenti (*Ponte delle Catene, Bastione dei Pescatori*), nomi geografici (*bacino dei Carpazi, Danubio*), nomi di piatti tipici (*coniglio alla paprika*), figure della storia ungherese (*buttero, cervo fatato, tribù*) ed altri.
Utilizzando invece un manuale destinato ad un pubblico plurilingue, ossia ad apprendenti di madrelingue varie, spetta al docente di adattare il libro alle esigenze degli allievi. L'adattamento in questo caso vuol dire arricchimento, tramite l'introduzione di termini meno comuni in Italia, ma ben noti a tutti i parlanti di madrelingua ungherese. A questo punto possiamo formulare la nostra ipotesi di lavoro: ci aspettiamo di trovare – confrontando sistematicamente i termini del glossario con una lista di frequenza ungherese – un certo numero di voci familiari per l'apprendente ungherese, ma assenti nel lessico del manuale *Progetto italiano 1*.
In assenza di una recente lista di frequenza rappresentativa, abbiamo ritenuto opportuno di effettuare il confronto con l'unico dizionario fondamentale della lingua ungherese. Per la descrizione di questo elenco rinviamo alla sez. 3.2.6, qui ricordiamo soltanto che i lemmi sono stati selezionati in base all'esperienza degli autori, in conformità ai bisogni comunicativi di turisti, studenti, lavoratori e commercianti che vogliono passare periodi più o meno lunghi in Ungheria. La lista contiene quindi il lessico che chiunque voglia cavarsela con la lingua ungherese, deve conoscere. Sfortunatamente gli autori non precisano il carico degli esponenti, ma secondo le nostre stime esse si aggirano attorno a 2.200 unità, il che permette un confronto diretto con il glossario del manuale, comprensivo di 2.819 parole ed espressioni.
I vocaboli, selezionati sulla base del concetto nozionale-funzionale, sono organizzati in due gruppi: nozioni generali, ovvero concetti che appaiono in tutte le situazioni; e nozioni specifiche, ossia termini legati a determinati temi e situazioni. Si prevede che sarà soprattutto questo secondo gruppo a contenere termini assenti nel manuale, anche perché le nozioni generali corrispondono, per gran parte, all'inventario contenuto nel *Threshold Level.*
Per il confronto abbiamo utilizzato la lista alfabetica in appendice alla prima parte, contenente tutte le unità lessicali, sia le nozioni specifiche, sia quelle generali (pp. 171-196).
Vediamo ora i risultati concreti. Nella lista dei sostantivi abbiamo trovato 278 termini assenti nel glossario. Per agevolare l'interpretazione dei risultati, riportiamo i sostantivi raggruppati secondo l'argomento cui sono legati. Cominciamo con gli alimenti:

> *csusza, felvágott, főzelék, gombaleves, gulyás, gulyásleves, halászlé, húsleves, krumplileves, liszt, margarin, mustár, pálinka, paprika, pörkölt, rántotta, tejföl, virsli, zöldségleves, zsír.*

È in questo gruppo che si trova il maggior numero di lemmi che difficilmente possono essere tradotti in italiano, giacché i loro referenti sono inesistenti o per lo meno poco conosciuti in Italia. Si pensi in particolare a *csusza, főzelék, pálinka* che non avendo un loro corrispondente italiano, possono essere parafrasati come *una specie di tagliatelle, contorno di verdure, una specie di grappa*. I termini *gulasch* e *paprika* sono ormai entrati nel lessico italiano, e anche gli altri cibi sono ben conosciuti in Italia, anche se non appartengono al gruppo degli alimenti principali.

[10] È opportuno porre in rilievo il carattere pionieristico di questo manuale, poiché si tratta del primo, e finora l'unico libro didattico destinato ad apprendenti ungheresi, che faccia riferimento ad una lista di frequenza della lingua italiana.

Accanto a bevande e piatti tipici si trovano frutti e verdure e i nomi di carni e pesci:

bab, barack, borsó, cseresznye, eper, káposzta, karfiol, málna, őszibarack, sárgabarack, sárgarépa, szilva, szőlő, uborka, zöldbab, zöldborsó, birkahús, karaj, keszeg, liba, marhahús, sertés, sertéshús, ponty.

Tutti questi alimenti sono diffusi anche in Italia, la differenza sta nell'importanza che essi hanno nell'alimentazione quotidiana. I pesci d'acqua dolce come l'*abramide* o la *carpa* non sono sicuramente tra quelli più consumati. Sono numerose le professioni ed attività:

ápolónő, balett-táncos, bíró, edző, fotóművész, gazdálkodó, gépírónő, hentes, hordár, idegenvezető, képviselő, jegyszedő, közgazdász, környezetvédő, miniszterelnök, miniszterhelyettes, munkanélküli, munkás, nagykövet, óvónő, pap, paraszt, pék, szobrász, teremőr.

In quale misura queste professioni ricalcano la società ungherese attuale, è discutibile, tuttavia l'elenco potrebbe dare dei suggerimenti sull'eventuale immissione di alcune professioni. Ci sono alcuni nomi di oggetti della vita quotidiana:

bajusz, bankjegy, bizonyítvány, bogár, bomba, borda, cd-lejátszó, ceruza, cigaretta, érme, fehérnemű, fogkefe, folyóirat, függöny, fürdőkád, hajkefe, hajszárító, hamutartó, hegedű, kanapé, kancsó, kefe, lemezjátszó, lepedő, műanyag, öngyújtó, pipa, rovar, textil, törülköző, űrlap, videó, zseb.

Si vedano i numi di luoghi ed istituzioni:

ABC, állatkert, áruház, belgyógyászat, benzinkút, bíróság, borozó, börtön, büfé, cirkusz, csomagmegőrző, diákszálló, ebédlő, egyesület, előszoba, étkezőkocsi, folyosó, földszint, főiskola, halastó, hálókocsi, kamra, kemping, középiskola, kikötő, műhely, nagykövetség, páholy, peron, óvoda, padlás, sebészet, szakszervezet, szántóföld, szemészet, terminál, váróterem, videotéka, vészkijárat, völgy.

Richiamiamo l'attenzione sugli istituti di formazione come *asilo* (*óvoda*), *scuola media superiore* (*középiskola*), *istituto superiore* (*főiskola*); ma anche su *enoteca* (*borozó*), che pur essendo conosciuto anche in Italia, nel LIP è registrato con rango 6505. Ecco i sostantivi legati alla circolazione:

autóút, beszállókártya, fék, felüljáró, gyalogos, helyjegy, jelzőtábla, kamion, kerékpáros, körforgalom, sebességkorlátozás, trolibusz, üzemanyag, zebra.

Il mezzo *filobus* (*trolibusz*), ad esempio, circola in poche città italiane, infatti, nel LIP non è registrato. Relativamente nutrito è il gruppo dei termini relativi alla criminalità:

áldozat, betörés, bűnöző, büntetés, emberrablás, emberrabló, gépeltérítés, gépeltérítő, géprabló, gyilkos, ítélet, kábítószer-kereskedő, tanú, terrorizmus, terrorista.

Non mancano i vari tipi di sport:

atlétika, birkózás, boksz, kézilabda, öttusa, röplabda, sakk, tenisz, vitorlázás, vívás, vízilabda.

Sono ovviamente le attività sportive più praticate, e quelle in cui gli ungheresi sono di tradizione talentati. Tra le unità di misura vanno menzionati:

deciliter, deciméter, dekagramm, köbcentiméter, köbméter, mázsa, milliméter, négyzetcentiméter, négyzetméter, négyzetkilométer.

Richiamiamo l'attenzione su *quintale* (*mázsa*), avente un rango 6505 nel LIP. Abbiamo alcuni nomi di malattie:

> *AIDS, gyomorrontás, hányinger, influenza, nátha, rák, tüdőgyulladás.*

Infine si riportano una serie di sostantivi dal significato astratto:

> *államvizsga, árfolyam, árnyék, balesetbiztosítás, barkácsolás, bélyeggyűjtés, betegbiztosítás, betét, biztosítás, borravaló, demokrácia, dzsessz, életrajz, ellátás, évtized, ezüst, fagy, fagypont, félpanzió, felvételi, fogadás, fogalom, forint, forróság, gazdaság, hajléktalan, hardver, házigazda, hobbi, horgászat, hulladék, illat, kabaré, kábítószer, kánikula, kémia, képzőművészet, keresztnév, kertészkedés, kézimunka, környezetvédelem, munkanélküliség, musical, nemzetiség, névnap, népzene, nyugdíj, oktatás, özvegy, postafiók, próza, szegénység, szennyezés, szobrászat, szoftver, szolgáltató, támadás, támadó, támogatás, táppénz, társadalombiztosítás, testnevelés, tisztelet, továbbképzés, ügyelet, választás, vallás, vereség, záróvizsga.*

Segniamo in quest'ultimo gruppo la parola *fiorino*, che ovviamente l'apprendente ungherese deve conoscere. È interessante e caratteristico delle condizioni climatiche dell'Ungheria la presenza simultanea dei termini *gelo* e *canicola*. Parte integrante della civiltà ungherese è la *musica folcloristica*, ma anche l'usanza di festeggiare l'*onomastico*. Notiamo ancora che tra le parole assenti nel glossario primeggiano i vocaboli legati alla vita pubblica ed ufficiale, alla vita civile e politica. È strano che tra le 2.800 parole ed espressioni proposte nel manuale non si trovino termini come *disoccupazione, senzatetto, droga, curriculum, hobby, nazionalità, democrazia, inquinamento* ecc. Per non parlare delle istituzioni come *ambasciata, magazzino, porto, campeggio, asilo*.
Il gruppo degli aggettivi è davvero ridotto rispetto a quello dei sostantivi:

> *anyagi, bolgár, büdös, demokratikus, dohányzó, drogos, durva, egyirányú, holland, hűvös, illatos, jómódú, kábítószeres, képtelen, különélő, liberális, lila, műszaki, narancssárga, nőtlen, ólommentes, paprikás, piszkos, rakott, rántott, rendszeres, román, sérült, sovány, szennyezett, szennyező, szocialista, szögletes, tilos, tisztelt, vastag.*

Richiamiamo l'attenzione sugli aggettivi relativi all'alimentazione (*alla paprica, sporco, stufato, fritto*); nomi di nazionalità (*rumeno, bulgaro, olandese*); colori (*arancio, viola*); aggettivi relativi alla politica (*socialista, democratico*); e poi una serie di aggettivi dal significato generico (*materiale, puzzolente, fumatore, drogato, rozzo, a senso unico, fresco, profumato* ecc.).
I verbi sono 51 in tutto:

> *aláír, atletizál, átutal, becsomagol, betöm, betör, betűz, birkózik, bokszol, dörög, edz, elromlik, elront, elszakad, elvámol, fagy, felrobbant, gyógyít, hegedül, kézilabdázik, leporol, letartóztat, lop, megbetegszik, megerőszakol, megbukik, megbüntet, megéget, megfázik, megjavít, megoperál, megöl, megszagol, megtámad, megtilt, megtölt, megüt, meggyógyít, meggyógyul, mosogat, öttusázik, röplabdázik, sakkozik, süllyed, takarít, teniszezik, tönkremegy, tönkretesz, villámlik, vív, vízilabdázik,*

Troviamo verbi legati ad attività sportive: *praticare l'atletica, il pugilato, il box, il pentathlon, la scherma, allenarsi, giocare a pallamano, a pallavolo, a scacchi, a tennis, pallanuoto*; e una serie di verbi generici come *firmare, versare, incartare, otturare, irrompere, fare lo spelling, tuonare, guastarsi, guastare, strapparsi* ecc.
Seguono infine due avverbi: *hiába, végeredményben*, di cui soprattutto il primo può essere utile per l'apprendente, infatti, nel LIP è registrato col rango 3514.
Dalle liste emerge che relativamente poche sono le parole di riferimento prettamente ungherese, ed esse si limitano quasi esclusivamente alla categoria dei sostantivi (piatti tipici, qualche festa, la moneta nazionale, i nomi di istituti di formazione). Del resto, e forse questo è l'aspetto più importante, si nota uno slittamento nel valore d'uso di certe parole ed espressioni. Un libro didattico per l'insegnamento dell'italiano ad ungheresi,

costruito a partire delle parole fondamentali della lingua ungherese, conterrebbe, infatti, un lessico leggermente diverso da quello dei manuali destinati ad un pubblico plurilingue. Per fare un esempio, le parole *zoo* e *circo*, registrati nel LIP con ranghi 6505 e 3361, rispettivamente, avrebbero poca probabilità di entrare in un libro didattico di primo livello. Muovendosi invece dalle parole che gli parlanti concepiscono come fondamentale nella propria madrelingua, queste voci dovrebbero figurare tra le 3.000 parole più importanti.
I concetti sopraelencati sono presenti nel vocabolario attivo degli studenti, che quindi in caso ideale dovrebbero essere in grado di esprimerle anche nella lingua straniera. Se questi elementi lessicali non compaiono a livello del libro di testo, come non figurano nel *Progetto italiano 1*, spetta all'insegnante di inserirli nel programma didattico, anche perché durante gli esercizi di produzione scritta ed orale gli alunni, parlando delle proprie esperienze, chiedono in ogni caso all'insegnante, come si chiama un determinato concetto in italiano. Pare quindi, che i manuali scritti per utenti plurilingui vadano integrati in conformità alle caratteristiche lessicali e culturali del Paese di provenienza degli apprendenti.

6.7 Suggerimenti per gli autori dei libri di testo

Potrà sembrare che il manuale sia criticato per alcune scelte lessicali. In realtà non si può dire che la scelta di certe parole fosse sbagliata in quel preciso contesto, ma si vuole riflettere se un determinato tipo di scelte, basate sulla competenza degli autori, sia funzionale in una grammatica di primo apprendimento di una L2, o se piuttosto non si debbano rifare a liste di frequenza e/o vocabolari di base, per evitare gli sforzi improduttivi da parte dell'apprendente.
Il VdB può costituire uno strumento utile per riequilibrare il lessico dei libri di testo, non soltanto a livello elementare, ma soprattutto in quelli superiori. Dal confronto con il VdB potrebbero nascere alcuni interventi sul lessico: bisognerebbe eliminare certi lemmi, immettere altri, infine riconsiderare l'importanza di alcuni. Il fatto che il 15% del glossario è costituito da parole non comprese nel VdB che include le circa 7.000 parole fondamentali della lingua italiana, pone in rilievo alcuni problemi. Da un lato nel manuale si trova un numero consistente di vocaboli ed espressioni, quindi sarebbe auspicabile l'espunzione di alcuni termini; d'altro lato mancano parole importanti che già in un primo livello di conoscenza dovrebbero entrare a far parte del patrimonio lessicale dei discenti.
Prima di tutto bisognerebbe esaminare le voci non appartenenti al VdB: una gran parte di esse, in quanto tecnicismi, esotismi o semplicemente di troppo bassa frequenza d'uso, dovrebbe essere eliminata. Si tratta dei termini contrassegnati da una crocetta nella lista alfabetica in appendice. E' ovvio che le parole relative alla grammatica e all'insegnamento in classe non possono non far parte di un manuale didattico, anche se, magari, sono parole poco usate nella vita quotidiana. Eppure esiste una serie di termini assenti nel VdB, la cui utilità nella prima fase dell'apprendimento andrebbe riconsiderata.
D'altra parte il VdB può essere fonte anche di immissioni di lemmi. Queste aggiunte però non vanno effettuate in modo meccanico. Non è detto che l'apprendimento delle tre fasce del VdB deve avvenire in modo sequenziale, ma operando una selezione appropriata tra i lemmi appartenenti alle tre fasce. Ovviamente, le poche parole

fondamentali mancanti vanno inserite nel manuale, poiché è questa la fascia alla quale occorre rivolgersi primariamente nella formazione di livello elementare. Riaffermiamo la proposta avanzata nella sez. 6.4 e confermata anche da Vedovelli (1995: 76), secondo cui la seconda fascia da prendere come riferimento sarebbe il VAD, proprio per la sua rilevanza come componente del sapere linguistico dei parlanti e per il fatto di esprimere nozioni dell'esperienza quotidiana, mentre il lessico AU dovrebbe diventare oggetto di apprendimento solo dopo il consolidamento dei due livelli precedenti.
Una proposta concreta è di integrare il glossario con quelle parole del VAD, che, in base alla verifica della ricchezza ricettiva del VdB, risultano conosciute dalla stragrande maggioranza degli alunni della terza media. Vedovelli e i suoi collaboratori in una ricerca hanno sottoposto l'intero VdB ad una verifica generalizzata, la quale ha permesso di identificare un nucleo centrale del lessico italiano.[11] Esso nasce dalla somma dei 2.000 termini fondamentali, più 1.100 voci che risultano dall'intersezione del VAU e del VAD e sono conosciute dal 100-97,8% degli alunni della terza media. Da un confronto non sistematico del glossario con i vocaboli di alta disponibilità, riportati in Gensini e Vedovelli (1983: 188-196) risulta che numerose sono le parole assenti. Ecco alcune parole AD assenti nel glossario, ma conosciute dalla quasi totalità (97-100%) degli allievi:

> *agnello, albicocca, ambulanza, appendicite, apriscatole, berretto, bidello, campeggio, capra, cascata, casalinga, colla, coperchio, cuccia, disinfettare, distributore, elettricità, garofano, gatta, graffiare, incartare, incollare, lavapiatti, macellaio, matita, materasso, muratore, nuoto, pallacanestro, parare, pedalare, pettinatura, pigiama, pollice, raffreddore, ragno, rivincita, scalzo, scoglio, secchio, sorpasso, spigolo, stampo, trapano, tuono* ecc.

L'elenco si potrebbe continuare a lungo, man mano che si scende con le percentuali di comprensione. Questi termini potrebbero contribuire ad arricchire il lessico degli studenti principianti.
A questo punto è legittimo chiedersi se un lessico revisionato sulla base del VdB fosse il più appropriato per scopi didattici. Il VdB, in realtà, rappresenta una potente base di riferimento per determinare il vocabolario fondamentale, ma non vogliamo affermare che esso basti per determinarlo. Per sapere che lingua insegnare agli studenti non italiani, ci vuole il VdB e qualcos'altro.
In un articolo che mette a confronto il LIP con il VdB De Palo (1997: 395-411) osserva che alcuni risultati del LIP, soprattutto quelli provenienti da conversazioni faccia a faccia e da quelle telefoniche, potrebbero contribuire a riequilibrare il VdB. Le integrazioni al VdB dovrebbero provenire da lemmi con rango fino a 2200, con un uso superiore a 6. In questo modo alcuni lemmi presenti nel glossario ma assenti nel VdB si sarebbero "salvati." Riportiamo l'elenco delle voci attestati nel LIP con uso superiore a 6 nelle conversazioni telefoniche e faccia a faccia. Si tratta in particolare di sostantivi (tra parentesi l'uso):

> *aggettivo* (6), *brano* (15), *ascolto* (12), *cassetta* (16), *descrizione* (11), *diretta* (11, *discoteca* (7), *foto* (22), *indicazione* (28), *moto* (7), *percentuale* (11), *schema* (15), *sintesi* (9); degli aggettivi *cronologico* (6), *specializzato* (9), *televisivo* (6); dei verbi *individuare* (26), *tutelare* (15); degli avverbi *assolutamente* (76), *certamente* (81), *direttamente* (65), *praticamente* (172), *probabilmente* (113), *sicuramente* (115), *veramente* (243), *ormai* (148).

A tutto questo va aggiunto che il VdB andrebbe integrato non solo con i lemmi del LIP che più massicciamente sono attestati nelle interazioni orali e informali, ma anche con le voci che l'evoluzione tecnica e sociale ha reso diffusi negli ultimi venti, quindici anni, come ad es. *euro, ipermercato, e-mail, televendita, ecc*.

[11] Per i dettagli della ricerca si rinvia a Gensini e Vedovelli (1983: 179-187).

Per ottenere quindi un lessico equilibrato pare quindi che non basti fare affidamento solamente al LIP, né al VdB, ma ci vuole l'uso combinato dei due strumenti, integrati, qualora necessario, da altre liste quali il LE, il VF o il LIF.
Nella scrittura dei libri di testo occorre considerare molti fattori, di cui il valore di frequenza degli elementi lessicali è solo uno. Non bisogna perdere di vista alcuni principi teorici generali: immettere nella lista solo parole che servono veramente alla verbalizzazione di situazioni comunicative quotidiane; non includere, come avviene nel nostro glossario, parole che senza un loro corrispettivo ben definito hanno basso valore semantico (*fumare* senza *sigaretta*); evitare di inserire voci che si possono facilmente sostituire o parafrasare *(t-shirt* vs. *maglietta, traversa* vs. *via secondaria, cileno* vs. *del Cile* ecc.); preferire i termini di registro meno formale (*finire* vs. *terminare*); i termini meno marcati stilisticamente (*prendere* vs. *pigliare*). Va rilevata l'importanza di introdurre anche quelle formule fisse o semifisse che, come abbiamo visto, costituiscono una parte non marginale della comunicazione, e la cui acquisizione è importante per un uso idiomatico e naturale della lingua. Esistono poi delle parole magari poco frequenti nella L2, specifiche della cultura o civiltà del Paese di provenienza dell'allievo che fanno parte del suo bagaglio lessicale fondamentale di L1. Un manuale davvero funzionale dovrebbe includere anche questi termini.
In sintesi, la scelta lessicale in un buon manuale didattico andrebbe operata tenendo presente almeno tre criteri: l'obiettivo più importante è che il libro rappresenti le strutture lessicali che sono effettivamente le più usuali nella L2. Gli strumenti più utili, per ovviare a questo problema, sono appunto i lessici fondamentali e di frequenza. In secondo luogo occorre considerare le indicazioni relative ai livelli di competenza lessicale, previsti dagli enti di certificazione. Solo in questo modo è possibile la coerenza tra gli obiettivi didattici e quelli formulati in sede di certificazione. In ultimo luogo, bisogna prendere in considerazione i bisogni lessicali dei destinatari. Nei libri destinati ad un pubblico multilingue, questo criterio viene meno, e in questi casi spetta all'insegnante di provvedere all'introduzione delle strutture lessicali che corrispondono a oggetti e concetti altamente disponibili nella L1 dei propri studenti.

7 Conclusioni

Le considerazioni sovra esposte nascono dall'intento, del tutto pratico, di compiere una prima esplorazione su quello che si potrebbe concretamente fare per la selezione del lessico fondamentale dell'italiano L2, facendo riferimento al patrimonio lessicale di una grammatica italiana per stranieri, messa a confronto con alcuni degli strumenti offerti dalla linguistica teorica e matematica.
Nella parte introduttiva della tesi abbiamo argomentato che nella didattica d'ogni materia che presenti una certa vastità e complessità, si cerca ragionevolmente, da parte dei docenti, il sistema più semplice che, a parità di sforzi a parte dell'apprendente, dia i risultati migliori rispetto a tutti gli altri. L'apprendimento di una lingua è un processo, appunto, che implica difficoltà tali che ben si comprende il tentativo di riduzione. Mentre gran parte degli studiosi di linguistica affrontò il problema generale della razionalizzazione delle conoscenze circa la struttura dalla lingua, dopo le prime iniziative degli anni Venti negli Stati Uniti, dagli anni Cinquanta in poi anche in Europa si fece largo il tentativo di semplificare la parte più fastidiosa per il discente, il lessico. Il primo capitolo è dedicato alla descrizione di due tipi di opere lessicografiche: i dizionari di frequenza e i lessici fondamentali, di cui i primi costituirono l'antecedente pratico e teorico. Nella seconda e terza parte si presentano diverse liste di frequenza e di base della lingua italiana e dell'ungherese, al fine di collocare dal punto di vista glottodidattico il ruolo di questi strumenti fra i riferimenti utilizzabili dall'insegnante nella programmazione della propria attività, e nella definizione degli obiettivi di competenza degli allievi. Nel capitolo 4 abbiamo additato che i dati di frequenza provenienti dall'analisi di larghi corpora, possono avere una serie di applicazioni di notevole utilità, in particolare possono contribuire al miglioramento della qualità dei prodotti lessicografici di uso comune. Di seguito abbiamo presentato le iniziative prese dal Consiglio d'Europa per l'omologazione delle competenze linguistiche relative ai singoli stadi d'apprendimento, e abbiamo esposto come si configurano concretamente i requisiti lessicali, stabiliti da enti di certificazione italiani. È emerso che per dare una definizione concreta delle competenze previste è indispensabile ricorrere a strumenti di riferimento quali dizionari fondamentali e di frequenza, altrimenti ci si trova di fronte ad una serie di criteri molto vaghi, in sostanza inutilizzabili per la valutazione e per la programmazione. Nel sesto capitolo, infine, si è cercato di esaminare le caratteristiche quantitative e qualitative del lessico proposto da uno dei manuali più diffusi nell'insegnamento dell'italiano L2 in Ungheria. Abbiamo argomentato che senza l'uso sistematico di strumenti di riferimento lessicali, nonostante la perspicacia degli autori non è possibile operare scelte che rispecchiano l'uso che i parlanti nativi fanno della propria lingua. Un dato rilevante è che ca. un quarto degli elementi lessicali proposti è assente o registra meno di tre occorrenze nel corpus LIP. È stato anche evidenziato che il lessico contenuto nel glossario non garantisce il superamento della prova di lettura di un esame di certificazione calibrato a livello A2. Per riequilibrare il lessico, abbiamo fatto delle proposte concrete riguardo l'immissione e l'espunzione di alcuni lemmi.
In definitiva, è fondamentale che i processi di apprendimento si nutrano di un input che potenzialmente è testimone della struttura generale della lingua obiettivo. Poiché le strutture lessicali di una lingua si depositano nelle liste di frequenza/uso e nelle loro interpretazioni costituite dal vocabolario di base, sono proprio tali lavori che vanno prese come elementi di riferimento durante il processo di sviluppo lessicale e, di conseguenza,

in sede di valutazione delle competenze lessicali. Nonostante non vi sia accordo tra i ricercatori né sulla classificazione delle diverse voci, né sui modi migliori per selezionarle, né sulla loro utilità per diversi tipi di studenti, appare molto importante per gli insegnanti avere familiarità con questo settore per ricercarvi non certezze, ma accresciuta consapevolezza dei problemi, più realistiche aspettative nei confronti dell'apprendimento, magari una conferma delle intuizioni e suggerimenti elaborati nella pratica didattica. Si tratta di un settore dinamico e aperto che offre molti stimoli all'insegnante che sappia adattarlo al suo contesto d'insegnamento.
Quale, quindi, tra le liste di frequenza e di base, andrebbe preso come strumento di verifica per definire gli obiettivi, controllare i testi, preparare prove di verifica ecc.? A nostro parere il primo strumento rimane il VdB, nonostante siano passati ormai più di un ventennio dalla sua prima pubblicazione. È stato dimostrato che, pur trattandosi di una lista basata su un repertorio dell'italiano scritto, quale il LIF, i suoi risultati rimangono validi anche per la lingua parlata, come dimostrato da vari contributi che mettono a confronto i lemmi del VdB con il LIP (cfr. ad es. Vedovelli 1995). In più, nel VdB la fascia del lessico di alta disponibilità è scelta in modo appropriato per presentare ciò che è vivo nella coscienza linguistica di un parlante italiano, ma spesso assente nei testi scritti o parlati. Il VdB dunque, si presenta come il più rappresentativo strumento per accedere al nucleo del lessico italiano d'uso comune contemporaneo. Proprio la sua articolazione interna in fasce può consentire la sua utilizzazione didattica.
Non crediamo che una lista lessicale debba essere utilizzata meccanicamente come elenco di parole da apprendere in modo meccanico, ovvero fuori dei testi e delle situazioni comunicative in cui esse vivono. Fatta questa premessa, però, ribadiamo la centralità di strumenti come le liste di frequenza e lessici fondamentali in relazione a problemi della didattica dell'italiano. Il docente ha il dovere di conoscere tali liste per conoscere la dimensione lessicale della lingua che insegna. Nell'insegnamento scolastico la selezione del materiale (lessicale) è operata a livello di libro di testo, e il successo del lavoro dell'insegnante dipende, fra l'altro, dalla chiarezza dei criteri che hanno guidato le scelte degli autori dei libri di testo, è quindi importante che gli insegnanti siano in grado di valutare criticamente tali scelte. D'altra parte, rispetto al passato, i testi proposti in classe che fungono da input all'apprendimento sono molto vari, cui si aggiunge una pluralità di fonti, tra cui Internet e i mass media. Occorre, quindi, servirsi di tali liste come termini di riferimento, e di strumenti come i programmi che calcolano automaticamente il grado di leggibilità di un testo, per verificare se i testi scelti per il processo didattico siano effettivamente rappresentativi dell'universo lessicale.
Delle liste di frequenza e dei vocabolari fondamentali bisogna ancor più tener conto, se si vuole stabilire il lessico attivo da raggiungere alla fine del corso. Tali liste, con i loro diversi livelli, segnalano diverse gerarchie di rilevanza e di funzionalità del lessico. Questi livelli andrebbro presi come punti di riferimento per stabilire gli obiettivi e per verificare gli esiti del processo formativo, e, di conseguenza, per fissare i livelli di lessico da coinvolgere in sede di certificazione linguistica.

Bibliografia

ALBANO LEONI F. (2003) Tre progetti per l'italiano parlato. In: MARASCHIO N., POGGI SALANI T. (a cura di) Atti del XXXIV Congresso SLI, *Italia linguistica anno mille, Italia linguistica anno duemila,* Roma, Bulzoni, pp. 675-683.

AMIZZONI M. (1991) *Calcolo automatico della leggibilità: l'indice GULPEASE*. Tesi di laurea, cattedra di Filosofia del linguaggio, Istituto di Filosofia, Università degli Studi di Roma "La Sapienza", Roma.

ANGELINI M., MÓRITZ GY. (1983) *Olasz nyelv II.,* Budapest, Tankönyvkiadó.

ARCAINI E. (1978) *Educazione linguistica come strumento e come fine*, Milano, Feltrinelli, pp. 240-257.

BARDEL C., NYSTEDT J. (a cura di) (2006) *Progetto dizionario svedese. Atti del primo colloquio*, Intellecta Docusys, Stoccolma.

BARKI P., GORELLI S., MACHETTI S., SERGIACOMO M. P., STRAMBI B. (2003) *Valutare e certificare l'italiano di stranieri: i livelli iniziali,* Perugia, Guerra.

BECCARIA G. L. (a cura di) (2004) *Dizionario di linguistica*, Torino, Einaudi.

BENUCCI A. (1994) L'italiano delle grammatiche italiane per stranieri. In: Giacalone Ramat A., Vedovelli M. (a cura di) Atti del XXVI Congresso SLI, *Italiano lingua seconda / lingua straniera*, Roma, Bulzoni, pp. 165-189.

BERGENHOLTZ H. (1990) Probleme der Selektion im allgemeinen einsprachigen Wörterbuch. In: HAUSMANN F. *et alii* (a cura di), pp. 772-779.

BERRETTA M. (1983) Recensione a: Nora Galli de' Paratesi, *Livello soglia per l'insegnamento dell'italiano,* Strasbourg, Consiglio d'Europa, 1981. In: "Vox Romanica", 42, pp. 243-247.

BERRETTA M. (1996) *La componente pragmatica nei modelli linguistici e le sue implicazioni per l'insegnamento delle lingue*, Torino, Giappichelli, pp. 49-69.

BETTONI C. (2001) *Imparare un'altra lingua*, Roma, Laterza, pp. 54-115.

BINI M. (1997) Le parole di "alta disponibilità" nell'italiano lingua straniera. In: DE MAURO T., LO CASCIO V., *Lessico e grammatica - Teorie linguistiche e applicazioni lessicografiche*, Bulzoni, Roma, pp. 17-24.

BOLASCO S. (2005) *La reperibilità statistica di tendenze diacroniche nell'uso delle parole*. In: DE MAURO T., CHIARI I. (a cura di) pp. 335-354.

BOLASCO S., BISCEGLIA B., BALOCCHI F. (2004) Estrazione automatica d'informazione dai testi. In: *Mondo digitale*, IV, 2, pp. 27-43. [disponibile on-line sul sito http://www.mondodigitale.net/Rivista/04_numero_due/Bolasco_p.27-43.pdf]

CALZOLARI N. (1992) Linguistica computazionale. In: MIONI A., CORTELAZZO M. (a cura di) *La linguistica italiana degli anni 1976-1986*, Roma, Bulzoni, pp. 409-433.

CARDINALE U. (1994) L'utilità del dizionario fondamentale nel panorama lessicologico contemporaneo. In: *Atti del seminario internazionale di studi sul lessico*, Bologna, Editrice Universitaria, pp. 183-193.

CASSESE S. (a cura di) (1993) *Codice di stile delle comunicazioni scritte ad uso delle amministrazioni pubbliche. Proposta e materiali di studio*, Roma, Istituto Poligrafico e Zecca dello Stato.

CHIARI I. (2004) *Informatica a lingue naturali. Teorie e applicazioni computazionali per la ricerca sulle lingue*, Roma, Aracne.

CHIARI I. (2007) *Introduzione alla linguistica computazionale*, Roma, Laterza.

CHIÙCHIÙ A., MINCIARELLI F., SILVESTRINI M. (1990) *In italiano*, Perugia, Guerra.

CILIBERTI A. (1994) *Manuale di glottodidattica, Firenze*, La Nuova Italia, 1994, pp. 55-150.
CORDA A., MARELLO C. (1999) *Insegnare e imparare il lessico,* Torino, Paravia.
D'ADDIO COLOSIMO W. (1976) *Lingua straniera e comunicazione. Problemi di glottodidattica*, Bologna, Zanichelli.
DE MAURO T. (1961) Statistica linguistica. In: *Enciclopedia Italiana,* Appendice III. 2. vol., Roma, pp. 820-821.
DE MAURO T. (1979) *Scuola e linguaggio,* Roma, Editori Riuniti, pp. 27-35.
DE MAURO T. (1980) *Guida all'uso delle parole,* Roma, Editori Riuniti. [1989[10], 1991[11], 1997[12]]
DE MAURO T. (1990) *Minisemantica*, Roma, Laterza, pp. 105-141.
DE MAURO T. (1994) *Capire le parole, Roma*-Bari, Laterza.
DE MAURO T. (2005) *La fabbrica delle parole. Lessico e problemi di lessicologia,* Torino, UTET.
DE MAURO T. (2006) Dizionari tra teorie e pratica. In: BARDEL C., NYSTEDT J. (a cura di), pp. 13-22.
DE MAURO T. (a cura di) (1994) *Come parlano gli italiani, Firenze*, La Nuova Italia.
DE MAURO T., CHIARI I. (a cura di) (2005) *Parole e numeri. Analisi quantitative dei fatti di lingua, Roma*, Aracne.
DE MAURO T., DE PALO M., DE RENZO F., FARES P., MANCINI F., VEDOVELLI M., VOGHERA M. (1992) Il lessico di frequenza dell'italiano parlato: LIP. In: Atti del XXV Congresso SLI, *Linee di tendenza dell'italiano contemporaneo*, Roma, Bulzoni, pp. 83-117.
DE MAURO T., FERRERI S. (2005) Quantità di lemmi nei dizionari. In: DE MAURO T., CHIARI I. (a cura di) pp. 297-306.
DE PALO M. (1997) Il Vocabolario di base a confronto con il Lessico italiano del parlato. In: DE MAURO T., LO CASCIO V., *Lessico e grammatica. Teorie linguistiche e applicazioni lessicografiche*, Roma, Bulzoni, pp. 395-411.
DE RENZO F. (2005) Nuove rilevazioni sul vocabolario di base. In: DE MAURO T., CHIARI I. (a cura di) pp. 215-232.
DELLA CASA M. (1993) Insegnare la grammatica. In: CORNO D. (a cura di) *Vademecum dell'educazione linguistica*, Firenze, La Nuova Italia, pp. 5-12.
DELLA VALLE V. (2005) *Dizionari italiani: storia, tipi, struttura*, Roma, Carocci.
DIADORI P. (2001) *Insegnare italiano a stranieri*, Firenze, Le Monnier.
F. FRANCESCHINI (a cura di) (2003) *Manuale di scrittura amministrativa*, Roma, Agenzia delle entrate. [Disponibile anche sul sito dell'Agenzia delle entrate: http://www1.agenziaentrate.it/documentazione/guide/scrittura_amministrativa/index.htm]
FERRERI S. (2006) Servono le marche d'uso in un dizionario bilingue? Problemi, compiti, applicazioni. In: BARDEL C., NYSTEDT J. (a cura di) pp. 119-144.
FÖLDES CS. (2004) A szótárak nagyságrendi osztályozásáról és elnevezéséről. In: FÓRIS Á., PÁLFY M. (a cura di) *A lexikográfia Magyarországon*, Budapest, Tinta Kiadó, pp. 29-38.
FÓRIS Á. (2000) Olasz gyakorisági szótárak. In: TASSONI L., FÓRIS Á. (szerk.), *Olasz nyelvi tanulmányok*, Pécs, Iskolakultúra, pp. 125-143.
FÓRIS Á. (2002) *Szótár és oktatás*, Pécs, Iskolakultúra.
GENSINI S., VEDOVELLI M. (1983) *Teoria e pratica del glotto-kit*, Milano, Franco Angeli.
GIACALONE RAMAT A. (1993) L'italiano di stranieri. In: SOBRERO A. (a cura di) *Introduzione all'italiano contemporaneo*, Roma-Bari, Laterza, pp. 341-410.

GIULIANI A., IACOBINI C., THORNTON A. M. (2005) La stratificazione diacronica del Vocabolario di base. In: DE MAURO T., CHIARI I. (a cura di) pp. 193-213.

GOUGENHEIM G., MICHÉA R., RIVENC P., SAUVAGEOT A. (1964) *L'élaboration du français fondamental*, Paris, Didier.

GREGO BOLLI G., SPITI M. G. (2000) *La verifica delle competenze linguistiche. Misurare e valutare nella certificazione CELI*, Perugia, Guerra.

GUIRAUD P. (1954) *Les caractères statistiques du vocabulaire*, Paris.

HALÁCSY P., KORNAI A., NÉMETH L., RUNG A., SZAKADÁT I., TRÓN V. (2003) Szógyakoriság és helyesírás-ellenőrzés. In: *Proceedings of the 1st Hungarian Computational Linguistics Conference*, Szeged, Szegedi Tudományegyetem, pp. 211-217.

HALÁCSY P., KORNAI A., NÉMETH L., RUNG A., SZAKADÁT I., TRÓN V. (2004) Creating open language resources for Hungarian, In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC2004)*, Lisbon. [disponibile sul sito www.mokk.bme.hu]

HAUSMANN F. J. (1990) Das Kinderwörterbuch. In: HAUSMANN F. *et alii* (a cura di) pp. 1365-1368.

HAUSMANN F. J., REICHMANN O., WIEGAND H. E., ZGUSTA L. (a cura di) (1989-1991) *Wörterbücher, Dicionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie. An international Encyclopedia of Lexicography. Encyclopédie internationale de lecixographie, voll. I-III.* Berlin - New York, Walter de Gruyter.

HEILMANN L. (1971) *Corso di linguistica teorica*, Milano, CELUC.

HERDAN G. (1964) *Quantitative Linguistics*, London, Butterworth & Co. [trad. it. 1971, Bologna, Il Mulino.]

IACOBINI C. (1999), Dizionari della lingua italiana su cd-rom. In: *Lingua e stile*, XXXIV, 4, pp. 541-568.

KATERINOV K., BORIOSI KATERINOV M. C. (1994) Le strutture della morfosintassi del lessico dell'italiano parlato. Implicazioni didattiche per l'italiano L2. In: GIACALONE RAMAT A., VEDOVELLI M. (a cura di) Atti del XXVI Congresso SLI, *Italiano lingua seconda / lingua straniera*, Roma, Bulzoni, pp. 191-197.

KORNAI A., HALÁCSY P., NAGY V., ORAVECZ CS., TRÓN V., VARGA D. (2006) Web-based frequency dictionaries for medium density languages. In: KILGARRIFF A., BARONI M. *Proceedings of the 2nd International Workshop on Web as Corpus,* ACL-06, pp. 1-9.

Közös Európai Referenciakeret: nyelvtanítás, nyelvtanulás, értékelés (2002) Strasbourg, Európa Tanács.

KROHN D. (1992) *Grundwortschätze und Auswahlkriterien. Metalexikographische und fremdsprachendidaktische Studien zur Struktur und Funktion deutscher Grundwortschätze*, Göteborg, Acta Universitatis Gothoburgensis.

KÜHN P. (1979) *Der Grundwortschatz. Bestimmung und Systematisierung*, Tübingen, Niemeyer.

KÜHN P. (1990) Das Grundwortschafzwörterbuch. In: HAUSMANN F. *et alii* (a cura di), pp. 1353-1364.

LENGYELNÉ MOLNÁR T. (2006) Gyakorisági szótárak. Magyarországi helyzetkép. In: *Könyvtári Figyelő,* 2006/1, pp. 45-58.

LEPSCHY G. C. (1978) Le parole più comuni? In: *Saggi di linguistica italiana*, Bologna, Il Mulino, pp. 55-59.

LO CASCIO V. (1994) Ricchezza e povertà dell'italiano parlato in Italia e all'estero. In: DE MAURO T. (a cura di) *Come parlano gli italiani,* Firenze, La Nuova Italia, pp. 51-69.

LO DUCA M. G. (2004) *Lingua italiana ed educazione linguistica,* Roma, Carrocci.
LUCISANO P. (a cura di) (1992) *Misurare le parole*, Roma, Kepos.
MACKEY W. F., SAVARD J. G. (1967) The Indices of Coverage: A New Dimension in Lexicometrics. In: IRAL 5, 71-121.
Magyar értelmező kéziszótár (2003) Budapest, Akadémiai Kiadó.
MARELLO C. (1994) Lessico, lessicologia, dizionari: un panorama italiano. In: *Atti del seminario internazionale di studi sul lessico*, Bologna, Editrice Universitaria, pp. 161-166.
MARELLO C. (1996) *Le parole dell'italiano. Lessico e dizionari, Bologna*, Zanichelli.
MARIN T., MAGNELLI S. (2006) *Progetto italiano 1. Corso multimediale di lingua e civiltà italiana*, Roma, Edilingua.
MARTIN W. (1990) The frequency dictionary. In: HAUSMANN F. *et alii* (a cura di) pp. 1314-1322.
MARTINET A. (1955) *Economie des changements phonétiques*, Berne, Frank.
MARTINET A. (1966) *Elementi di linguistica generale,* Laterza, Bari.
MASTIDORO N. (1991) *Rilevamento automatico del tasso di vocabolario di base*. Tesi di laurea, Facoltà di Lettere e Filosofia, Università degli Studi di Roma "la Sapienza", Roma.
MASTIDORO N. (1992) Il sistema Èulogos per la valutazione automatica della leggibilità. In: LUCISANO P. (a cura di), pp. 125-141.
MASTIDORO N., Amazzoni M. (2005) Strumenti automatici di analisi e gestione testuale. In: DE MAURO T., CHIARI I. (a cura di) pp. 417-438.
MAZZOTTA P. (a cura di) (2002) *Europa, lingue e istruzione primaria. Plurilinguismo per il bambino italiano-europeo,* Torino, UTET.
MERZAGORA G. M. (1983) *La lessicografia,* Bologna, Zanichelli.
MIGLIORINI B. (1951) *Che cos'è un vocabolario?,* Firenze, Le Monnier.
MISANO G., COSSUTTA F. (1977) Contributi ad una indagine del lessico italiano di base. In: EBNETER T. (a cura di) *Italiano d'oggi. Lingua nazionale e varietà regionali*, Trieste, LINT, pp. 283-293.
MOREAU R. (1962) Au sujet de l'utilisation de la notion de fréquence en linguistique. In: *Cahiers de lexicologie*, 3, pp. 140-158.
MÓRITZ GY., SZABÓ GY. (1993) *I primi due passi 1/A*, Budapest, Tankönyvkiadó.
MÓRITZ GY., SZABÓ GY. (1994) *Arrivederci 1/B*, Budapest, Tankönyvkiadó.
NEUBAUER F. (1990) Vocabulary Control in the Definitions and Examples of Monolingual Dictionaries. In: HAUSMANN F. *et alii* (a cura di) pp. 899-905.
ODGEN C. K. (1937) *A Short Guide to Basic English,* Cambridge, Orthological Institute.
Olasz-magyar szószedet. Supplemento per studenti ungheresi (2006) Roma, Edilingua.
PAPP F. (2006) A debreceni thészaurusz. In: KLAUDY K. (a cura di) *Papp Ferenc olvasókönyv*, Budapest, Tinta Kiadó.
PATOTA G., PIZZOLI L. (2004) *PLIDA. Progetto Lingua Italiana Dante Alighieri, Firenze, Società Dante Alighieri.* [disponibile sul sito www.ladante.it]
PIEMONTESE M. E. (1998) *Capire e farsi capire. Teorie e tecniche della scrittura controllata*, Tecnodid, Napoli.
PIEMONTESE M. E. (2005) Misurazioni quantitative degli stili personali e indici di leggibilità, In: DE MAURO T., CHIARI I. (a cura di) pp. 377.397.
PIEMONTESE M. E., SAPONARO CIOFFI A. (1994) L'informazione scientifica in "Due Parole" In: DE MAURO (a cura di) *Studi sul trattamento linguistico dell'informazione scientifica*, Roma, Bulzoni, pp. 273-297.
PRAT ZAGREBELSKY M. T. (1998) *Lessico e apprendimento linguistico*, Quaderni del LEND, 13, Firenze, La Nuova Italia.

Quadro comune europeo di riferimento per le lingue. Apprendimento, insegnamento, valutazione. Strasburgo, Consiglio d'Europa.

Reference Level Descriptions for National and Regional Languages (2005) Strasbourg, Council of Europe.

ROSIELLO L. (1965) Consistenza e distribuzione statistica del lessico poetico di Montale. In: *Rendiconti*, 11, pp. 397-421.

RUDOLF R., HARTMANN K. (1990) *The Dictionary as an Aid to Foreign-Language Teaching*. In: HAUSMANN F. *et alii* (a cura di), pp. 181-189.

RUSSO D. (2005) La rilevazione dei gradi di notorietà dei lemmi del Vocabolario AD. In: DE MAURO T., CHIARI I. (a cura di) pp. 233-246.

RUSSO T. (2005) Un lessico di frequenza della LIS. In: DE MAURO T., CHIARI I.(a cura di) pp. 277-290.

SCHARNHORST J. (1995) Der Wortschatz unter dem Aspekt von Zentrum und Peripherie. In: POHL I./ERHARDT H. (HRSG.): *Wort und Wortschatz. Beiträge zur Lexikologie,* Tübingen, Niemeyer, pp. 63-75.

SCHUMACHER H. (1998) Über den Nutzen von Grundwortschätzen. In: JUNG U., *Praktische Handreichung für Fremdsprachenlehrer*, Berlin, Peter Lang, pp. 161-168.

SERIANNI L. (1999) *Dizionari di ieri e oggi*, Garzanti, Milano.

SERRA BORNETO C. (1998) *C'era una volta il metodo. Tendenze nella didattica della lingue straniere, Roma*, Carracci.

SGROI S. C. (1981) I lessici fondamentali e di frequenza della lingua italiana (1927-1981) In: *Quaderni di semantica,* II, 2, pp. 281-295.

SGROI S. C. (1982) *La linguistica e l'insegnamento delle lingue straniere,* Roma, Armando, pp. 79-115.

SGROI S. C. (1994) I lessici fondamentali e di frequenza della lingua italiana (1927-1993) In: *Diglossia, prestigio e varietà della lingua italiana,* Enna, Il Lunario, pp. 207-245.

SHEILS J. (2005) (a cura di) *Relating Language Examinations to the Common European Framework of Reference for Languages. Reading and Listening Items and Tasks. Pilot Samples.* Strassbourg, Council of Europe.

STAMMERJOHANN H. (1994) L'italiano L2: dalla descrizione alla didattica. In: DE MAURO T. (a cura di), pp. 71-81

STOPPELLI P., PICCHI E. (2001) *Letteratura italiana Zanichelli*. CD-ROM, Bologna, Zanichelli.

STRAMBI B. (2004) Strumenti di verifica delle competenze linguistiche e comunicative degli) apprendenti iniziali. In: L. MADDII (a cura di), *Apprendimento e insegnamento dell'italiano L2 in età adulta,* Atene, Edilingua, pp. 189-201.

THORNTON A. M. (1992) Gli studi sulla leggibilità e la riscrittura in Italia. In: LUCISANO P. (a cura di), pp. 45-53.

THORNTON A. M., IACOBINI C., BURANI C. (1997) *BDVDB – una base di dati sul vocabolario di base della lingua italiana,* Roma, Bulzoni. [2a edizione riveduta ed ampliata]

TRIM, J. M. L. (1978) *Developing a Unit/Credit System of Adult Language Learning,* Strasbourg, Council of Europe

VEDOVELLI M. (1993) *Confronti fra il LIP e le altre liste di frequenza dell'italiano.* In: DE MAURO *et alii,* pp. 119-147.

VEDOVELLI M. (1994) L'italiano parlato dagli italiani e l'italiano appreso dai non italiani. In: DE MAURO T. (a cura di), pp. 87-98.

VEDOVELLI M. (1995) *La lingua italiana d'uso. Il lessico del parlato*, Sagep, Genova.

VEDOVELLI M. (2002a) *Guida all'italiano per stranieri. La prospettiva del Quadro comune europeo per le lingue,* Roma, Carrocci.
VEDOVELLI M. (2002b) *L'italiano degli stranieri. Storia, attualità e prospettive,* Roma, Carrocci.
VEDOVELLI M. (a cura di) (2006) *Linee Guida CILS*, Perugia, Guerra. [disponibile sul sito www.unistrasi.it]
WILLIS D. (1990) *The Lexical Syllabus. A New Approach to Language Teaching*, London-Glasgow, Collins.
ZIPF G. K. (1949) *Human Behaviour and The Principle of Least Effort*, Cambridge, Mass.

Dizionari italiani

ALINEI M. (1962) *Dizionario inverso italiano. Con indici e liste di frequenza delle terminazioni*, The Hague, Mouton.
ALINEI M. (1968-78) *Spogli elettronici dell'italiano delle origini e del duecento,* Bologna, Il Mulino, 20 voll.
ALINEI M. (1973) *Spogli elettronici dell'italiano letterario contemporaneo,* vol I.: Moravia A., *La ciociara*, vol. II: Calvino I., *Il sentiero dei nidi di ragno*, vol. III: Cassola C., *Ferrovia locale*, Bologna, Il Mulino.
BALDELLI I., MAZZETTI A. (1974) *Vocabolario minimo della lingua italiana per stranieri,* Firenze, Le Monnier. [1978[2] con "Dizionarietto italo-somalo"; 1982[3] con "Dizionarietto arabo"]
BALDELLI I., MAZZETTI A., FALCINELLI M., SERVADIO B. (2005) *Dizionario fondamentale della lingua italiana per stranieri,* Firenze, Le Monnier.
BATINTI A., TRENTA W. (1982) *Ricerche sul lessico di base dell'italiano contemporaneo: contributo all'insegnamento dell'italiano come lingua prima e come lingua seconda*, Perugia, Guerra.
BATTAGLIA S., BARBERI SQUAROTTI G. (1961-2002) *Grande dizionario della lingua italiana*, Torino, UTET.
BERTINETTO P. M., BURANI C., LAUDANNA A., MARCONI L., RATTI D., ROLANDO C., THORNTON A. (1995) *Corpus e Lessico di Frequenza dell'Italiano Scritto* (CoLFIS) [disponibile sul sito: http://www.istc.cnr.it/material/database/colfis/]
BORTOLINI U., TAGLIAVINI C., ZAMPOLLI A. (1971) *Lessico di frequenza della lingua italiana contemporanea,* Milano, Garzanti-IBM.
CANNELLA M. (a cura di) (1991) *Il primo Zanichelli. Vocabolario elementare di italiano*, Bologna, Zanichelli.
CASELLI M. C., CASADIO P. (1995) *Il Primo Vocabolario del Bambino. Guida all'uso del questionario MacArthur per la valutazione della comunicazione e del linguaggio nei primi anni di vita,* Milano, Franco Angeli.
CATTANEO A. (1996) *Guida didattica DIB Dizionario di base dellalingua italiana*, Torino, Paravia.
CRESTI E., MONEGLIA M. (a cura di) (2004) *Campionamento Corpus Stammerjohann*. [Consultabile al sito: http://lablita.dit.unifi.it/corpora/descriptions/stam/]
DE MAURO T. (1999) *Grande dizionario italiano dell'uso*, 6 voll, Torino, UTET. (GRADIT)
DE MAURO T. (2000) *Il dizionario della lingua italiana per il terzo millennio*, Torino, Paravia.
DE MAURO T. (2002) *Il dizionario dei sinonimi e contrari*, Torino, Paravia.

De Mauro T. (2003) *Nuove parole italiane dell'uso,* VII vol, Torino, UTET, Supplemento a De Mauro (1999).

De Mauro T. *et alii* (1980) *Il vocabolario di base della lingua italiana*. In: De Mauro T., pp. 146-172 [1989[10], 1991[11], 1997[12]; rist. in *Codice di stile* 1993: 118-152]

De Mauro T., Mancini F., Vedovelli M., Voghera M. (1993) *Lessico di frequenza dell'italiano parlato,* Milano, Etaslibri.

De Mauro T., Moroni G. (1996) *Dizionario di base della lingua italiana, Torino*, Paravia. (= DIB)

De Mauro T., Moroni G. (1997a) *DAIC. Dizionario avanzato dell'italiano corrente, Torino*, Paravia.

De Mauro T., Moroni G. (1997b) *Prime parole. Dizionario illustrato di base della lingua italiana*, Torino, Paravia.

Deho D. (1988) *Italienischer Mindestwortschatz. Die wichtigsten Wörter und Wortverbindungen zum Lernen und Nachschlagen,* Ismaning, Hueber.

Devoto G., Oli G. C. (2002) *Dizionario Devoto-Oli compatto. Dizionario fondamentale della lingua italiana,* Firenze, Le Monnier.

Devoto G., Oli G. C. (2004) *Dizionario della lingua italiana*. *Edizione 2004-2005 con cd-rom*, Serianni L., Trifone M. (a cura di), Firenze, Le Monnier.

Devoto G., Oli G. C. (2007) *Dizionario della lingua italiana*. *Edizione 2008,* Serianni L., Trifone M. (a cura di), Firenze, Le Monnier.

Dizionario fondamentale della lingua italiana (2005) Novara, De Agostini.

Dizionario illustrato per l'uso essenziale della lingua (1978) Torino, SEI. (=D.I.L.)

Dizionario medio italiano (2007) Milano, Garzanti.

Egger P. K. *et alii* (1977a) *Grundwortschatz der italienischen Umgangssprache: für die einfache Laufbahn,* Bozen, Athesia. [1985[4]]

Egger P. K. *et alii* (1977b) *Grundwortschatz der italienischen Umgangssprache: für die mittlere und gehobene Laufbahn,* Bozen, Athesia. [1997[9]]

Gabrielli A. (1981) *Il mio primo dizionario illustrato*, Milano, Mondadori. [Traduzione del *Golden Book Illustrated Dictionary* (1962) New York, Western Publishing Company.]

Galli de' Paratesi N. (1981) *Livello soglia per l'insegnamento dell'italiano come lingua straniera*, Strasbourg, Consiglio d'Europa.

Giovanelli P. (1966) *Grund- und Aufbauwortschatz. Italienisch*, Stuttgart, E. Klett.

Grande dizionario italiano 2008 (2007), Milano, Garzanti.

Guida didattica DAIC. Dizionario avanzato dell'italiano corrente (1997) Torino, Paravia.

Il grande dizionario Garzanti della lingua italiana (2005) Milano, Garzanti.

Juilland A., Traversa V. (1973) *Frequency Dictionary of Italian Words,* The Hague, Mouton.

Katerinov K., Boriosi Katerinov M. C., Sciarone A. G. (1991) *Calcolo di rendimento di liste base: italiano parlato, italiano scritto, livello soglia. Ricerche di linguistica computazionale,* Perugia, Edizioni Guerra.

Knease, T. M. (1931) *An Italian Word List from Literary Sources*, The Sate University of Iowa, Toronto, Toronto U. P.

Laudanna A., Thornton A., Brown G., Burani C., Marconi L. (1995) Un corpus dell'italiano scritto contemporaneo dalla parte del ricevente. In Bolasco S., Lebart L., Salem A. (a cura di), *III Giornate internazionali di Analisi Statistica dei Dati Testuali*. Vol. I, Roma, CISU, pp.103-109.

MARCONI L., OTT M., PESENTI E., RATTI D., TAVELLA M. (1994) *Lessico elementare. Dati statistici sull'italiano letto e scritto dai bambini delle elementari,* Bologna, Zanichelli.

MERZ G. (1987) *Langenscheidts Grundwortschatz Italienisch*, München, Langenscheidt.

MIGLIORINI B. (1943) *Der grundlegende Wortschatz der Italienischen. Die 1500 wesentlichsten Wörter,* Marburg, Elwert.

PUTZER O. (a cura di) (1998a) *Bilinguismo. Zweisprachigkeit. Lessico di base. Grundwortschatz*, Bolzano, Arkadia.

PUTZER O. (a cura di) (1998b) *Bilinguismo. Zweisprachigkeit. Lessico approfondito. Aufbauwortschatz*, Bolzano, Arkadia.

REISKE H. (1974) *VHS-Italienisch Unterricht und VHS-Zertificat Italienisch*, Frankfurt, Pädagogische Arbeitsstelle-Deutscher Volkhochschulverband.

RINALDI P., BARCA L., BURANI C. (2004) Caratteristiche semantiche, grammaticali e di frequenza delle parole del "Primo Vocabolario del Bambino." In: *Psicologia clinica dello sviluppo, 8,* 119-143. [la lista di frequenza è scaricabile dal sito: http://www.istc.cnr.it/material/database/pvbvarless.htm]

ROSSELLI R., EYNARD R. (1996) *Dizionario di base della lingua italiana. Con sinonimi-contrari, schede didattiche,* Torino, SEI.

RUSSO G. A. (1947) A combined Italian Word List. In: *Modern Language Journal*, XXXI, April, 1947, pp. 218-240.

RUSSO J. A. (1962) An Italian Idiom List. In: *Modern Language Journal,* XLVI, January, 1962, pp. 14-21.

SABATINI F., COLLETTI V. (1997) *Dizionario italiano Sabatini Colletti*. *DISC*, Firenze, Giunti.

SAVINO E. (a cura di) (1995) *Il mio primo vocabolario*, Milano, Mursia.

SCIARONE A. G. (1977) *Vocabolario fondamentale della lingua italiana,* Bergamo, Minerva Italica.

SCIARONE A. G. (1995) *Vocabolario fondamentale della lingua italiana,* Perugia, Guerra.

SKINNER L. H. (1935) A Comparative Study of the Vocabulary of Forty Five Italian Textbooks. In: *Modern Language Journal*, XX, November, 1935, 2, pp. 67-84.

STAMMERJOHANN H. (1970), *Strukturen der Rede. Beobachtungen an der Umgangssprache von Florenz*Studi di Filologia Italiana, XXVIII, pp. 295-397.

THOMPSON M. E. (1927) *A Study in Italian Vocabulary Frequency*, University of Iowa Master's Thesis (non pubblicato).

Vocabolario Elettronico della lingua italiana. Il vocabolario del 2000 (1989) Milano, IBM Italia. (=VELI)

ZINGARELLI N. (2001) *Lo Zingarelli minore. Vocabolario della lingua italiana*, Bologna, Zanichelli.

ZINGARELLI N. (2005) *Vocabolario della lingua italiana 2006*, Bologna, Zanichelli.

ZINGARELLI N. (2007) *Vocabolario della lingua italiana 2008*, Bologna, Zanichelli.

Dizionari ungheresi

ARADI A., ERDŐS J., STURCZ Z. (2000) *Küszöbszint. Magyar mint idegen nyelv,* Strasbourg, Európa Tanács.

BÁCSI J., KEREKES J. (2003) Az első osztályos olvasókönyvek szóanyagából készült gyakorisági szótár: "Van szó". In: *Módszertani közlemények*, 43. évf. 2. sz. p. 53-58.

BALÁZS G., P. EŐRY V., KISS G., J. SOLTÉSZ K., T. SOMOGYI M. (1994) *Jókai szótár*, Budapest, Unikornis Kiadó.
BEKE J. (szerk.) (2004) *Zrínyi Miklós magyar életművének szókészlete*, Budapest, Argumentum.
BENKŐ L. (1972) *Juhász Gyula költői nyelvének szótára*, Budapest, Akadémiai Kiadó.
CSER J. (1939) *A magyar gyermek szókincse. Gyakorisági és korszótár*, Budapest, Magyar Pedagógiai Társaság.
CSIRIKNÉ CZACHESZ E., CSIRIK J. (1986) *Újságnyelvi gyakorisági szótár*, Szeged, Juhász Gyula Tanárképző Főiskola.
CSIRIKNÉ CZACHESZ E., CSIRIK J. (2002) *10-16 éves tanulók írásbeli szókincsének gyakorisági szótára*, Budapest, BIP.
FÜREDI M., KELEMEN J. (1989) *A mai magyar nyelv széppprózai gyakorisági szótára (1965-1977*), Budapest, Akadémiai Kiadó.
GÓSY M. (1984) *Hangtani és szótani vizsgálatok hárméves gyermekek nyelvében*, Budapest, Akadémiai Kiadó.
ITTZÉS N. (a cura di) (2006) *A magyar nyelv nagyszótára*, Budapest, Akadémiai Kiadó.
JAKAB L., BÖLCSKEI A. (1993) *Csokonai-szókincstár*, Debrecen, Kossuth Lajos Tudományegyetem.
JAKAB L., BÖLCSKEI A. (2000) *Balassi-szótár*, Debrecen, Kossuth Lajos Tudományegyetem.
JUHÁSZ J., SZŐKE I., O. NAGY G., KOVALOVSZKY M. (szerk.) (1972) *Magyar értelmező kéziszótár*, Budapest, Akadémiai Kiadó.
KÖVÉRNÉ NAGYHÁZI B. (2004) Aradi A., Erdős J., Sturcz Z., *Küszöbszint: magyar mint idegen nyelv (recenzió).* In: *Hungarológiai Évkönyv 5*, pp. 272-279.
Magyar Nemzeti Szövegtár (Corpus nazionale ungherese) http://corpus.nytud.hu/mnsz/
Magyar Történeti Szövegtár (Corpus storico ungherese) www.nytud.hu/hhc/
NEMES Z. (1932) *A magyar nyelv képző-, jel- és ragstatisztikája*, Az Egységes Magyar Gyorsírás Könyvtára 57., Szeged.
NEMES Z. (1933) *A magyar parlamenti nyelv leggyakoribb szavai*. Az Egységes Magyar Gyorsírás Könyvtára 66., Szeged.
NEMES Z. (1941) *Szóstatisztika egymillió szótagot felölelő újságszövegek alapján.* Az Egységes Magyar Gyorsírás Könyvtára 190., Szeged.
PAPP F. (a cura di) (1969) *A magyar nyelv szóvégmutató szótára*, Budapest, Akadémiai Kiadó.
PUSZTAI F. (szerk.) (2003) *Magyar értelmező kéziszótár*, Budapest, Akadémiai Kiadó.
SIMONYI ZS. (1896) A betűk és hangok előfordulásának aránya. In: *Magyar Nyelvőr*. XXV. p. 325.
SOLTÉSZ K., SZABÓ D., WACHA I., GÁLDI L. (a cura di) (1973-1987) *Petőfi Sándor életművének szókészlete I-IV.*, Budapest, Akadémiai Kiadó.
SZENDE T. (1973) *Spontán beszédanyag gyakorisági mutatói*, Budapest, Akadémiai Kiadó.
Szószablya gyakorisági szótár (Dizionario di frequenza WordSword) BME Szociológia és Kommunikáció Tanszék - Média Oktató és Kutató Központ (Centro di Formazione e di Ricerca dei Media della Cattedra di Sociologia e Comunicazione dell'Università Tecnica di Budapest) http://mokk.bme.hu/resources/webcorpus/index_html

Dizionari di altre lingue

Advanced Learner's English Dictionary (2001) Glasgow, Harper Collins.

BAKONYI H. (1934) *Die gebräuchlichsten Wörter der deutschen Sprache für Fremdsprachenunterricht stufenmäßig zusammengestellt,* München.
BALDEGGER M., MÜLLER M., SCHNEIDER G., NÄF A. (1980) *Kontaktschwelle Deutsch als Fremdsprache*. Berlin, Langenscheidt.
BUCHANAN M. A. (1927) *A Graded Spanish Word Book*, Toronto.
Cambridge Advanced Learner's Dictionary (2003) Cambridge, Cambridge University Press.
Cambridge International Dictionary of English (2001) Cambridge, Cambridge University Press.
Collins Cobuild English Dictionary for Advanced Learners (2003) London, Harper Collins.
Collins Cobuild English Language Dictionary (1987) London, Harper Collins. [1995[2], 2001[3], 2003[4], 2006[5]]
Concise Learner's Dictionary (2003) London, Harper Collins.
COSTE D., COURTILLON J., FERENCZI V., MARTINS-BALTAR M., PAPO E. (1976) *Un niveau-seuil*, Strasbourg, Conseil de l'Europe.
DOLCH E. W. (1931) Grade Vocabularies. In: *Journal of Educational Research*, 16, 1, pp. 16-26.
EATON H. S. (1934) *Comparative Frequency List.* New York, International Auxiliary Language Association.
ELDRIDGE R. C. (1911) *Six Thousand Common English Words*, Niagara Falls.
ESTOUP J. B. (1907) *Gammes sténographiques*, Paris.
FAUCETT L. W., PALMER H. E., THORNDIKE E. L., WEST M. P. (1936) *Interim report on vocabulary selection for the teaching of English as a foreign language*, London, King and Son.
FLOOD W. E., WEST M. (1962) *An Elementary Scientific and Technical Dictionary*, London.
GAMBLE W. (1861) *Two lists of selected characters containing all in the Bible and twenty-seven other books,* Shanghai.
GOUGENHEIM G. (1958) *Dictionnaire fondamental de la langue française*, Paris.
GRAHAM E. (1965) *Science Dictionary in Basic English*, London.
HENMON V. A. C. (1924) *A French Word Book Based on a Count of 400.000 Running Words*, Madison.
JUILLAND A., BRODIN D., DAVIDOVITCH C. (1970) *Frequency Dictionary of French Words,* Paris, Mouton.
JUILLAND A., CHANG RODRIGUEZ E. (1964) *Frequency Dictionary of Spanish Words.* The Hague, Mouton.
JUILLAND A., EDWARDS M. H., JUILLAND I. (1965) *Frequency Dictionary of Rumanian Words,* The Hague, Mouton.
KÄDING F. W. (1899) *Häufigkeitswörtebuch der deutschen Sprache,* Berlin, Steglitz.
KENISTON H. (1920) Common words in Spanish. In: *Hispanica 3*, pp. 85-96.
KENISTON H. (1933) *A Basic List of Spanish Words and Idioms*, Chicago.
Longman Dictionary of American English (2001) New York, Longman.
Longman Dictionary of Contemporary English (2005) London, Longman.
MATORÉ G. *et alii* (1963) *Dictionnaire du vocabulaire essentiel. Les 5000 mots fondamentaux*, Paris.
MORGAN B. Q. (1928) *German Frequency word Book. Based on Käding's "Häufigkeitswörterbuch der Deutschen Sprache"*, New York.
Oxford English Dictionary (2004) Oxford, Oxford University Press.

PRESCOTT D. A. (1929) Le Vocabulaire des Enfants et des Manuels de Lecture. In: *Extrait des Archives de Psychologie*, 21, pp. 83-84.
SILIAKUS H. J., SCHLICK W. (1965) Untersuchungen zur deutschen Worthäufigkeit. In: *Deutschunterricht für Ausländer* 15, 46-52.
SLAGTER P. (1980) *Un nivel umbral*, Estrasburgo, Consejo de Europa.
STEGER H., KEIL M. (1972) Wortliste. In: STEGER H. (a cura di) *Das Zertifikat Deutsch als Fremdsprache*, Bonn/Frankfurt, DVV, pp. 13-64.
STERN K. (ed.) (2000) *Longman American idioms dictionary*, New York, Longman.
THORNDIKE E. L. (1921) *The Teacher's Word Book*, New York.
THORNDIKE E. L., LORGE I. (1944) *The Teacher's Word Book of 30.000 Words*, New York.
VANEK J. A. (1979) *The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults*, London, Longman.
VANEK J. A., ALEXANDER L. G. (1975) *The Threshold Level English,* Strasbourg, Council of Europe.
VANEK J. A., ALEXANDER L. G., FITZPATRICK M. A. (1977) *Waystage, an Intermediary Objective Below Threshold Level,* Strasbourg, Council of Europe.
VANEK J. A., TRIM J. L. M. (1980) *Waystage English*, Pergamon Press.
VANEK J. A., TRIM J. L. M. (1991) *Threshold Level 1990*, Cambridge, CUP.
VANEK J. A., TRIM J. L. M (1996) *Vantage Level*, Strasbourg, Council of Europe.
VANDER BEKE G. E. (1929) *French Word Book*, New York.
VERLÉE L. (1954) *Basis-Woordenboek voor de Franse Taal*, Amsterdam.
WEST M. (1935) *Definition Vocabulary*, University of Toronto, Department of Educational Research.
WEST M. (1953) *A General Service List of English Words, London*, Longman.

Appendice

Lista A

Questa lista contiene informazioni sui 2.819 lemmi ricavati dal glossario italiano-ungherese, appartenente al manuale *Progetto italiano 1* (Marin e Magnelli 2006). Tra i lemmi si trovano parole ed espressioni polirematiche, riportati in ordine alfabetico. Le informazioni sui lemmi sono riportate in otto colonne.

– in colonna 1 il lemma. Sono stati registrati come entrate separate gli omonimi grammaticali. Per la distinzione degli omonimi lessicali bisogna consultare i traducenti ungheresi. È indicato inoltre l'accento, qualora questo non cade sulla penultima sillaba, l'articolo e/o il genere grammaticale dei sostantivi nei casi dubbi. Le abbreviazioni *f.* e *m.* indicano, rispettivamente, il genere femminile e maschile.
– in colonna 2 il traducente ungherese. Nel caso dei lemmi polisemici e omonimi lessicali, le accezioni o significati sono separati da punto e virgola.
– in colonna 3 la categoria grammaticale, codificata con lettere secondo la tabella seguente:

Ag	aggettivo	**AgPr**	polirematica avverbiale
Art	articolo		
Avv	avverbio	**AvPr**	polirematica aggettivale
C	congiunzione	**CPr**	polirematica congiunzionale
I	interiezione	**IPr**	polirematica interiettiva
Nu	numerale		
P	pronome	**PPr**	polirematica pronominale
Pz	preposizione	**PzPr**	polirematica preposizionale
S	sostantivo	**SPr**	polirematica sostantivale
V	verbo	**VPr**	polirematica verbale

– in colonna 4 l'unità in cui il lemma appare per la prima volta. Nel caso dei lemmi polisemici e omonimi lessicali, le indicazioni relative alla fonte seguono l'ordine delle accezioni o significati e sono separati da punto e virgola. Le marche utilizzate per rinviare alla fonte, sono riassunte nella tabella sottostante:

Struttura delle unità 0 - 11			
Libro dei testi		***Quaderno degli esercizi***	
Sezioni A-H	**L**	Esercizi relativi all'unità	**Q**
Conosciamo l'Italia		Test finale	**QTF**
Glossario		Test di ricapitolazione	**QTR**
Autovalutazione	**LAV**	(ogni 3 unità)	
Autovalutazione generale	**LAVG**	Test finale generale	**QTFG**
Appendice grammaticale	**LAP**		

Un termine che figura nell'appendice grammaticale all'unità 6, avrà quindi la marca 6LAP. Tutte le voci che compaiono nei brani audio del Libro o del Quaderno, sono stati stampati in corsivo e sono segnalati come LA o QA, rispettivamente.
– in colonna 5 troviamo il rango d'uso dei lemmi secondo il LIP (De Mauro *et al.* 1993). Nel caso dei lemmi polirematici, invece del rango d'uso, è riportata la

frequenza assoluta registrata nel corpus LIP. Sono state contrassegnate con una x le polirematiche assenti nel LIP, ma presenti sulla lista della società Èulogos.

– in colonna 6 sono indicati con una crocetta (#) i lemmi (solo parole semplici, non polirematiche) assenti nel VdB (De Mauro *et al.* 1980)

***

a bocca aperta:	tátott szájjal, csodálkozva	AvvPr	7L	x	
a causa di:	miatt	PzPr	2L	x	
a dire la verità:	az igazat megvallva	AvvPr	2L	x	
a fiori:	virágos	AgPr	9L		
a letto:	ágyba	PzPr	2L	x	
a livello:	valamilyen szinten	AvvPr	7L		
a lungo:	hosszasan	AvvPr	7L		
a mano:	kézzel	AvvPr	8L	7	
a pezzi:	darabokban	AvvPr	8L		
a piedi:	gyalog	AvvPr	2L		
a posto:	rendben	AvvPr	6L	2	
a presto:	hamarosan	AvvPr	1L	1	
a prima vista:	első látásra	AvvPr	9Q	x	
a righe:	csíkos	AgPr	9L		
a stasera:	viszlát ma este	Ipr	2Q		
a testa:	fejenként	AvvPr	5Q	2	
a turno:	felváltva	AvvPr	1L	x	
a:	-nak, -nek; -ba, -be; -ban, -ben; -ra, -re; -on, -en, -ön; -hoz, -hez, -höz; -ig; -kor	Pz	0L	5	
abbassare:	lejjebb vesz, csökkent	V	8Q	2141	
abbastanza:	eléggé, meglehetősen	Avv	1L	258	
abbellire:	díszít	V	5L		#
abbigliamento:	ruházat	S	3L	5101	
abbinamento:	párosítás	S	6L	6505	#
abbinare:	összepárosít	V	0L	3309	#
abbonamento:	előfizetés	S	10L	6069	#
abbonato:	előfizető	S	10L		
abbottonato:	begombolt	Ag	9L		
abbracciare:	megölel, átölel	V	3L	2185	
abbreviazione:	rövidítés	S	3L		#
abilità:	képesség	S	3L	2516	
abitante, *l' (m.)*:	lakos	S	3L		
abitare:	lakik	V	1L	800	
abitazione:	lakhely	S	2LAV	2744	
abito da sera:	estélyi ruha	SPr	8L	x	
abito:	ruha	S	3L	2939	
abituale:	szokásos	Ag	7L		
abituarsi (a):	hozzászokik valamihez	V	9Q	1205	
abitudine, *l' (f.)*:	szokás	S	0LAP	2298	
accanto (a):	valami mellett	PzPr	1L	x	
accendere:	bekapcsol, meggyújt	V	4L	1940	
accento:	hangsúly	S	0LAP	3662	
accessorio*:*	kiegészítő	S	9L	4577	#
accettare:	elfogad	V	2L	766	
accidenti!:	a fenébe	I	8L	3733	
accogliente:	kellemes, barátságos	Ag	4L		
accompagnare:	elkísér	V	8L	1028	
accompagnato:	valami által kísért	Ag	7L	6505	
accordo:	egyetértés	S	7L	196	
accorgersi:	észrevesz	V	9Q	877	
aceto balsamico:	balzsamecet	SPr	8L		
aceto:	ecet	S	6L		
acqua in bocca:	egy szót se róla	AvvPr	9L	x	

acqua:	víz	S	3Q	479	
acquistare:	vásárol	V	5L	1154	
acquisto:	megvásárlás	S	2L	1420	
ad alta voce:	hangosan	AvvPr	11L		
ad esempio:	például	AvvPr	4L	x	
ad essere sincero:	őszintén szólva	AvvPr	10L		
ad un certo punto:	egy bizonyos ponton	AvvPr	8L		
adattato:	átalakított	Ag	11QTGF		
adatto:	megfelelő, alkalmas	Ag	1L	2024	
addirittura:	egyenesen	Avv	7L	383	
addobbare:	feldíszít	V	5L		#
addormentarsi:	elalszik	V	9L	2426	
adesso:	most	Avv	0L	94	
adulto:	felnőtt	S	5L	4002	
aereo*:*	repülő	S	0L	1813	
aerobica:	aerobik	S	11L	6505	#
aeroporto:	reptér	S	2L	4859	
affascinante:	elbűvölő	Ag	7L	4555	
affatto:	egyáltalán nem	Avv	6L	1603	
affermazione:	állítások	S	1L	2816	
affettare:	felszeletel	V	6L		
affittare:	kibérel	V	6Q	2526	
affitto:	albérleti díj	S	2L	4639	
affollato:	zsúfolt	Ag	5L	6505	
affrontare:	szembenéz valamivel	V	4L	642	
affumicato:	füstölt	Ag	6L		
africano:	afrikai	Ag	1Q		
agenda:	határidőnapló	S	2L	3466	
agente, *l'* (*m./f.*):	ügynök	S	4L	3010	
agenzia:	ügynökség, iroda	S	1L	1404	
agevolazione:	kedvezmény	S	5L	6505	#
aggettivo*:*	melléknév	S	0L	2530	#
aggiungere:	hozzáad	V	6L	671	
aggiunta:	kiegészítés	S	3L		
agitato:	viharos	Ag	5L	6505	
aglio:	fokhagyma	S	6L	6505	
agosto:	augusztus	S	3L	2096	
agricolo:	mezőgazdasági	Ag	8L	3199	
agroalimentare:	mezőgazdasági és élelmiszeripari	Ag	8L	6505	#
ai ferri:	rostonsült	AgPr	6LA		
aiutare:	segít	V	8L	442	
aiuto:	segítség	S	3L	1226	
al cento per cento:	száz százalékig	AvvPr	3Q	1	
al contrario:	ellekezőleg	AvvPr	3L		
al dente:	fogravaló tészta	AgPr	4L		
al forno:	sütőben készült	AgPr	6L		
al massimo:	legfeljebb	AvvPr	6L		
al più presto:	minél hamarabb	AvvPr	6Q	x	
al posto (di):	valaki helyében	PzPr	11L		
al solito:	általában	AvvPr	4Q		
alba:	hajnal	S	11L	6144	
albergo:	szálloda	S	0LAP	1421	
albero genealogico:	családfa	SPr	6L	x	
albero*:*	fa	S	0L	1795	
alcuno:	néhány	P	0L	612	
alfabeto:	ábécé	S	0L	6505	
alimentare:	táplálkozási	Ag	3L	3313	
alimentazione:	étkezés	S	8L	4168	
alimento:	élelmiszer	S	8L		
all'arrabbiata:	arrabbiata módra (csípősen)	AvvPr	6L		
all'improvviso:	hirtelen	AvvPr	5LA		
all'interno (di):	valamin belül	PzPr	5L	x	
all'italiana:	olasz módra	AvvPr	7L		

all'ora:	óránkénti	AgPr	5L		
alla bolognese:	bolognai módra	AvvPr	6L		
alla fine:	a végén	AvvPr	0L		
alla grande:	nagyszerűen	AvvPr	9Q		
alla moda:	divatos	AvvPr	9L	x	
alla rinfusa:	ömlesztve	AvvPr	7Q	x	
alla spina:	csapolt	AvvPr	4L		
allegria:	vidámság	S	8QTF	3765	
allegro:	vidám	Ag	1L	4555	
allo stesso tempo:	egyszerre	AvvPr	8L		
allora:	hát akkor, nos	C	1L	39	
alluminio:	alumínium	S	6L		
almeno:	legalább	Avv	3L	268	
alta società:	felső rétegek	SPr	9L	x	
altare, *l'* (*m.*):	oltár	S	10L	6505	
alternativa:	választási lehetőség	S	6L	1847	#
altissimo:	nagyon magas	Ag	10L	6505	
alto:	magas	Ag	0L	440	
altrettanto:	szintén, hasonlóképp	Avv	9L	1983	
altrimenti:	különben	Avv	10L	551	
altro:	másik	Ag	6L	54	
altro:	más, egyéb	P	0L	77	
altrui:	másé	Ag	11L		#
alzarsi:	felkel	V	4L	500	
amalfitano:	Amalfi környéki	Ag	5QTF		#
amare:	szeret	V	2L	975	
amaro:	keserű	Ag	0LAP		
amatissimo:	nagyon kedvelt	Ag	11L	5516	#
ambiente, *l'* (*m.*):	környezet	S	6L	851	
americano:	amerikaiak	Ag	0LA	1371	
amica:	barátnő	S	1L	1166	
amichevole:	baráti, barátságos	Ag	3L		#
amicizia:	barátság	S	9L	1370	
amico:	barát	S	0LA	1497	
ammettere:	bevall	V	4LAP	1948	
amministrazione:	igazgatás	S	10L	779	
ammirare:	csodál	V	8Q	6505	
ammirato:	csodált	Ag	7L	6505	
amore, *l'* (*m.*):	szerelem	S	0Q	469	
analfabeta:	analfabéta	Ag	11QTGF		
analisi, *l'* (*f.*):	analízis, elemzés	S	0LAP	1034	
anche se:	noha, habár	CPr	5L		
anche:	is	C	1L	28	
ancora:	még	Avv	1L	91	
andare a ruba:	gyorsan elkel	VPr	11L	2	
andare d'accordo:	jól kijön egymással	VPr	10Q		
andare in onda:	adásba kerül	VPr	10L	1	
andare matto (per):	bolondul valamiért	VPr	11L		
andare via:	elmegy	VPr	4L	2	
andare:	megy	V	1L	26	
andata e ritorno:	menettérti	AgPr	5L	1	
angolo:	sarok	S	6L	1464	
animale, *l'* (*m.*):	állat	S	8L	1408	
anniversario:	évforduló	S	5L	3157	
anno*:*	év	S	0L	75	
annoiarsi:	unatkozik	V	9Q	6505	
annunciare:	bejelent	V	8L	2318	
antenna:	antenna	S	10L		
antichissimo:	nagyon régi	Ag	8L		
antico:	régi	Ag	4L	1399	
antipasto:	előétel	S	6L	6505	#
antipatico:	ellenszenves	Ag	1L	4281	
anzi:	sőt	C	4L	376	

anziano:	idős	Ag	5L	2565	
anziché:	helyett	C	5L	2137	
anzitutto:	mindenekelőtt	Avv	4L	6402	
aperitivo:	aperitif	S	4L		
aperto:	nyitott	Ag	0L	630	
apertura:	nyitás	S	3L	2288	
apparecchio:	készülék	S	10L	1297	
appartamento:	lakás	S	2L	6505	
appartenere (a):	tartozik valahova	V	4L	1498	
appassionare:	fellelkesít, felkavar	V	11L	3460	
appena:	éppen	Avv	2L	586	
appena:	amint	C	5L	2776	
appendere:	kifüggeszt, fellógat	V	4LAP	6505	
appendice, *l'* (*f.*):	függelék	S	0L	5914	#
apprezzare:	értékel	V	6L	1769	
apprezzato:	kedvelt, amit értékelnek	Ag	6L		
appuntamento:	találkozó	S	2L	980	
appunto:	pontosan, ahogy mondod	Avv	3L	124	
appunto:	jegyzet	S	3L	944	
aprile, *l'* (*m.*):	április	S	3L	1277	
aprire:	nyit	V	1L	332	
arabo:	arab	Ag	6L	3021	
arancia:	narancs	S	4L	6505	
aranciata:	narancslé	S	8Q	6505	
archeologia:	régészet	S	11QTGF		#
archeologo:	régész	S	0LAP		#
architetto:	építész	S	5L	3968	
architettura:	építészet	S	7QTF	4719	
argentino:	argentín	Ag	0L	6505	
argomento:	téma, tárgy	S	3L	409	
aria condizionata:	légkondicionálás	SPr	11QTGF	1	
armadio:	szekrény	S	3L	2398	
armato:	fegyveres	Ag	5L	3320	
aroma:	aroma	S	4L		#
arrabbiarsi:	megharagszik, feldühödik	V	9L	1525	
arrabbiato:	mérges	Ag	6L	4054	
arredamento:	bútorzat, berendezés	S	10QA	6505	
arricchire:	gazdagít	V	6L	3041	
arrivare:	érkezik	V	1L	108	
arrivederci:	viszontlátásra	I	1L	636	
arrivederLa:	viszontlátásra (önöző forma)	I	1L	6505	#
arrivo:	érkezés	S	4Q	2134	
arrosto:	sült	S	5L		
arte, *l'* (*f.)*:	művészet	S	0L	651	
articolo*:*	névelő; cikk	S	0L; 9L	484	
artigianale:	kézműves	Ag	8L	5419	
artista:	művész	S	7L	1707	
artistico:	művészeti	Ag	6L	2020	
ascensore:	lift	S	2L	6505	
ascesa:	felemelkedés, mennybemenetel	S	5L	6505	#
ascoltare:	hallgat	V	0L	331	
ascolto:	hanganyag	S	3L	1618	
aspettare:	vár	V	1L	170	
aspetto:	megjelenés; szempont	S	1L; 8L	445	
aspirina:	aszpirin	S	3Q	6505	#
assaggiare:	megkóstol	V	10Q	3484	
assaporare:	megízlel	V	8L		
assegnato:	kijelölt	Ag	5L		
assistere:	részt vesz	V	11L	1054	
assolutamente:	mindenféleképpen	Avv	8L	418	
assoluto:	abszolút	Ag	8L	1157	
assomigliare:	hasonlít	V	6L	5640	
assumere:	felvesz	V	1L	1081	

assunzione:	mennybemenetel	S	5L	3399	
atmosfera:	légkör	S	2L	2484	
attento:	figyelmes, óvatos	Ag	5Q	601	
attenzione*:*	figyelem	S	0L	371	
attesa:	várakozás	S	5L	1789	
attimo**:**	pillanat	S	6LAV	210	
attività:	feladat	S	0L	613	
attivo:	működő	Ag	5L	3120	
attore, *l'* (*m.*):	színész	S	0LAP	1928	
attrarre:	vonz	V	2LAP	5285	
attraversare:	átszel, keresz+B509tül megy	V	7LAV	2971	
attraverso:	keresztül	C	7L	473	
attrice, *l'* (*f.*):	színésznő	S	7L	6505	
attualità:	aktualitás	S	10L	1645	
augurare:	kíván	V	5L	1033	
augurio:	jókívánság	S	7Q	1257	
aula:	terem	S	4L	1715	
aumentare:	megnövekedik, nő	V	8L	1329	
aumento:	emelkedés, növekedés	S	5L	1059	
ausiliare, *l'* (*m.*):	segédige	S	4L		#
australiano:	ausztrál	Ag	0L	6505	
austriaco:	osztrák	Ag	0QTF	6505	
auto, *l' (f.)*:	gépkocsi	S	0L	2656	
autobiografico:	önéletrajzi	Ag	7L		#
autobus, *l' (m.)*:	busz	S	0L	2913	
automatico:	automata	Ag	2L	2683	
automobilismo:	autóversenyzés	S	10L	6210	#
autore, *l' (m.)*:	szerző	S	11L	649	
autostrada:	autópálya, sztráda	S	11L	4153	
autovalutazione:	önértékelés	S	1LAV		#
autunno:	ősz	S	3L	5357	
avanti Cristo (a.C.):	időszámítás előtt	AgPr	6L		
avanti:	előre	Avv	5L	214	
avere bisogno (di):	szüksége van valamire	VPr	6LA	x	
avere da fare:	dolga van	VPr	2LA		
avere fame:	éhes	VPr	4L		
avere fretta:	siet	VPr	6L		
avere in mente:	gondol valamire	VPr	11L	1	
avere intenzione (di):	szándékozik	VPr	9L		
avere paura:	fél	VPr	5Q		
avere ragione:	igaza van	VPr	1L	x	
avere sete:	szomjas	VPr	4L		
avere torto:	téved	VPr	11LAV		
avere voglia:	kedve van	VPr	2L	3	
avere**:**	van valamije	V	0L	13	
avvenimento:	történés, esemény	S	4L	1724	
avvenire:	bekövetkezik, megtörténik	V	7L	3564	
avventura:	kaland	S	7L	2744	
avvenuto:	megtörtént	Ag	4L	6505	
avverbio:	határozó	S	4L	6505	#
avvicinarsi:	közeledik	V	9Q	1723	
avvocato:	ügyvéd	S	3LAV	497	
azienda:	vállalat	S	4Q	860	
azione*:*	cselekvés, tett	S	0L	992	
azzurro:	kék	Ag	0L	1787	
baciare:	csókol	V	3L	6505	
bacio:	csók	S	3L	827	
bacione:	csók	S	3L	827	
bacon:	angolszalonna	S	6L		#
bagaglio:	poggyász	S	5L	5205	
bagno:	fürdőszoba	S	0L	1153	
balcone:	erkély	S	2L		
ballare:	táncol	V	2L	2183	

bambino:	gyerek	S	1QTF	270	
banale:	banális	Ag	11L	6505	
banana:	banán	S	8L		
banca:	bank	S	2L	773	
banco**:**	pult	S	4L	1224	
bancomat:	készpénzautomata	S	7Q	6505	#
bandiera:	zászló	S	6L	3320	
bar:	bár	S	1L	2011	
barba:	szakáll	S	9Q	6505	
barca:	bárka	S	5L	2209	
barista, *il/la*:	pultos	S	0LAP	6505	
basato (su):	valamin alapuló	AgPr	9L		
base, *la:*	alap	S	6L	388	
basilica:	bazilika	S	7LAV		#
basilico:	bazsalikom	S	6L	6210	
basso:	alacsony	Ag	1L	1824	
bastare:	elég	V	8L	195	
battere:	megdönt	V	7L	1051	
batteria:	dobfelszerelés	S	11L	3692	
battesimo:	keresztelő	S	11QTR	4288	
battuta:	mondat	S	10L	943	
be':	nos	I	4L	6505	#
beato:	jó neki	Ag	4QA	2914	
bellezza:	szépség	S	7L	1561	
bellissimo:	gyönyörű	Ag	5L	126	
bello:	szép	Ag	0L	126	
ben arrivato:	isten hozott	AgPr	1L		
bene:	jól	Avv	1L	45	
benissimo:	nagyon jó, kiváló	Avv	4L	45	
benvenuto:	Isten hozott benneteket	Ag	0L	5516	#
benzina:	benzin	S	5Q	4931	
bere:	iszik	V	2L	1584	
bevanda:	ital	S	4L	6505	
bianco:	fehér	Ag	4L	606	
bibita:	üdítő	S	4L	6505	
biblioteca:	könyvtár	S	1L	1974	
bicchiere:	pohár	S	0LA	2349	
bici, *la:*	bicikli	S	6L	4759	
bicicletta:	kerékpár	S	2L	4759	
biglietteria:	jegypénztár	S	5L		
biglietto d'auguri:	üdvözlőlap	SPr	10Q		
biglietto:	jegy	S	2L	989	
binario:	vágány	S	5L	2368	
biologico:	bio	Ag	8L	5799	#
biondo:	szőke	Ag	1L	3939	
biro, *la*:	golyóstoll	S	11L	6505	#
birra:	sör	S	2QTR	4281	
birreria:	söröző	S	3Q		#
biscotto:	keksz	S	6L	6505	
bisognare:	kell, szükséges	V	3L	152	
bisogno:	szükség	S	5L	288	
bistecca:	bifsztek	S	6LA	5419	
blu:	kék	Ag	1Q	2294	
bocca:	száj	S	1L	2769	
bolognese:	bolognai	Ag	6L	6505	
bordo:	utastér	S	5L	4480	
borsa*:*	táska	S	0L	985	
bottiglia:	üveg	S	0LA	2556	
box:	boksz, fülke	S	3L	4326	#
braccio:	kar	S	1L	2270	
brano:	részlet	S	5L	1384	#
brasiliano:	brazil	Ag	0L	5145	
bravissimo:	nagyon jó, nagyon ügyes	Ag	5L	311	

bravo!:	ügyes, jól van	I	1L	6505	
breve:	rövid	Ag	0L	786	
brevemente:	röviden	Avv	5L	1919	
brioche:	briós	S	6LAV		#
bruno:	barna	Ag	1L	3103	
bruschetta:	fokhagymás pirítós paradicsommal és bazsalikommal	S	6L		#
brutto:	csúnya	Ag	1L	623	
buca delle lettere:	postaláda	SPr	3L		
bufala:	bivaly	S	8L		#
bugia:	hazugság	S	4L		
buonanotte:	jó éjt	I	1LA	1868	
buonasera:	jó estét	I	1LA	576	
buongiorno:	jónapot	I	0LA	406	
buono:	jó	Ag	1L	166	
burro:	vaj	S	6L	6322	
busta:	boríték	S	3L	1627	
cabina:	fülke	S	3L	6505	
cadere:	esik	V	0LAP	539	
caffè corretto:	kávé likőrrel	SPr	4L	x	
caffè macchiato:	kávé kis tejjel	SPr	4L		
caffè:	kávé; kávéház	S	0L; 4L	699	
caffelatte:	tejeskávé	S	4L		#
caffettiera:	kávéfőző	S	4L	6505	#
caimano:	kaljmán, aligátor	S	7L		#
calamità:	csapás, katasztrófa	S	3L	6505	#
calciatore:	futballista	S	7Q	6505	
calcio:	foci	S	0L	1125	
calcolare:	kiszámít	V	9QA	2606	
caldo:	meleg	Ag	2Q	755	
call center:	telefonos ügyfélszolgálat	SPr	5L		
calma:	nyugalom	S	7Q	3029	
calmo:	nyugodt	Ag	5L	5640	
calza:	harisnya	S	5L	2817	
calzatura:	cipő, lábbeli	S	9L		
calzone:	töltött pizzatáska	S	6L	5552	
cambiare:	változtat, változik	V	1QTF	241	
cambiarsi:	átöltözik	V	9L	241	
camera da letto:	hálószoba	SPr	2L	5	
camera:	szoba	S	2L	561	
cameriere:	pincér	S	4L	6505	
camerino:	próbafülke	S	9L	6402	
camicetta:	blúz	S	9L	6144	
camicia:	ing	S	9L	2206	
camino:	kandalló	S	3L		
camminare:	gyalogol	V	4L	1554	
camomilla:	kamilla	S	4L		
campagna:	vidék	S	2L	615	
campanile:	haragntorony	S	3QA	6505	
campionato:	bajnokság	S	4L	6505	
campione:	bajnok	S	0Q	2438	
campo:	térség, mező, tér	S	3LAV	550	
canadese:	kanadai	Ag	1Q	6505	
canale:	csatorna	S	3L	1353	
candidato:	jelölt	S	7L	3228	
cane:	kutya	S	0Q	589	
canone:	díj	S	10L	4615	#
cantante:	énekes	S	2L	3324	
cantare:	énekel	V	5Q	2493	
cantautore:	énekes dalszerző	S	11L	6505	
canzone, *la:*	dal	S	0L	1374	
capace:	képes	Ag	11QTGF	1027	
capello:	haj	S	1L	1644	

capire:	ért	V	1L	67	
capo:	ruhadarab	S	9L	590	
Capodanno:	újév	S	5L	6505	
capolavoro:	remekmű	S	7L	6322	
cappotto:	kabát	S	9L	6505	
cappuccino:	kapucinus	Ag	4L		
cappuccino:	kapucsínó	S	0L	6505	
cappuccio:	kapucsínó	S	4L	6505	
caprese:	capri	Ag	6L		#
carabiniere:	csendőr	S	3L	3524	
carattere:	jellem, személyiség	S	1L	1148	
caratteristica:	jellegzetesség	S	10L	1066	
caratterizzato:	valami által jellemzett	Ag	6L		#
carbonara:	carbonara tészta	S	6LA		#
carbone:	szén	S	5L	6505	
carino:	helyes	Ag	1L	1023	
carissimo:	drága	Ag	3L	422	
carne, la:	hús	S	6LA	1109	
Carnevale:	karnevál, farsang	S	5L	6505	
caro:	kedves, drága	Ag	1L	422	
carriera:	karrier	S	2L	6505	
carrozza:	kocsi	S	5L	6505	
carta di credito:	hitelkártya	SPr	3Q	1	
carta igienica:	WC-papír	SPr	8L	6	
carta:	kártya	S	4L	384	
cartina:	térkép	S	1L	3188	
cartolina:	képeslap	S	4L	2607	
cartone animato:	rajzfilm	SPr	9Q	23	
casa:	ház, otthon	S	0L	117	
caso:	eset	S	3L	141	
cassa:	pénztár	S	4L	1614	
cassetta per le lettere:	postaláda	SPr	3L		
cassetta:	kazetta	S	10L	1338	
cassetto:	fiók	S	3L	1945	
castano:	barna	Ag	1L		#
castello:	vár, kastély	S	3QA	3646	
categoria:	kategória	S	10L	682	
cattivo:	rossz	Ag	5L	1129	
cattolico:	katolikus	Ag	5L	5253	
causa:	ok	S	2L	697	
cavaliere:	lovag	S	5L	6210	
cavallo:	ló	S	5L	1160	
cavatappi:	dugóhúzó	S	6L		
cavolo:	káposzta	S	8L	2503	
celebrare:	ünnepel	V	5L	6505	
celebre:	híres	Ag	7L	4951	
celeste:	világoskék	Ag	7L	3856	
cellulare:	mobiltelefon	S	3L	2813	#
cena:	vacsora	S	0L	1167	
cenare:	vacsorázik	V	2QTR	6505	
cenone:	ünnepi vacsora	S	5L		
centesimo:	cent	S	5L	6210	#
centinaio:	kb. száz	S	10LA	4728	
centrale:	központi	Ag	3Q	794	
centro commerciale:	bevásárlóközpont	SPr	8Q		
centro:	központ	S	0LA	449	
cercare (di):	keres; próbál	V	1Q; 2L	159	
cereale:	gabonafélék	S	6L		#
certamente:	bizonyára	Avv	3Q	393	
certificare:	igazol	V	8L	6505	
certo:	persze	Avv	2L	112	
cesto:	kosár	S	11QTGF	6505	
che bello!:	de jó!	Ipr	1L		

che cosa:	mi, mit	PPr	1L	370	
che:	melyik, milyen	Ag	1L	72	
che:	amely	C	0L	12	
che:	mi, micsoda	P	0L	10	
chef, lo:	konyhafőnök	S	6LA		#
chi:	ki, kicsoda	P	0LA	79	
chiacchierare:	cseveg	V	4L	4228	
chiamare:	hív	V	3L	95	
chiamarsi:	hívják valahogyan	V	0L	95	
chiamata:	hívás	S	3L	1455	
chiarire:	tisztáz	V	1L	1593	
chiaro:	világos, egyértelmű	Ag	7L	307	
chiave, *la*:	kulcs	S	0L	1589	
chiedere:	kérdez	V	0L	122	
chiedersi:	felten magának a kérdést	V	9Q	122	
chiesa:	templom	S	3L	1042	
chilo:	kiló	S	4L	941	
chilometro orario:	kilométer per óra	SPr	5L		
chilometro:	kilometer	S	5L	2217	
chirurgo:	sebész	S	0LAP	6505	
chissà:	ki tudja	Avv	5Q	1019	
chitarra:	gitár	S	4Q	6505	
chiudere:	zár	V	1L	428	
chiuso:	csukott	Ag	7L	1321	
ci:	ott, oda	P	2L	18	
ciao:	szia	I	0L	153	
cibo:	étel	S	4L	4607	
ciclone:	ciklon	S	7L		#
cielo:	ég	S	5L	1503	
cileno:	chilei	Ag	7L		
cinema:	mozi	S	0L	1030	
cinematografico:	mozival, filmekkel kapcsolatos	Ag	7Q	5914	
cinese:	kínai	Ag	10L	2905	
cintura:	öv	S	4Q	3783	
ciò:		P	8L	498	
cioccolata:	kakaó	S	4L		
cioccolatino:	bonbon	S	7L	6505	
cioccolato:	csokolédé	S	4L	6505	
cioè:	azaz, úgymint	C	1L	51	
circa:	körülbelül	Avv	3QTF	971	
circolazione:	forgalom	S	4L	1276	
circostanza**:**	körölmény	S	11L	2449	
citare:	idéz, megemlít	V	7L	1057	
città:	város	S	0LAP	415	
cittadino:	polgár	S	3L	2615	
civile:	polgári	Ag	3L	801	
classe, *la:*	osztály	S	0L	401	
classico:	klasszikus	Ag	1Q	1056	
classifica:	slágerlista	S	11L	3494	
cliente:	ügyfél, vásárló	S	8L	1470	
clientela:	vevőkör	S	9L	6505	#
clima:	klíma, éghajlat	S	0LAP	2093	
coccolare:	dédelget	V	10L	6505	#
codice di avviamento postale, *il*:	postai irányítószám	SPr	3L		
cogliere:	szed, megragad	V	2LAP	1389	
cognome:	vezetéknév	S	0LA	1011	
coinvolgere:	érint	V	11L	1174	
colapasta:	tésztaszűrő	S	6L		#
colazione:	reggeli	S	2L	1602	
collaborare:	együttműködik	V	7L	1885	
collaborazione:	együttműködés	S	8L	2150	
collega, *la/il:*	kolléga	S	0L	634	
collegare:	összeköt	V	5L	1596	

collezione:	gyűjtemény, válogatás	S	11L	6210	#
colloquio:	interjú	S	4L	2231	
colonna sonora:	filmzene	SPr	7L	6	
colonna:	oszlop, hasáb	S	1L	6069	
colorato:	színes	Ag	8QTF	3826	
colore:	szín	S	2L	423	
coloro:	azok, akik	P	11QTF		
colpa:	hiba	S	6L	1344	
coltello:	kés	S	6L		
combinazione, *la:*	kombináció	S	0L		
come mai:	hogy hogy	CPr	4L	45	
come no:	hogyne	AvvPr	5L	29	
come:	hogyan	Avv	1L	78	
come:	mint	C	0L	38	
comfort:	komfort	S	5L		#
comico:	komikus	S	7L	5623	
cominciare:	elkezd	V	0LAP	215	
commedia:	komédia, vígjáték	S	7LA	2118	
commentare:	kommentál	V	10L	3607	
commento:	megjegyzések	S	10Q	1622	
commerciale:	kereskedelmi	Ag	3L	1239	
commercio:	kereskedelem	S	4L	2085	
commessa:	eladónő	S	9L	6505	
comodamente:	kényelmesen	Avv	5L	5419	
comodità:	kényelem	S	5L	3389	
comodo:	kényelmes	Ag	2L	1318	
compagnia**:**	társaság	S	3QTF	1419	
compagno:	társ	S	0L	5050	
compleanno:	születésnap	S	2LA	1857	
completare:	kitölt	V	0L	3161	
completo:	teljes	Ag	4L	1192	
completo:	öltöny	S	9L	6505	
complimento:	gratulálok, elismerésem	S	1L	2017	
comporre:	zenét szerez, szöveget ír	V	11L	1629	
comportamento:	viselkedés	S	6L	743	
comportarsi:	viselkedik	V	7L	1299	
compositore:	zeneszerző	S	7L	5623	#
composto:	összetett	Ag	5L	5914	
comprare:	venni	V	2L	475	
comprensione:	szövegértés	S	9L	3467	
compreso:	magába foglalva	Ag	3L	2160	
computer:	számítógép	S	5LAV	2127	
comunale:	városi, önkormányzati	Ag	3L	1003	
comune:	közös	Ag	1L	838	
comunicare:	kommukál	V	3L	968	
comunque:	minden esetre	C	3L	92	
con:	-val, -vel	Pz	0L	27	
concedere:	átenged	V	4LAP	1510	
concerto:	koncert	S	2L	1214	
concludere:	befejez,lezár	V	3L	837	
concluso:	befejezett, lezárt	Ag	7L		
concorso:	verseny	S	10L		
condimento:	ízesítés	S	6L		#
condizionale, *il*:	feltételes mód	S	11L	6505	#
conduttore, *il*:	műsorvezető	S	10L	6505	#
conferenza:	előadás	S	6L	1489	
conferma:	megerősítés, visszaigazolás	S	5L	3047	
confermare	ellenőriz	V	2L	890	
confermato:	megerősített	Ag	11L		
confezione:	csomagolás, kiszerelés	S	8L	3105	
confrontare:	összehasonlít	V	10L	2026	
congiuntivo:	kötőmód	S	9L		#
congratulazione:	gratuláció	S	10L	6505	#

coniugazione*:*	igeragozás	S	1L		#
conoscenza:	ismeretségek	S	6L	753	
conoscere:	ismer	V	0L	154	
conquistare:	meghódít	V	6L	2861	
consegnare:	bead, átad	V	8LA	1321	
conseguenza:	következmény	S	3L	917	
conservare:	tárol	V	8L	2559	
conservarsi:	eláll	V	8L	2559	
considerare:	tart, vél	V	6L	367	
considerarsi:	tartja magát valamilyennek	V	9Q	367	
consigliare:	tanácsol	V	3L	1229	
consiglio:	tanács	S	4L	540	
consonante, *la*:	mássalhangzó	S	0L		
consultare:	igénybe vesz, segítségül hív	V	0Q	2324	
consumare:	fogyaszt	V	8Q	2122	
consumatore:	fogyasztó	S	11QTF	2067	#
consumazione:	fogyasztás	S	4L		
consumo:	fogyasztás	S	4L	4812	
contare:	számít	V	11QA	1164	
contemporaneo:	egyidejű	Ag	7L	2180	
contenere:	tartalmaz	V	5L	1537	
contenitore:	tartó, edény	S	8L		#
contento:	elégedett	Ag	1L	700	
contenuto:	tartalom	S	8L	1029	
contesto:	kontextus, szövegkörnyezet	S	9L	2133	
continente, *il:*	kontinens	S	10L	3806	
continuamente:	folyamatosan	Avv	8L	1925	
continuare:	folytat	V	1L	291	
conto:	számla	S	6Q	201	
contorno:	köret	S	6LA		
contrario:	ellenkező	Ag	1LAV	1749	
contratto:	szerződés	S	7L	449	
contro:	valami ellen szóló érv	S	10L		
controllare:	ellenőriz	V	1LAV	826	
controllato:	ellenőrzött	Ag	8L	6505	
controllo**:**	ellenőrzés	S	5L	796	
convalida:	érvényesítés	S	2L	6505	#
convalidare:	érvényesít, kezel	V	2L	6505	#
conveniente:	kedvező	Ag	8L	3107	
conversazione:	beszélgetés	S	10L	2314	
convincere:	meggyőz valakit valamiről	V	8L	1234	
convivere:	együtt él	V	11Q	2647	
coordinato:	koordinált	Ag	3L		#
coperto:	borult (ég)	Ag	5L	4156	
copia:	másolat, példány	S	8Q	1765	
coppa:	kehely	S	4L	4620	
coppia:	pár	S	0L	2158	
coprire:	lefed	V	5L	1283	
coprirsi:	betakarózik	V	11QTR	1283	
coraggio:	bátorság	S	11Q	1672	
cornetto:	croissant	S	4L	6505	
corpo:	testület	S	3L	1007	
correggere:	kijavít	V	4L	1946	
correre:	fut (futott)	V	4L	1194	
corretto:	helyes	Ag	0L	1834	
corrispondere:	tartozik	V	0L		
corsa:	út, utazás	S	2L	2400	
corso:	tanfolyam	S	0LA	290	
cortesia*:*	udvariasság	S	1L	1116	
corto:	rövid	Ag	1L	1574	
cosa:	mi, micsoda	P	0L	104	
cosa:	dolog	S	1L	41	
coscia:	comb	S	8L	3349	

così:	így	Avv	1LA	65	
così:	így	C	3L	462	
cosiddetto:	úgynevezett	Ag	5L	1284	
cosmetico:	kozmetikum	S	8LAV		#
costare:	kerül valamennyibe	V	2L	725	
costiera:	tengerpart	S	5QTF		#
costituire:	épít	V	0L	890	
costo:	ár	S	3L	1149	
costoso:	drága	Ag	2L	5799	
costume da bagno, *il:*	fürdőruha	SPr	7L	1	
costume, il:	jelmez; szokás, életmód	S	5L; 10QA	1820	
cotone, *il:*	pamut	S	9L	2689	
cotto:	főtt	Ag	4QA	6505	
cottura:	főzés	S	6L	6505	
cravatta:	nyakkendő	S	9L	5669	
creare:	alkot	V	4L	570	
creato:	megalkotott, tervezett	Ag	5L	6505	
credere:	hisz	V	3L	106	
credito:	hitel	S	3Q	4207	
crema:	krém	S	4L	4960	
cremoso:	krémes	Ag	8L		#
crescere:	nő, nevel	V	4LAP	940	
crisi, *la*:	krízis	S	0LAP	1274	
critica:	kritika	S	5Q	1517	
crociera:	tengeri körutazás	S	5L	6505	#
cronaca:	rovat	S	10L	2176	
cronologico:	időrendi	Ag	8L	2427	#
cruciverba, *il*:	keresztrejtvény	S	1QTF	6505	#
crudo:	nyers	Ag	4L	6144	
cubetto:	kockákra	S	6L		
cucchiaino:	kiskanalak	S	4QA	3216	#
cucchiaio:	kanál	S	6L	4118	
cucina*:*	konyha	S	0L	1460	
cucinare:	főz	V	5L	2568	
cuffia:	fülhallgató	S	11L	4541	
cugino:	unokatestvér	S	6L	2561	
culturale:	kulturális	Ag	6L	904	
cuocere:	főz	V	6L	4555	
cuore, *il:*	szívügyek	S	6L	808	
cura:	odafigyelés	S	4Q	1900	
cuscino:	párna	S	3L	4182	
d'accordo!:	rendben!	Ipr	2L	183	
d'altra parte:	másrészről	CPr	3L	14	
da nessuna parte:	sehova	AvvPr	6L		
da parte:	félre	AvvPr	6L		
da qualche parte:	valahová	AvvPr	5L		
da solo:	egyedül	AvvPr	2L		
da:	óta; -tól, -től; -ból, -ből; -hoz, -hez, -höz; -nál, -nél	Pz	1L	19	
dal vivo:	élőben	AvvPr	11L	1	
dappertutto:	mindenütt	Avv	5L	2225	
dare del Lei:	magáz	VPr	1L	2	
dare del tu:	tegez	VPr	1L	2	
dare fastidio (a):	idegesít	VPr	10L		
dare una mano:	segít valakinek	VPr	3Q	5	
dare:	ad	V	1L	56	
data:	dátum, időpont	S	4L	584	
dato:	adat	S	0L	609	
davanti (a):	valami előtt	PzPr	3L	x	
davanti:	előtt	Avv	3L	470	
davvero:	igazán	Avv	0L	614	
debole:	gyenge	Ag	5L	1848	

decaffeinato:	koffeinmentes	Ag	4L		#
decidere:	dönt	V	4L	432	
decidersi:	dönt, elszán magát	V	9Q	432	
decina:	kb. tíz	Nu	10L		
decisione:	döntés	S	4Q	1320	
deciso:	határozott	Ag	4L	1918	
decodificatore, *il*:	dekóder	S	10L		#
dedicare:	szán valamit valamire	V	10L	1236	
definizione:	definíciók	S	2QTF	1249	
del tutto:	teljesen	AvvPr	6L	1	
delicato:	finom	Ag	8L	2488	
deludere:	csalódást okoz	V	4LAP	3939	
denominazione:	elnevezés	S	8L	4054	#
dente, *il*:	fog	S	4L	2743	
dentifricio:	fogkrém	S	8L	6505	
dentista*:*	fogorvos	S	4L	4785	
dentro:	valamin belül	Pz	3L	456	
depliant, il:	prospektus	S	5QA	4894	#
derby, *il*:	itt: küzdelem	S	3Q	6505	#
descrivere:	leír	V	1L	1152	
descrizione:	leírás	S	2L	1753	#
deserto:	kihalt	Ag	7L	6505	
deserto:	sivatag	S	7L	2353	
desiderare:	kíván,óhajt	V	6L	2121	
desiderato:	kívánt	Ag	3L	6505	
desiderio:	kívánság	S	6L	3433	
designer:	formatervező	S	5L		#
desinenza:	végződés	S	0L		#
destinatario:	címzett	S	3L	4585	#
destinazione:	úti cél	S	5L	4326	#
destra:	jobb (oldal)	S	1L	541	
determinativo:	határozott	Ag	0L		#
detersivo:	mosószer	S	8L		#
dettagliatamente:	részletekbe menően	Avv	9L		#
di conseguenza:	következésképpen	CPr	3L	3	
di niente:	nincs mit, szívesen	Ipr	3L		
di nuovo:	újra, ismét	AvvPr	0L		
di provincia:	vidéki	AgPr	4L		
di seconda mano:	használt	AgPr	4L		
di solito:	általában	AvvPr	1Q	x	
di:	néhány	Art	3L	3826	
di:	-ról, -ról; -ból, -ből; -nak a, -nek a -nak a; -é	Pz	0L	2	
dialogo:	párbeszéd	S	0L	1690	
diario*:*	napló	S	1L	5023	
dicembre, *il:*	december	S	3L		
dieta:	étrend	S	8L	1414	
dietologo:	dietetikus	S	11L		#
dietro:	mögött	Avv	3L	744	
dietro:	mögött, mögé	Pz	3L	759	
difendere:	megvéd	V	4LAP	1004	
difendersi:	megvéd magát	V	9L	1004	
difetto:	hiba	S	11QTGF	3699	
differenza:	különbség	S	1L	496	
differire:	különbözik	V	5L	6505	#
difficile:	nehéz	Ag	0L	364	
difficoltà:	nehézség	S	3L	554	
diffondersi:	elterjed	V	6L	2337	
diffusione**:**	elterjedés, terjesztés	S	10L	2337	
diffuso:	elterjedt	Ag	3L	2249	
digitale:	digitális	Ag	10L		#
digitare:	tárcsázik, billentyűzik	V	3L		#
dimagrire:	lefogy	V	11L	3887	

dimenticare:	elfelejt	V	4Q	494	
dimenticarsi (di):	elfelejtkezik valamiről	V	9Q	494	
diminuzione:	csökkenés	S	5LA	5470	#
dipendere:	függ valamitől	V	4L	495	
dire:	mond	V	1L	16	
direttamente:	közvetlenül	Avv	5LAV	459	
diretto:	közvetlen	Ag	5L	1077	
direttore, *il*:	igazgató	S	1L	947	
direttrice, *la*:	igazgatónő	S	4L	3783	
dirigere:	irányít, vezet	V	4LAP	1962	
disaccordo:	egyet nem értés	S	7L		#
disappunto:	csalódottság	S	8L		#
disco:	lemezek	S	5QTR	1968	
discoteca:	diszkó	S	1L	2309	#
discount:	diszkonáruház	S	8L		#
discussione:	viták, veszekedések	S	4Q	646	
discutere:	vitatkozik	V	4L	742	
disegnare:	rajzol	V	2L	1776	
disegno:	rajz	S	0L	558	
dispiacere, *il:*	sajnálkozás	S	10L	625	
dispiacere:	sajnál	V	2L	6505	
disponibilità:	lehetőség	S	9L	1323	
disporre (di):	rendelkezik valamivel	V	10L	2366	
disposto:	hajlandó	Ag	9L	4114	
distanza:	távolság	S	3Q	1155	
distinguere:	megkülönböztet	V	4LAP	2283	
distrarre:	elvon, elterel	V	2LAP	5505	
distrattamente:	figyelmetlenül, szórakozottan	Avv	10QTF		#
distratto:	szórakozott	Ag	7L	5253	
distruggere:	lerombol	V	4LAP	1519	
distrutto:	letört, fáradt	Ag	6LA	4931	
disturbare:	zavar	V	3L	2172	
dito:	ujj	S	1L	2532	
divano:	dívány	S	3L	4575	
diventare:	válik valamivé	V	4L	226	
diversamente:	eltérően	Avv	7Q	1969	
diverso:	különféle	Ag	3L	164	
divertente:	szórakoztató	Ag	4L	1158	
divertirsi:	szórakozik, mulat	V	9L	885	
dividere:	elválaszt, szétoszt	V	4LAP	585	
dividersi:	szétoszlik	V	9L	585	
divorzio:	válás	S	7L		
dizionario:	szótár	S	6L	6505	
doccia:	zuhany	S	0LA	3545	
docente, *il*:	tanár, oktató	S	9L	3313	#
documentario:	dokumentumfilm	S	10L	6505	#
documento:	dokumentum	S	8Q	595	
dolce, il	édesség	S	4L	4931	
dollaro:	dollár	S	7L	4214	#
domanda	kérdés	S	0L	249	
domani:	holnap	Avv	1L	2712	
domattina:	holnap reggel	S	10L	2042	
domenica:	vasárnap	S	2L	468	
domestico:	házi	Ag	4L	2056	
donna:	nő	S	1L	229	
dono:	ajándékok	S	5L	3416	
dopo:	után	Avv	1L	194	
dopo:	után	Pz	3L	167	
dopodomani:	holnapután	Avv	2Q	2453	
doppio:	dupla	Ag	0L	3108	
dormire:	alszik	V	1L	567	
dormita:	alvás	S	9Q		
dottore, *il*:	doktor	S	3L	680	

dove:	hol	Avv	0L		
dove:	ahol	C	2L	86	
dovere:	kell	V	2L	36	
dozzina:	tucat	Nu	8L	6505	
drammatizzare:	eljátszik	V	4L		#
dritto:	egyenes	Ag	10LA	3729	
dubbio:	kétely	S	3L	785	
due:	kettő	Nu	0L	58	
dunque:	tehát	C	3L	231	
duomo:	dóm	S	3QA	3980	#
durante:	valami közben	Pz	5L	559	
durare:	tart időben	V	4L	946	
duro:	kemény	Ag	7L	1244	
e:	és	C	0L	6	
eccetera:	stb.	Avv	0LAP	184	
eccezionale:	kivételes	Ag	7L	1848	
eccezione:	kivétel	S	0LAP	5357	
ecco:	íme	Avv	0L	55	
economia:	közgazdaságtan	S	4L	887	
economico:	olcsó	Ag	6L	481	
edicola:	újságárus bódé, kioszk	S	1L	6505	
egizio:	egyiptomi	Ag	6L		#
elegante:	elegáns	Ag	9L	3167	
elegantemente:	elegánsan	Avv	9L	6505	
eleganza:	elegancia	S	9L	4747	
elementare:	általános, elemi	Ag	4L	2441	
elemento:	elem, szó	S	4L	505	
elencato:	felsorolt	Ag	11L		
elettronico:	elektronikus	Ag	3L	6144	
eliminare:	kiküszöböl	V	5L	978	
e-mail, *la*:	e-mail	S	1L		#
emergenza:	vészhelyzet	S	3L	3202	
energetico:	energiadús	Ag	8L		#
enfasi, *l'* (*f.*):	nyomaték (beszédben)	S	0LAP		#
enorme:	hatalmas, óriási	Ag	6L	1322	
ente, *l'* (m.):	intézmények, szervezetek	S	3L	1506	
entrambi:	mindkettő	Ag	7L	2945	
entrare:	belép, bemegy	V	2L	199	
entrata:	bejárat	S	3Q	1784	
entusiasmo:	lelkesedés	S	7Q	3081	
entusiasta:	rajongó	Ag	8Q	6505	
Epifania:	Vízkereszt	S	5L		
episodio:	epizód, rész	S	10L	2260	
eppure	pedig	C	7LA	1456	
equilibrato:	kiegyensúlyozott	Ag	8L	4541	#
ereditare:	örököl	V	9L	5914	
errore, *l'* (*m.*):	hiba	S	0L	889	
esagerare:	túlozni	V	9L	3193	
esagerato:	túlzott	Ag	11L	5189	
esame, *l'* (*m.*):	vizsga	S	2L	483	
esattamente:	pontosan	Avv	4QTF	984	
esatto:	igaz, helyes	Ag	2L	762	
escludere:	kizár	V	4LAP	1139	
esempio:	példa	S	0L	115	
esercizio:	gyakorlat	S	0LA	1942	
esigente	igényes	Ag	10L	6505	
esigenza:	igény	S	5L	824	
esistente	létező	Ag	2L	6505	
esistere:	létezik	V	1L	373	
esperienza:	tapasztalat, élmény	S	5QA	490	
esplodere:	felrobban	V	4LAP	4143	
esponente:	képviselő	S	11L	3284	
esporre:	kifejt, kiállít	V	2LAP	3022	

esportare:	exportál	V	6L	6505	
esportazione:	export	S	9L	5552	
espressione:	kifejezés	S	1L	1934	
espresso:	eszpresszó	S	0L	6505	
esprimere:	kifejez	V	3L	522	
esprimersi:	kifejezi magát	V	9L	522	
essere	van, létezik	V	0L	3	
essere abituato (a):	hozzá van szokva valamihez	VPr	7Q		
essere costretto (a):	kénytelen valamit csinálni	VPr	10QTF		
essere giù:	letört	VPr	7L		
essere in giro:	az utcákat járja	VPr	7L		
essere stufo (di):	elege van valamiből	VPr	10L		
esso:	az	P	10L	44	
estate, *l'* (*f.*):	nyár	S	3L	1127	
esterno:	kint	Ag	4L	2181	
estero:	külföld	S	2L	3066	
estivo:	nyári	Ag	5L	6505	
estraneo:	idegen, oda nem illő	Ag	3LAV	3193	
età:	életkor	S	1L	830	
etnico:	etno	Ag	11L	3593	#
etrusco:	etruszok	Ag	6L	6505	
etto:	tíz dekagramm	S	8L		
euro:	euró	S	2L		#
europeo:	európai	Ag	6Q	938	#
Eurostar, *l' (m.)*:	eurostar	S	5L		
evento:	esemény	S	7L	2197	#
eventuale:	esetleges	Ag	4Q	1039	
evidente:	nyilvánvaló, evidens	Ag	11L	1362	
evitare:	elkerül	V	3L	1085	
extralarge:	XL-es méretű	Ag	9LAV	6505	#
extravergine:	extraszűz	Ag	6L		#
fa schifo!:	undorító!	Ipr	10L		
fa:	ezelőtt	Avv	4L	296	
facile:	könnyű	Ag	0LA	501	
facilità:	könnyűség	S	9Q	6505	
facilitazione:	könnyítés, engedmény	S	5L	6505	
facilmente:	könnyen	Avv	9Q	1908	
falso:	hamisak	Ag	1L	1975	
fame da lupi:	farkaséhség	SPr	6LA	x	
fame, *la:*	éhség	S	0Q	1263	
famiglia:	család	S	0L	377	
famigliare:	családi	Ag	4L	1528	#
familiare, *il*:	családtag, rokon	S	11Q	3778	
famoso:	híres	Ag	1L	620	
fan:	rajongó	S	10L	6505	
fantascienza:	sci-fi	S	7L	6505	#
fantasia:	fantázia, képzelet	S	6L	3045	
fantastico:	fantasztikus	Ag	6L	1442	
farcito:	töltött	Ag	5L		#
fare di testa sua:	a saját feje után megy	VPr	2L		
fare il filo (a):	udvarol valakinek	VPr	6L		
fare le ore piccole:	éjszakázik	VPr	9L		
fare lo zapping:	kapcsolgat a csatornák közt	V	10L		
fare male:	megárt	VPr	3Q		
fare parte di:	részt vesz valamiben	VPr	4L	44	
fare quattro passi:	sétál egy kicsit	VPr	11LAV		
fare:	tesz, csinál	V	1L	15	
farfalla:	olasz tésztaféle	S	6L	5552	
farmacia:	gyógyszertár	S	1L	6505	
farsi la barba:	borotválkozk	VPr	9Q	x	
farsi male:	megsérül	VPr	9Q		
fascicolo:	füzet	S	10L	4472	#
fatto:	dolog	S	9Q	139	

favore*:*	szívesség	S	2L	443	
febbraio:	február	S	3L	1447	
febbre, *la*:	láz	S	10Q	2343	
fedele:	hű	Ag	8QA	3161	
felice:	boldog	Ag	7L	1378	
felicità:	boldogság	S	8QTF	5357	
femminile:	nőnem	Ag	0L	3543	
fenomenale:	fenomenális	Ag	7L		#
ferie, *le*:	szabadság	S	5QTF	2505	
fermarsi:	megáll	V	5L	399	
fermata:	megálló	S	0L	6505	
fermo:	mozdulatlan, álló	Ag	10Q	1184	
Ferragosto*:*	augusztus 15., Mária mennybemenetele	S	5L	6505	#
ferroviario:	vasúti	Ag	5L	6505	
festa:	ünnep, buli	S	2L	588	
festeggiare:	ünnepel	V	5L	1606	
festival:	fesztivál	S	4L		
fetta biscottata:	kétszersült, pirítós	SPr	6L	x	
fetta:	szelet	S	6L	3502	
fettuccina:	olasz tésztaféle	S	6L		#
fiaba:	mese	S	10L	3033	
fidanzato*:*	vőlegény	S	1L	3080	
fidarsi (di):	bízik valakiben	V	9L	2073	
fiducia:	bizalom	S	10L	1075	
figlia:	lánya valakinek	S	3Q	1284	
figlio:	fia valakinek	S	6L	263	
figurarsi:	elképzel, gondol	I	3L	846	
fila:	sor	S	10LA	1921	
filetto:	vesepecsenye	S	6L	4894	
film, *il*:	film	S	0LAP	299	
filosofia:	filozófia	S	6Q	1242	
filosofo:	filozófus	S	10L	3173	#
finale, *la:*	döntő	S	10L	4939	
finale:	záró	Ag	0L	1073	
finalmente:	végre	Avv	2QTR	1811	
finanziato:	fizetett	Ag	10L		#
fine settimana, *il:*	hétvége	SPr	2L	2	
fine, *la:*	vége valaminek	S	0L	220	
finestra:	ablak	S	0L	2128	
finezza:	finomság, kifinomultság	S	9L		#
finire:	befejez	V	1L	176	
fino (a):	valameddig	PzPr	3L	1	
finora:	eddig	Avv	6L	1763	
fioraio:	virágárus	S	8L	6505	
fiore, *il*:	virág	S	6Q	1172	
firma:	cég	S	9L	1337	
firmare:	jegyez	V	9L	847	
firmato:	márkás	Ag	9L	3223	
fisso:	fix, stabil	Ag	3L	1669	
focaccia:	lepényféle	S	6L		#
foglia:	falevél	S	3Q	5088	
fondare:	alapít	V	10L	1881	
fondato:	alapított	Ag	10L		
fontana:	kút	S	1LAV	3472	
forchetta:	villa	S	6L	6505	
forestale:	erdészeti	Ag	8L	6505	#
forma:	alak, forma	S	0LAP	533	
formaggio:	sajt	S	6L	1852	
formale:	formális, hivatalos	Ag	1L	2458	
formare:	alkot, képez	V	0L	1278	
formazione:	képzés	S	4L	939	
formulare:	alkot, képez	V	2L	3277	#

fornire:	megad	V	5L	993	
forno:	sütő	S	6L	6505	
forse:	talán	Avv	1L	135	
forte!:	nem semmi!	I	7L		
forte:	erős	Ag	4L	519	
fortuna:	szerencse	S	2L	873	
fortunato:	szerencsés	Ag	1L	1637	
forza!:	hajrá!	I	10L		
foto, *la*:	fotó	S	0L	1078	#
fotografia:	fénykép	S	10L	538	
fra l'altro:	többek közt	AvvPr	9L	18	
fra:	között; múlva	Pz	1L; 5L	186	
francese:	francia	Ag	0Q	1135	
franchising:	fanchising	S	9L		#
francobollo:	bélyeg	S	3L	6505	
frase, *la*:	mondat	S	0L	704	
frate, il:	szerzetesek	S	4L	5552	
fratellino:	öccs	S	6L	5552	
fratello:	fiú testvér	S	0L	656	
freddo:	hideg	Ag	4L	1706	
frequentare:	jár valahova, valakihez	V	6L	1752	
fresco:	friss	Ag	6L	4939	
fretta:	sietség	S	4L	2299	
friggere:	olajban süt	V	6L	3472	
frigorifero:	hűtőszekrény	S	7L	6505	
fronte, *la:*	homlok	S	1L	549	
frutta:	gyümölcs	S	6L	1726	
fruttivendolo:	zöldséges	S	8L	6505	#
frutto di mare:	a tenger gyümölcse	SPr	6L		
frutto di stagione:	idénygyümölcs	SPr	6L	x	
frutto:	gyümölcs	S	6L	1825	
fumare:	dohányzik	V	1Q	1377	
fungo:	gomba	S	6LA	3699	
funzionare:	működik	V	5L	537	
fuoco:	tűz	S	0LAP	1545	
fuori città:	városon kívül	AvvPr	8L		
fuori:	bárban, nem otthon	Avv	4L	203	
furto:	lopás	S	4L		
fusillo:	olasz tésztaféle	S	6L		#
futuro:	jövőbeli	Ag	5L	2800	
futuro:	jövő	S	5L	1084	
galleria:	galéria, képtár	S	3QA	4913	
gallina:	tyúk	S	11QTGF	6505	
gambero:	rák	S	8L		
gara:	verseny	S	5L	1416	
garage, *il*:	garázs	S	3Q	4472	
garanzia:	garancia	S	8L	1986	
gatto:	macska	S	0L	922	
gel, il:	zselé	S	8QA	6505	#
gelato:	fagylalt	S	0L	4812	
geloso:	féltékeny	Ag	6Q	6505	
generale:	általános	Ag	3L	1064	
generalmente:	általában	Avv	3L	2079	
genere, *il:*	műfaj	S	7L	310	
geniale:	zsális	Ag	7L		
genio:	géniusz, zs	S	7L	6505	
genitore, *il*:	szülő	S	2L	857	
gennaio:	január	S	3L	1434	
gente, *la*:	emberek	S	2L	175	
gentile:	kedves	Ag	1L	2278	
genuino:	természetes	Ag	8L	5101	
geografico:	földrajzi	Ag	8L	2477	
gestito (da):	valaki által üzemeltetett	Ag	3L		#

ghiaccio:	jég	S	4L	1992	
già:	már	Avv	1L	85	
giacca:	blézer	S	9L	2860	
giallo:	sárga; krimi	Ag	0L; 7L	2303	
giapponese:	japán	Ag	3Q	2482	
giardino:	kert	S	1Q	2666	
ginnastica:	torna	S	5QTR	5101	
ginocchio:	térd	S	8L	3759	
giocare:	játszik	V	2L	557	
giocatore, *il*:	játékos	S	1Q	1502	
Giochi Olimpici:	olimpiai játékok	SPr	4L	x	
gioia:	öröm	S	8L	1194	
gioiello:	ékszerek	S	9L	4723	
giornalaio:	újságárus	S	3L	6505	#
giornale, *il:*	újság	S	0L	314	
giornaliero:	napi	Ag	4L	4998	
giornalista*:*	újságíró	S	2L	1701	
giornata:	nap	S	0Q	453	
giorno:	nap	S	1L	100	
giostra:	lovagi torna	S	5L		
giovane*:*	fiatal	S	0L	820	
giovanile:	fiatalos	Ag	9L	3007	#
giovedì, *il*:	csütörtök	S	2L	715	
girare:	fordul; forgat	V	2L; 7L	665	
giro:	kör	S	5L	489	
gita:	kirándulás	S	2L	2028	
giù:	le	Avv	6Q	374	
giubbotto:	dzseki, kabát	S	9L		
giugno:	június	S	3L	791	
giungere:	elér, megérkezik	V	4L	1580	
giungla:	dzsungel	S	11QA		#
giustificare:	magyaráz	V	6L	2974	
giusto:	helyes, igaz	Ag	0L	244	
globale:	globális	Ag	0L	4131	#
glorioso:	dicső	Ag	7L	5914	
glossario:	szószedet	S	2L		#
gnocco:	burgonyás galuska	S	6L	6505	#
gnomo, *lo:*	törpe	S	0LAP		#
godere (di):	élvez valamit	V	4L	4998	
gola:	torok	S	0Q	5669	
golf, *il*:	kardigán	S	9L	3733	#
golfo:	öböl	S	3Q	1046	
gonna:	szoknya	S	0L	6505	
governo:	kormány	S	10L	599	
graduale:	fokozatos	Ag	5LA		#
grammaticale:	nyelvtani	Ag	0LAP	6505	#
grammo:	gramm	S	6L	1961	
Gran Premio:	nagydíj	SPr	10L	1	
granché:	valami nagy dolog	Avv	7L	3543	#
grande schermo:	mozivászon	SPr	7L	x	
grande:	nagy	Ag	1L	125	
grasso:	zsíros	Ag	8L	5732	
grattugia:	reszelő	S	6L		#
grattugiare:	reszel	V	6L	6505	
grattugiato:	reszelt	Ag	6L		#
gratuitamente:	ingyen	Avv	10L	4913	#
gratuito:	ingyenes	Ag	3L	5640	#
grave:	komoly	Ag	2L	855	
grazie (a):	valaminek köszönhető	PzPr	6L	17	
grazie:	köszönöm	I	1L	130	
greco:	görög	S	0LAP	2698	
grigio:	szürke	Ag	6L	1657	
grosso:	nagy	Ag	8L	284	

gruppo:	csoport	S	3LAV	277	
guadagnare:	pénzt keres	V	8QTR	1253	
guanto:	kesztyű	S	3L		
guardare:	néz	V	1L	90	
guerra:	háború	S	5L	294	
guidare:	vezet (járművet)	V	6Q	2310	
gusto:	ízlés	S	0L	3253	
gustoso:	ízletes	Ag	8L		#
idea:	gondolat, ötlet	S	0Q	269	
ideale:	ideális	Ag	2L	1931	
idolo:	bálvány	S	11L		
idratante:	hidratáló	Ag	8QA		#
ieri:	tegnap	Avv	4L	282	
ignoto:	ismeretlen	Ag	7L	6505	
illuminato:	kivilágított	Ag	5L	6505	
illustrazione:	illusztráció	S	4L	6505	
imbucare:	bedob (levelet postaládába)	V	3L	6505	#
immaginare:	elképzel	V	1L	638	
immagine, *l'* (f.):	kép	S	0L	547	
imparare:	megtanul	V	1Q	1045	
impaziente:	türelmetlen	Ag	7L		
impegnato:	elfoglalt	Ag	9L	2769	
impegno:	elfoglaltság	S	2L	708	
impensabile:	elképzelhetetlen	Ag	4L		#
imperativo:	felszólító mód	S	10L		#
imperatore:	császár	S	7L	4652	
imperfetto:	imperfetto (igeidő)	S	7L		#
impero:	birodalom	S	9L	1860	#
impersonale:	személytelen	Ag	8L		#
impiegato:	alkalmazott	S	5L	2483	
importante:	fontos	Ag	1L	158	
importanza:	fontosság	S	11L	707	
importare:	számít, fontos	V	8L	934	
importato:	importált	Ag	4L		
impossibile:	lehetetlen	Ag	2LA	1010	
improvviso:	hirtelen	Ag	3L	2475	
in altri termini:	más szóval	CPr	3L		
in assoluto:	abszolút, egyértelműen	AvvPr	8L		
in base (a):	valami alapján	PzPr	6L	x	
in contante:	készpénzben	AvvPr	9L		
in continuazione:	folyton	AvvPr	7L		
in corso:	folyamatban lévő	AgPr	7L		
in diretta:	egyenes adásban	AvvPr	10L	15	
in fondo:	hátul	AvvPr	9L		
in fretta:	gyorsan, sietve	AvvPr	4L		
in gamba:	belevaló, ügyes	AgPr	10L		
in genere:	általában	AvvPr	2L		
in media:	átlagosan	AvvPr	10L		
in ogni caso:	minden esetben	AvvPr	6L		
in particolare:	különösen	AvvPr	3L		
in piedi:	állva	AvvPr	4L	1	
in pratica:	gyakorlatban	AvvPr	4L	x	
in punto:	pontosan	AvvPr	4L	x	
in qualche modo:	valahogyan	AvvPr	8LA		
in tutto:	összesen	AvvPr	4L		
in:	-ba, -be; -ban, -ben; -ra, -re; -on, -en, -ön; alatt (idő)	Pz	0L	9	
incarico:	megbízás	S	0LAP	2931	
incasso:	bevétel	S	7L	3175	
incendio:	tűzeset	S	3L		
incertezza:	bizonytalanság	S	3L	3350	
incidente, *l'* (*m.*):	baleset	S	7Q	2179	
includere:	magába foglal	V	5L	2410	

incontrare:	találkozik	V	2QTF	544	
incontro:	találkozás	S	1L	555	
incredibile:	hihetetlen	Ag	4L	731	
incrocio:	kereszteződés	S	10LA	6505	
indagine, *l'* (*f.*):	kutatás, felmérés	S	11L	1727	
indeciso:	bizonytalan	Ag	6LA	6505	
indeterminativo:	határozatlan	Ag	1L		#
indicare	mutat	V	0LAP	631	
indicativo:	kijelentő mód	S	1L		#
indicazione:	jelzés	S	8L	886	#
indietro:	hátra, vissza	Avv	7Q	906	
indimenticabile:	felejthetetlen	Ag	7L	6505	
indiretto:	részes	Ag	10L	4165	
indirizzo:	lakcím	S	6Q	839	
individuare:	megkülönböztet	V	8L	924	
indossare:	visel, hord	V	9L	5008	
infanzia:	gyermekkor	S	3L	3810	
infatti:	valóban	C	4L	133	
infelice:	boldogtalan	Ag	9Q	6505	
infine:	végül	Avv	2L	2216	
infinito:	főnévi igenév	Ag	2L	2843	
influenza:	hatás, befolyás	S	11L	3245	
influenzato:	befolyásolt	Ag	11L	6505	#
informale:	informális, laza, baráti	Ag	1L	5914	#
informare:	tájékoztat	V	8L	867	
informarsi:	tájékozódik	V	9LAV	867	
informato:	jól értesült	Ag	3L		
informazione, *l' (f)*:	információ	S	1L	482	
ingegnere:	mérnök	S	3L	1823	
inglese:	angol	Ag	0L	1926	
ingrediente, *l'* (m.):	hozzávaló	S	6L	3320	#
ingresso:	bejárat	S	11QTGF	1938	
iniziale:	kezdeti	Ag	5L	1894	
inizialmente:	kezdetben	Avv	11LAVG	2166	
iniziare:	elkezd	V	2QTR	447	
inizio:	kezdet	S	1L	410	
innamorare:	szerelmet kelt	V	7L	5101	
innamorarsi (di):	beleszeret valakibe	V	9L	5101	
inoltre:	továbbá	Avv	2L	1910	
insalata:	saláta	S	6L	2394	
insegna:	felirat	S	4L		
insegnante:	tanár	S	1L	893	
insegnare:	tanít	V	10L	622	
inserire:	beilleszt	V	1L	770	
inserto:	melléklet	S	10L	6505	#
insieme (a):	valakivel együtt	PzPr	4L	x	
insieme:	együtt	Avv	2L	251	
insistere:	ragaszkodik valamihez	V	4LAP	1216	
insomma:	végeredményben	C	4L	71	
intanto:	közben	Avv	5Q	386	
intelligente:	értelmes	Ag	1L	1485	
intenso:	mozgalmas	Ag	4L	5328	
Intercity, *l' (m.)*:	intercity	S	5L		#
interessante:	érdekes	Ag	1L	716	
interessare (a):	érdekel valakit	V	10L	341	
interessato:	érdekelt	Ag	3L	1899	
interesse, *l'* (*m.*):	érdeklődés; érdek	S	7Q; 8QTR	435	
internazionale:	nemzetközi	Ag	7L	1172	
internet point:	internet kávézó	SPr	3L		
interno:	belső	Ag	5L	729	
Interno:	belügy	S	3L	404	
intero:	egész, teljes	Ag	4L	1302	

interpretare:	alakít	V	7L	1250	
interpretazione:	alakítás, színészi játék	S	7L	3744	
interprete, *l' (m./ f.)*:	előadó	S	7L	5357	
interregionale:	tartományok közti	Ag	5L	6505	#
interrompere:	megszakít	V	10L	1354	
interrotto:	megszakított	Ag	7L	6505	
interurbano:	távolsági hívás	Ag	3L	6505	#
intervallo:	szünet	S	10L	6210	
intervento:	beavatkozás, közbelépés	S	3L	438	
intervista:	interjú	S	2L	1636	
intervistare:	interjút készít	V	2L	3440	#
intimo:	legközelebbi	Ag	2L	3216	
intorno (a):	valami körül	PzPr	3L	x	
intransitivo:	tárgyatlan	Ag	4L		#
introdurre:	bevezet	V	6L	1360	
introduttivo:	bevezető	Ag	0L	4422	#
inutile:	felesleges, hiábavaló	Ag	9L	822	
invece (di):	valami helyett	PzPr	4L		
invece:	viszont	Avv	2L	99	
inventare:	feltalál	V	4L	843	
invenzione:	találmány	S	6L	3514	
invernale:	téli	Ag	4L	4198	
inverno:	tél	S	3L	2833	
investire:	elüt, elgázol	V	7LAV	2065	
inviare:	elküld	V	10L	1733	
invitare:	meghív	V	2L	750	
invito:	meghívás	S	2L	2009	
involtino:	göngyölt hús	S	6L		#
io:	én	P	0L	11	
ipermercato:	hipermarket	S	8L		#
ipotesi, *l'* (*f.*):	hipotézis, feltételezés	S	0LAP	840	
ipotetico:	feltételes	Ag	5L	5552	#
irlandese:	ír	Ag	1Q		
ironico:	ironikus	Ag	7L	5914	
irregolare:	rendhagyó	Ag	0L	6505	
irregolarità:	szabálytalanság	S	11L		#
irrinunciabile:	amiről nem lehet lemondami	Ag	4L		#
iscriversi:	beiratkozik	V	9Q	1559	
isola, l':	sziget	S	0LA	3078	
istruzione:	utasítás	S	6L	3274	
italiano:	olasz	Ag	0L	325	
italiano:	olasz	S	1L	433	
italo-americano:	olasz-amerikai	Ag	11QA		#
jeans, *i:*	farmer	S	7L	4024	#
l'altro ieri:	tegnapelőtt	AvvPr	4L		
là:	ott	Avv	9L	168	
ladro:	tolvaj	S	7L	2401	
lago:	tó	S	2L	4124	
lamentarsi:	panaszkodni	V	10L	1484	
lampada:	lámpa	S	3L	4868	
lana:	gyapjú	S	9L	2668	
lancetta:	óramutató	S	2L		#
lasagna:	lasagne	S	6LA		
lasciare:	hagyni	V	2L	160	
lasciarsi:	elhagy egymást	V	9L	160	
lato:	oldal	S	10L	752	
latte, il:	tej	S	0LA	2351	
lattina:	alumínium italdoboz	S	4L	6505	
lattuga:	saláta	S	8L	6505	
laurea:	diploma	S	5L	1523	
laurearsi:	diplomát szerez	V	9Q	5732	
laureato:	diplomás	Ag	4L	5177	
lavanderia:	mosoda	S	6Q	6505	

lavarsi:	megmosakszik	V	4L	862	
lavatrice, la:	mosógép	S	8QA	2804	
lavorare:	dolgozik	V	0L	174	
lavorato:	megmunkált	Ag	9L		
lavorazione:	feldolgozás	S	8L	5036	
lavoro pubblico:	állami beruházás	SPr	3L		
lavoro:	munka	S	1L	118	
legato:	összetartó	Ag	6L	2558	
leggenda:	legenda	S	6L	6505	#
leggendario:	legendás	Ag	4L		#
leggere:	olvas	V	0L	183	
leggero:	könnyű	Ag	0LA	1459	
legionario:	légionárius	S	10L		#
lei:	ő (nőnem); Ön	P	0L; 1LA		
lente, la:	lencse	S	9QA	6505	
lento:	lassú	Ag	6L	228	
lettera:	levél	S	0L	357	
letteratura:	irodalom	S	11QTGF	1012	
Lettere, *le*:	bölcsészkar	S	1L	2594	
letto:	ágy	S	2L	765	
lettore, *il*:	olvasó	S	10Q	2145	
lezione:	tanóra	S	1Q	663	
lì:	ott	Avv	4L	110	
liberarsi:	megszabadul	V	9Q	1651	
libero:	szabad	Ag	2L	397	
libertà:	szabadság	S	4Q	821	
libreria:	könyvespolc	S	0L	6505	
libro:	könyv	S	0L	264	
liceo:	gimnázium	S	6Q	1349	
limite, *il*:	korlát	S	9L	664	
linea:	vonal	S	2L	356	
lingua:	nyelv	S	0L	478	
linguaggio:	nyelvezet	S	10QA	644	
linguine:	olasz tésztaféle	S	6L		
liquore, *il*:	likőr	S	4L		
lirico:	opera-	Ag	8L	6505	#
lista:	lista	S	4L	1376	
listino:	árlista	S	4L	2597	
litigare:	veszekszik	V	4Q	2802	
litro:	liter	S	7L	5253	
livello:	szint	S	5L	413	
locale, *il:*	helyiség	S	4L	2045	
locale:	helyi	Ag	5L	950	
località:	helység	S	3LAV	3573	
localizzare:	elhelyez, leírja a helyzetét	V	4LAV	4081	#
locandina:	plakát	S	11L		#
logico:	logikus	Ag	10L	915	
londinese:	londoni	Ag	9Q		
lontano (da):	távol valamitől	Avv	5L	1711	
lontano:	távol	Ag	1QTF	1087	
loro:	ők	P	0L	252	
lotto:	lottó	S	5L	6505	
luce:	fény	S	0L	930	
luglio:	július	S	7L	1500	
lui:	ő (hímnem)	P	0L		
luminoso:	világos	Ag	2L		
lunedì, *il*:	hétfő	S	0L	583	
lungo:	hosszú	Ag	1L	385	
luogo:	hely	S	2L	626	
lupo:	farkasok	S	6LA	3939	
lusso:	luxus	S	9L	5158	
lussuoso:	luxus	Ag	1L		#
ma dai!:	na menj már!	Ipr	2L		

ma:	de	C	0L	23	
macché!:	ugyan már!	I	4L	3605	
macchina:	autó, berendezés	S	0L	304	
macchinetta:	berendezés	S	2L		
macellaio:	hentes	S	8LAV	6505	
madre, *la:*	anya	S	2L	486	
maestro:	mesterek	S	6L	4771	
magari:	esetleg	Avv	0LA	340	
maggio:	május	S	3L	895	
maggiore**:**	nagyobb	Ag	9L	669	
magico:	varázs-, varázslatos	Ag	10L	3545	
maglia:	pulóver	S	9L	4747	
maglietta:	póló	S	7L	3887	#
mai:	soha	Avv	1Q	134	
male, *il:*	fájdalom	S	4L	689	
male:	rosszul	Avv	0Q	312	
malgrado:	valami ellenére	Pz	10L	6505	
mamma:	mama	S	0L	283	
manager:	menedzser	S	7L	6505	#
mancanza:	hiány	S	6L	1175	
mancare:	hiányzik	V	1L	353	
mandare a quel paese:	elküld melegebb éghajlatra	VPr	11L		
mandare:	küld	V	3L	212	
mangiare:	eszik	V	1L	206	
manica:	ruhaujj	S	9L	6144	
maniera:	módszer	S	8L	350	
mannaggia!:	a francba	I	7L	1542	#
mano, la:	kéz	S	0LA	230	
mantenere:	eltart, megtart	V	2LAP	635	
mappa:	térkép	S	5L		#
marca:	márka	S	8L	1905	
marchio:	márka	S	8L	4695	#
marciapiede, *il*:	járda	S	4L	5205	
mare, *il:*	tenger	S	2L	686	
margherita:	margaréta	S	0L	6505	
marinara:	pizza tengerész módra	S	6L		#
marito:	férj	S	6L	825	
marmellata:	lekvár	S	8L	6322	
marocchino:	marokkói	Ag	0L	6505	
marrone:	barna	Ag	9L	6505	
martedì, *il*:	kedd	S	2L	1041	
marzo:	március	S	3L	1210	
maschera:	maszk	S	0L	5669	
mascherarsi:	álarcot ölt, beöltözik	V	5L		#
maschile:	himnem	Ag	0L	4792	
maschio:	vártorony	S	3QA	3530	
mass media, *i*:	tömegkommunikációs eszközök	SPr	10L	1	
massimo:	legnagyobb	Ag	6L	2640	
matematica:	matematika	S	4Q	1915	
matrimonio:	házasság, esküvő	S	6Q	1474	
mattina:	reggel	S	1L	437	
matto:	bolond	S	11L	2831	
maturazione:	érlelés	S	8L	6505	
maturo:	érett	Ag	8L	6210	
mazzo:	csokor	S	8L		
meccanico:	autószerelő	S	3L	4518	
Medicina:	orvostudomány	S	1L	2491	
medicinale, *il:*	gyógyszer	S	8L	6505	#
medico:	orvos	S	0LAP	1864	
medio:	közepes	Ag	4L	875	
Medioevo:	középkor	S	6L	4455	#
mediterraneo:	mediterrán	Ag	7L	5914	
meglio:	legjobb	S	0LA	3416	

meglio:	jobban	Avv	8L	187	
mela:	alma	S	6L	3958	
melodico:	dallamos	Ag	11L		#
membro:	családtag	S	6L	3652	
meno male:	még jó	Ipr	8L	1	
meno:	kevésbé	Avv	2L	132	
mensa:	menza	S	4L		
mensile, il:	havilap	S	10QA	4711	#
mente, *la:*	elme, ész	S	11L	863	
mentre:	míg, mialatt, miközben	C	2L	324	
menù:	menü	S	4L	6505	#
meraviglioso:	csodálatos	Ag	11L	1496	
mercato:	piac	S	8L	526	
mercoledì, *il*:	szerda	S	2L	1110	
merenda:	uzsonna	S	6L	3642	
mescolare:	kever	V	6L	6505	
mese, *il:*	hónap	S	2L	193	
messaggio:	üzenet	S	5Q	790	
mestolo:	merőkanál	S	6L		
metà:	fele valaminek	S	6QTF	987	
metallo:	fém	S	9QA	5008	
meteo:	időjárásjelentés	S	5LA		#
metodo:	módszer	S	8L	1015	
metro:	méter	S	10L		
metrò:	metró	S	1L	713	#
metropolitana:	metró	S	2L	4809	
mettere:	tesz, rak	V	0L	68	
mettersi (a):	felvesz ruhát; elkezd	V	7LA; 9L	68	
mettersi d'accordo:	megegyezik	VPr	9Q	x	
mettersi in testa:	a fejébe vesz valamit	VPr	11QTGF		
mettersi insieme:	összejön	VPr	9L		
mezzanotte:	éjfél	S	2L	1739	
mezzo:	fél	Ag	0L	451	
mezzo:	eszköz	S	1L	232	
mezzogiorno:	dél	S	2L	1508	
mica:	egyáltalán, ám	Avv	8L	425	
microfono:	mikrofon	S	11L	1212	
miele, *il*:	méz	S	6L		
miglialio:	kb. ezer	Nu	10L	2615	#
miglioramento:	javulás	S	5LA	2994	#
migliorare:	javul	V	5Q	1248	
migliore:	legjobb	Ag	1L	568	
milanese:	milánói	Ag	4L	3271	
minerale:	ásványi, ásvány-	Ag	4L	4555	
mini:	mini, rövid	Ag	0L	6505	#
ministero:	minisztérium	S	3L	1025	
ministro:	miniszter	S	10Q	1979	
minuto:	perc	S	1L	405	
mio:	enyém	Ag	1L	98	
miopia:	rövidlátás	S	9QA		#
miracolo:	csoda	S	10LAV	4660	
mistero:	rejtély	S	9Q	3127	
misto:	vegyes	Ag	6L	1016	
mite:	enyhe	Ag	8L		
mittente, *il:*	feladó	S	3L	6505	
mobile, *il*:	bútor	S	3L	2670	
mobile:	mobil	Ag	4Q	3424	
moda:	divat	S	0L	2424	
modale:	módbeli	Ag	0L		#
modalità:	mód	S	2L	3450	#
modella:	modell, manöken	S	1Q		
modello:	példa	S	0Q	411	
moderato:	közepes	Ag	5L	6505	

moderno:	modern	Ag	0L	1449	
modo:	mód	S	4L	127	
moglie, *la*:	feleség	S	3Q	511	
molto:	sok	Ag	0L	207	
molto:	nagyon	Avv	1L	73	
momento:	pillanat	S	2L		
monaco:	szerzetes	S	4L	6505	
mondiale:	világ-	Ag	5L	5253	
mondo:	világ	S	3L	319	
montagna:	hegy	S	2L	1367	
montatura:	szemüvegkeret	S	9QA		#
monumento:	műemlék, látványosság	S	3QA	4583	
morbido:	puha	Ag	9L	6505	
morire:	meghal	V	2LAP	320	
mortadella:	mortadella felvágott	S	8L	6505	#
mosso:	hullámos	Ag	5L		
mostra:	kiállítás	S	2L	2265	
mostrare:	megmutat	V	11QTR	1952	
mostro:	szörny	S	7L	4081	
motivare:	indokol	V	7L	4960	
motivo:	ok	S	1L	465	
moto, *la*:	motor	S	0LAP	2236	
motorino:	robogó	S	2L	4837	
movimento:	mozgás	S	4L	532	
mozzarella:	mozzarella sajt	S	4L	6505	#
mucca:	tehén	S	8L	2767	
mulino:	malom	S	8L		
multicolore:	sokszínű	Ag	9L		
muovere:	mozgat	V	4LAP	645	
muoversi:	mozog	V	9Q	645	
museo:	múzeum	S	0LA	2335	
musica:	zene	S	0L	652	
musicale:	zenei	Ag	6L	2669	
musicista:	zenész	S	11L		
napoletano:	nápolyi	Ag	1Q	1256	
nascere:	születik	V	4L	416	
nascita:	születés	S	3L	1190	
nascondere:	elbújtat, elrejt	V	4LAP	1403	
nascosto:	rejtett	Ag	1LAV	2604	
naso:	orr	S	1L	2648	
Natale, *il:*	karácsony	S	5L	1396	
natura:	természet	S	2L	970	
naturale:	természetes	Ag	6LA	976	
naturalmente:	természetesen	Avv	5L	346	
nave, *la*:	hajó	S	5L	2342	
nazionale, *la*:	válogatott	S	2L	3661	
nazionale:	nemzeti	Ag	5L	535	
nazione:	nemzet	S	10L	1844	
nazista, *il*:	náci	S	7L		#
ne:	arról, abból	P	2L	69	
né:	sem	C	7Q	407	
nebbia:	köd	S	5LA		
necessario:	szükséges	Ag	2Q	772	
necessità:	szükség	S	5L	898	
negativo:	tagadó	Ag	10L	998	
negoziante:	eladó	S	8L	4191	
negozio:	üzlet	S	2L	1170	
nemico:	ellenség	S	7L	4231	
nemmeno:	sem	Avv	5LA	2945	
neorealismo:	neorealizmus	S	7L		#
neppure:	sem	Avv	7LA	4002	
nero:	fekete	Ag	1L	603	
nervoso:	ideges	Ag	7L	2384	

nessuno:	semmilyen	Ag	1Q	369	
nessuno:	semmi, senki, egy se	P	3L	238	
neve, *la:*	hó	S	5L	2339	
nevicare:	havazik	V	4L	4555	
niente:	semmi	P	3L	248	
nipote, *il/la:*	unoka, unokaöccs, unokahúg	S	6L	2108	
nipotino:	kisunoka	S	6L		
no:	nem	Avv	0L	25	
nobilissimo:	nemes	Ag	8L	4639	
noi:	mi	P	0L	29	
nome, *il:*	név	S	0L	217	
non:	nem	Avv	1L	8	
nonna:	nagymama	S	0L	2040	
nonno:	nagypapa	S	6L	2058	
nonostante:	valami ellenére	C	7QTF	2479	
Nord, *il*:	észak	S	3Q	672	
normale:	normális	Ag	7Q	572	
nota:	megjegyzés; hangjegy	S	0L; 0L	1424	
notare:	észrevesz	V	0L	721	
notizia:	hír	S	1L	580	
notiziario:	hírműsor	S	10L		#
noto:	ismert	Ag	5L	2350	
notte, *la:*	éjszaka	S	0L	628	
novembre, *il:*	november	S	3L	2284	
novità:	újsonság	S	3Q	1065	
numerato:	számozott	Ag	0L	6505	
numero cardinale:	tőszám	SPr	2L		
numero ordinale:	sorszám	SPr	2L		
numero:	szám	S	0L	146	
numerosissimo:	számos	Ag	11L	2113	
nuoto:	úszás	S	11L	6505	
nuovamente:	újból	Avv	5L	3783	
nuovo:	új	Ag	0L	171	
nutritivo:	tápláló	Ag	6L		#
nutrizionale:	táplálkozási	Ag	8L		#
nuvola:	felhő	S	5LA		
nuvolosità:	felhősség	S	5LA		#
nuvoloso:	felhős	Ag	5L		
o:	vagy	C	0L	53	
obbligo:	kötelezettség	S	0LAP	1624	
occasione:	alkalom	S	6L	899	
occhiale, *l' (m.)*:	szemüveg	S	6Q	4859	#
occhiali da sole:	napszemüveg	SPr	7Q	x	
occhiali da vista:	látószemüveg	SPr	1L	x	
occhiata:	pillantás	S	11L	2430	
occhietto:	szemecske	S	8QTF		
occhio:	szem	S	1L	474	
occuparsi (di):	foglalkozik valamivel	V	9L	429	
occupato:	megszállt	Ag	7L	2943	
oceano:	óceán	S	7L	6505	
offendere:	megsért	V	4LAP	2282	
offendersi:	megsértődik	V	9Q	2282	
offerta:	akció, ajánlat	S	5L	1251	
offerto:	kínált	Ag	5L	6505	
offeso:	sértett	Ag	11L	6505	
offrire:	kínál	V	1L	745	
oggetto:	tárgy	S	0L	1162	
oggi:	ma	Avv	1L	123	
ogni:	minden, mindegyik	Ag	0L	162	
ognuno:	mindenki	P	4L	607	
olio:	olaj	S	6L	2279	
oliva:	olívabogyó	S	6L	3117	
oltre (a):	valamin túl	PzPr	5L	x	

ombelico:	köldök	S	11L		
ombrello:	esernyő	S	5LAV	3354	
onesto:	őszinte, tisztességes	Ag	11QTGF	3703	
opera:	opera; alkotás	S	0L; 7L	810	
opportunità:	lehetőség, alkalom	S	10L	1432	#
opportuno:	megfelelő	Ag	7Q	1375	
opposizione:	ellentét, szembeállítás	S	3L	2070	
oppure:	vagy	C	1L	197	
ora:	most	Avv	1L	128	
ora:	óra	S	1L	177	
orale:	szóbeli	Ag	4L	2576	#
oralmente:	szóban	Avv	1L	4172	#
orario:	nyitva tartás, időbeosztás	S	1L	746	
ordinare:	rendel	V	4L	3235	
ordinazione:	rendelés	S	6L	6144	#
ordine, *l'* (*m.*):	sorrend, rend	S	0L	375	
orecchino:	fülbevaló	S	8L	3841	
oretta:	szűk óra	S	3Q	177	
organizzare:	szervez, rendez	V	2L	466	
organizzazione:	szervezet	S	9L	460	
organo:	kiadvány	S	10L	1745	
orgoglioso:	büszke	Ag	11L	5419	
originale:	eredeti	Ag	7L	3506	
origine, *l'* (*f.*):	eredet	S	0L	1453	
orizzontale:	vízszintes	Ag	2LAV	3673	
ormai:	immár	Avv	6Q	243	
oro:	arany	S	8L	1541	
orologio:	óra	S	1Q	2412	
ortopedico:	ortopéd szakorvos	S	11L	5205	#
ospedale, *l'* (*m.*):	kórház	S	2L	996	
ospitale:	vendégszerető	Ag	4L	6505	#
ospitare:	vendégül lát	V	4L	2126	
ospite, *l' (m. / f.)*:	vendég	S	7L	3113	
osservare:	megfigyel	V	0L	1650	
osteria:	pub, vendéglő	S	6L	6505	
ottenere:	elér	V	6L	727	
ottimista:	optimista	S	9L	4054	
ottimo:	kitűnő	Ag	2L	1017	
ottobre, *l'* (*m.*):	október	S	2L	1871	
ovvero:	vagyis	C	10L	2791	
ovviamente:	nyilvánvalóan	Avv	2L	394	
pacchetto:	csomag	S	0L	4960	
pacco:	csomag	S	3L	1643	
padre, *il*:	apa	S	4Q	616	
paese, *il:*	ország	S	2L	358	
pagamento:	fizetés	S	4L	1080	
pagare:	fizet	V	0LA	228	
pagina:	oldal	S	0L	434	
paio:	pár	S	8L	726	#
palazzo:	épület	S	1L	900	
palcoscenico:	színpad	S	11L	3196	
palestra:	edzőterem	S	1L	6505	
palio:	lóverseny Sienában	S	5L	6505	
pancetta:	szalonna	S	6L		
pancia:	has	S	6L	2173	
pandoro:	veronai kuglóf	S	5L		#
pane, *il:*	kenyér	S	6L	1778	
panetteria:	pékség	S	8L		
panettone, *il*:	panettone, karácsonyi kuglóf	S	5L		#
panino:	szendvics	S	0L	2510	
paninoteca:	szendvicsbár	S	6L		#
panna:	tejszín, tejszínhab	S	4L	6144	
pannacotta:	tejszínpuding	S	4L		#

panorama, *il*:	panoráma, kilátás	S	0LAP	4288	
pantalone, *il*:	nadrág	S	9L	4269	#
papà, *il*:	papa	S	6L	1262	
parabolico:	parabola	Ag	10L		#
paradiso:	paradicsom	S	7L	6505	
parcheggiare:	parkol	V	3QTF	5205	
parcheggio:	parkoló	S	2L	3590	
parco:	park	S	5QTR	1670	
parecchio:	jócskán	Avv	9L	3508	
parecchio:	számos, sok	Ag	6L	1090	
parente, *il/la*:	rokon	S	5L	1768	
parentela:	rokonság	S	6L	6505	
parentesi, *la*:	zárójel	S	9L	1050	
parere, il:	vélemény	S	6LA	1150	
parete, *la:*	fal	S	3L	2514	
pari (a):	valaminek megfelelő	PzPr	4L		
parlare:	beszél	V	1L	62	
parlato:	beszélt	Ag	4L	6505	
parmigiano:	parmezán sajt	S	6L	6505	#
parodia:	paródia	S	7L	6505	#
parola:	szó	S	0L	236	
parrucchiere, *il*:	fodrász	S	10Q	6505	
parte, *la:*	rész	S	1L	81	
partecipare (a):	részt vesz valamiben	V	10L	662	
partenza:	indulás	S	5L	1814	
participio passato:	befejezett melléknévi igenév	SPr	4L	x	
particolare:	különleges, eltérő	Ag	0L	315	
particolarità:	különlegesség, eltérés	S	2L	5516	#
particolarmente:	különösen	Avv	5L	1246	
partire (per):	indul, elutazik	V	1L	216	
partita:	meccs, mérkőzés	S	1L	1068	
partitivo:	részelő	Ag	3L		#
partito:	pártok	S	10L	802	
Pasqua:	húsvét	S	5L	1588	
passante:	járókelő	S	11L		
passaporto:	útlevelek	S	6Q	2890	
passare:	eltölt (időt)	V	2L	151	
passatempo:	szabadidős tevékenységek	S	10L	6505	
passato prossimo:	közelmúlt	SPr	4L		
passato:	múlt	S	4L	1479	
passeggero:	utas	S	2L	6505	
passeggiare:	sétál	V	4L	6505	
passeggiata:	séta	S	4QTF	3545	
passione:	szenvedély	S	4L	2784	
passo:	lépés	S	11LAV	735	
pasta:	tészta	S	6LA	1560	
pasticceria:	cukrászda	S	8L	6505	
pasto:	étkezés	S	4L	2084	
patata:	burgonya	S	8L	1991	
patente, *la:*	jogosítvány	S	6Q	4472	
patria:	haza	S	10L	6505	
patrimonio:	vagyon, örökség	S	8L	3695	
paura:	félelem	S	5Q	546	
pausa:	szünet	S	6L	3140	
pazienza:	türelem	S	9L	909	
peccato:	kár	I	7L	1268	
pecorino:	juhsajt	S	8L	6505	#
peggio:	rosszabb	Avv	11L	1615	
pelle, *la:*	bőr	S	9L	2069	
pelliccia:	szőrme	S	9L	193	
pendente:	ferde	Ag	3QA		
penisola:	félsziget	S	5LA	6505	#
penna:	toll; tésztaféle	S	0LA; 6L	2131	

pensare:	gondol	V	2L	76	
pensiero:	gondolatok	S	9Q	1128	
pentola a pressione:	kuktafazék	SPr	6L	x	
pentola:	fazék	S	6L	5669	
penultimo:	utolsó előtti	Ag	0LAP	6505	#
pepe, *il:*	bors	S	6L	6505	
per caso:	véletlenül	AvvPr	4QTF	1	
per esempio:	például	AvvPr	1L	49	
per favore:	kérem, legyen szíves	AvvPr	2L	5	
per fortuna:	szerencsére	AvvPr	2L	3	
per forza:	még jó	AvvPr	8L	1	
per niente:	egyáltalán nem	AvvPr	6Q		
per piacere:	kérem	AvvPr	11L	3	
percentuale, *la:*	százalékos arány	S	3L	1672	#
perché:	miért	C	1L	21	
perciò:	ezért	C	3L	1031	
perdere:	elveszít	V	4L	255	
perdersi:	eltéved	V	9Q	255	
perdonare:	megbocsát	V	11Q	3023	
perfetto:	tükéletes	Ag	3L	834	
pericolo:	veszély	S	3L	1719	
pericoloso:	veszélyes	Ag	11QTR	2226	
periferia:	külváros	S	2L	3593	
perifrasi, *la*:	körülírás	S	0LAP		#
periodo ipotetico:	feltételes mondat	SPr	5L		
periodo:	időszak	S	5L	323	
permesso:	engedély	S	6L	1693	
permettere:	megenged	V	1L	563	
permettersi:	megengedmagának	V	9L	563	
però	de, azonban	C	3L	52	
però!:	nem mondom, ez igen!	I	5L		
persona:	személy	S	1L	113	
personaggio:	személyiség	S	7L	861	
personale, *il:*	személyzet	S	5L	1892	
personale:	személyes	Ag	1L	788	
pesante:	nehéz	Ag	1L	1334	
pescare:	horgászik	V	2L	3780	
pesce, *il:*	hal	S	0L	1949	
pescivendolo:	halbolt	S	8L		#
pessimista, *il/la*:	pesszimista	S	0LAP	4054	#
pesto:	bazsalikomos szósz	S	6L	6505	
pettinarsi:	fésülködik	V	9Q		
pezzo:	darab	S	0L	452	
piacere, *il*:	öröm	S	0L	513	
piacere:	tetszik	V	1Q	265	
piacevole:	kellemes, barátságos	Ag	4L	2536	
pian piano:	apránként	AvvPr	9L	11	
piangere:	sír	V	4L	1364	
pianista:	zongorista	S	7L		#
piano:	lassan	Avv	2L	734	
piano:	emelet	S	2L	348	
pianoforte, *il:*	zongora	S	11L	5419	
pianta:	növény	S	3L	1281	
pianura:	síkság, alföld	S	8L	5914	
piatto**:**	tányér; fogás	S	5Q; 6L	1213	
piazza:	tér	S	3LAV	518	
piccante:	csípős	Ag	6L	5867	
piccolo:	kicsik	Ag	0L	189	
picnic, *il*:	piknik	S	5Q	6505	#
piede, *il*:	láb	S	8QTF	637	
pieno (di):	tele	Ag	4L	902	
pillola:	tabletta, pirula	S	8L		
pioggia:	eső	S	0LA	2230	

piovere:	esik	V	5L	2901	
pirateria:	kalózkodás	S	11QTF		#
piscina:	medence, uszoda	S	9Q	3633	
pittore:	festő	S	6Q	3112	
più che altro:	leginkább	CPr	4L		
più o meno:	körülbelül, többé-kevésbé	AvvPr	4L	13	
più:	többé	Ag	1L	1161	
più:	több, többé	Avv	0L	43	
piuttosto:	inkább, meglehetősen	Avv	5L	621	
pizza:	pizza	S	0L	1791	
pizzaiolo:	pizzasütő	S	6L	6505	#
pizzeria:	pizzéria	S	6L	3426	
plurale, *il*:	többes szám	S	0L		
pneumatico, *lo:*	gumiabroncs	S	0LAP	6505	
pochino:	kevéske	Avv	10Q	60	
pochissimo:	nagyon kevés	Ag	4L	457	
poco:	kevéssé	Avv	1Q	60	
poco:	kevés	P	2L	1784	
poesia:	költemény	S	11L	1471	
poeta, *il*:	költő	S	7L	3193	
poi:	aztán	Avv	1L	37	
poiché:	mivel	C	11L	1960	
polemica:	vita	S	9L	1218	
politico:	politikai	Ag	8L	280	
polizia:	rendőrség	S	3L	1717	
poliziotto:	rendőr	S	10Q	6505	
pollice, il:	hüvelyk (mértékegység)	S	10L	6505	
pollo:	csirke	S	5L	3460	
poltrona:	fotel	S	3L	2933	
pomeriggio:	délután	S	1L	491	
pomodoro:	paradicsom	S	4L	3073	
ponte, *il:*	híd	S	2LAV	2032	
popolare:	népi	Ag	5L	1223	
popolo:	nép	S	6L	811	
porre:	helyez, tesz	V	2LAP	412	
porta:	ajtó	S	0Q	504	
portacenere, *il:*	hamutartó	S	8L		
portare:	hoz; visel	V	1L; 8L	102	
portata:	étel, fogás	S	6L	2184	
portineria:	portásfülke	S	11QTGF	6505	
positivo:	pozitív	Ag	10L	714	
posizione:	hely	S	4L	553	
possessivo:	birtokos	Ag	3L	6505	#
possesso:	birtoklás	S	6LAV	2838	
possibile:	lehetséges	Ag	1L	4326	
possibilità:	lehetőség	S	5L	347	
posta elettronica:	elektronikus posta	SPr	3L		
posta:	posta	S	3L	3338	
postino:	postás	S	7L		
post-laurea:	dimploma utáni, posztgraduális	Ag	11L		#
posto di lavoro:	munkahely	SPr	4L	x	
posto:	hely	S	4L	260	
potente:	hatalmas, erős	Ag	9L	5158	
potere:	képes, tud, -hat, -het	V	0L	31	
povero:	szegény	Ag	6L	591	
pranzare:	ebédel	V	3L		#
pranzo:	ebéd	S	2L	1516	
pratica:	gyakorlat	S	4L	596	
praticamente:	gyakorlatilag	Avv	7L	221	
precedente:	előző	Ag	0Q	754	
precisare:	pontosít	V	8L	2695	
preciso:	pontos	Ag	4L	639	
preferibile:	jobb, ajánlott	Ag	8L	6210	#

preferire:	előnyben részesít, szeret	V	0L	675	
preferito:	kedvenc	Ag	4L	6505	
prefisso:	előhívószám	S	3L	4728	#
prego:	kérem	I	1L	521	
premere:	megnyom	V	11L	4770	
premiare:	díjaz	V	11L	5357	
premio:	díj	S	7L	1731	
prendere il sole:	napozik	VPr	7Q		
prendere in giro:	ugrat	VPr	10L	6	
prendere parte (a):	részt vesz valamiben	VPr	7L	x	
prendere:	fog, vesz	V	1L	96	
prenotare:	előjegyez, lefoglal	V	5L	2142	
prenotazione:	helyfoglalás	S	3L	3223	
preoccuparsi:	aggódík	V	4Q	703	
preparare:	elkészít	V	4L	508	
prepararsi:	felkészül, elkészül	V	9L	508	
preparazione:	elkészítés	S	4L	1553	
preposizione articolata:	névelős elöljárószó	SPr	3L		
preposizione:	elöljárószó	S	2L		#
presentare:	bemutat	V	0L	247	
presentarsi:	bemutatkozik	V	9Q	247	
presentatrice:	bemondónő	S	10L	6505	#
presente, *il:*	jelen	S	1L	2143	
presente:	jelen lévő	Ag	3L	426	
presepe, *il*:	jászol, betlehem	S	5L	6505	#
presidente, *il:*	elnök	S	11L	523	
pressione:	nyomás	S	6L	1587	
presso:	-nál, -nél	Pz	5L	1521	
prestare:	nyújt, kölcsönöz	V	3Q	1638	
prestigioso:	rangos	Ag	11L	3361	#
prestito:	kölcsön	S	8Q	2671	
presto:	hamar, gyorsan	Avv	1L	678	
prevedere:	előrelát	V	5L	717	
previsione:	előrejelzések, jóslatok	S	5L	5050	
previsto:	előre látott, előre tervezett	Ag	11L	2416	
preziosissimo:	nagyon értékes	Ag	8L	5101	
prezzo fisso:	fix ár	SPr	9L		
prezzo:	ár	S	3L	387	
prima (di):	valami előtt	PzPr	2L	41	
prima:	előbb, korábban	Avv	1Q	97	
primato:	elsőbbség	S	8L	6402	
primavera:	tavasz	S	3L	4747	
primo:	első	Nu	3L	93	
primo:	első fogás	S	6LA		
principale:	főbb	Ag	5L	1293	
privato:	magán	Ag	2L	1367	
pro:	valami mellett szóló érv	S	10L	6322	#
probabilmente:	valószínűleg	Avv	3L	305	
problema, *il*:	probléma	S	0LAP	87	
procedura:	eljárás	S	8L	1520	
processo:	folyamat	S	8L	618	
prodotto:	termék	S	6L	647	
produrre:	termel, gyárt	V	2LAP	684	
produttore:	termelő, gyártó	S	9L	3027	
produzione:	gyártás, előállítás	S	8L	1105	
professore, *il*:	tanár; professzor	S	0Q; 2L	378	
professoressa:	tanárnő, professzor asszony	S	3L	1063	
profumo:	parfüm; illat	S	2Q; 4L	4194	
progetto:	terv	S	5L	677	
programma, *il*:	program	S	0LAP	542	
promessa:	ígéret	S	5L	4527	
promettere:	megígér	V	4L	2871	
pronome diretto:	tárgyas névmás	SPr	8L		

pronome partitivo:	részelő névmás	SPr	8L		
pronome, *il:*	névmás	S	1L		#
pronto!:	haló	I	1L	395	
pronto:	gyors, azonnali	Ag	1L	724	
pronuncia:	kiejtés	S	0L	6505	
pronunciare:	kiejt	V	0L	4172	
proporre:	javasol	V	2LAP	463	
proposta:	javaslat, bemutatkozó	S	11L	480	
proposto:	ajánlott, javasolt	Ag	0L	6505	
proprietà:	tulajdon	S	10L	1228	
proprietario:	tulajdonos	S	8L	1097	
proprio:	éppen, pontosan	Avv	4L	80	
proprio:	saját	P	3L	4350	
prosciutto:	sonka	S	4L		
prossimo:	következő	Ag	2L	3276	
protagonista, *il/la*:	főszereplő	S	1L	2124	
protetto:	védett	Ag	8L	6505	
protezione civile:	polgári védelem	SPr	3L		
protezione:	védelem	S	3L	3306	
provare:	megpróbál	V	3L	313	
provenire:	származik, ered	V	7L	2918	
proverbio:	közmondás	S	5L	6505	
provincia:	körzet	S	3L	685	
provocatorio:	provokatív	Ag	9L	3826	#
provvedere:	gondoskodik valamiről	V	5L	1922	
psicologico:	pszichológiai	Ag	7L	1612	
psicologo:	pszichológus	S	0LAP	4671	
pubblicità:	reklám	S	7L	1310	
pubblicitario:	reklám-	Ag	10L	2539	
pubblicizzato:	reklámozott	Ag	7L		#
pugno:	marok, ököl	S	7L		
pulire:	takarít, tisztít	V	1L	1744	
pulito:	tiszta	Ag	2L	3024	
pullman, *il*:	turistabusz	S	5LAV	2521	
pullover, *il:*	pulóver	S	9L		
pulsante, *il*:	gomb	S	11L	6505	#
puntata:	rész (folytatásos tévéfilmben)	S	10L	6505	
punto di incontro:	találkozóhely	SPr	4L		
punto di ritrovo:	találkozóhely	SPr	4L		
punto:	feladat	S	2L	107	
puntuale:	pontos	Ag	5LAV	6069	
può darsi:	lehet	VPr	4LAV		
pure:	is; csak nyugodtan	Avv	5Q; 10L	148	
pure:	is	C	11L	1058	
purtroppo:	sajnos	Avv	1L	295	
qua:	itt	Avv	3L	101	
quaderno:	füzet	S	0Q	2223	
quadro:	kép	S	3L	756	
qualche:	néhány	Ag	2L	114	
qualcosa:	valami	P	2L	185	
qualcuno:	valaki	P	1L	237	
quale:	mely	Ag	0L	388	
quale:	melyik	P	0L	129	
qualità:	minőség	S	5L	897	
qualsiasi:	bármelyik	Ag	6L	641	
quando:	mikor	Avv	1L		
quando:	amikor	C	2L	59	
quantità:	mennyiség	S	5L	1280	
quanto:	hány, mennyi	Ag	0L	293	
quanto:	hány	P	5L	137	
quasi:	majdnem	Avv	1Q	259	
quello:	az	Ag	6L	64	
quello:	az	P	1L	33	

questo:	ez	Ag	5L	20	
questo:	ez	P	0L	30	
qui:	itt	Avv	1L	61	
quindi:	tehát	C	2L	3202	
quinto:	ötödik	Nu	2L	683	
quiz, il:	kvíz	S	3QA	2691	#
quotidiano:	mindennapi	Ag	2L	1488	
quotidiano:	napilap	S	10L	3361	
rabbia:	düh	S	8L	6505	
raccogliere:	gyűjt	V	2LAP	828	
raccolta:	gyűjtemény, válogatás	S	6L	1758	
raccomandarsi:	ajánl	V	11L	954	
raccontare**:**	elmesél	V	2L	439	
racconto:	elbeszélés	S	7L	1621	
raddoppiare:	megkettőz	V	10L		
radio, *la:*	rádió	S	4L	556	
radiotelevisione:	rádió és televízió	S	10L		#
raffinatezza:	kifinomultság, választékosság	S	9L		#
raffinato:	kifinomult	Ag	5L	6505	
ragazza:	lány	S	0L	499	
ragazzo:	fiú	S	0L	239	
raggiungibile:	elérhető	Ag	3L		#
ragione, *la:*	értelem	S	1L	318	
ragionevole:	kedvező	Ag	5QA	6069	
ragù:	ragu	S	6L		
rammarico:	sajnálkozás	S	8L		#
rapido:	gyorsvonat	S	5L		
rapido:	gyors	Ag	5L	1747	
rapporto:	kapcsolat	S	6L	287	
rappresentare:	képez	V	7L	789	
raramente:	ritkán	Avv	11L	6505	
rata	részlet (összeg)	S	8L	3768	
re, *il*:	király	S	0LAP	1656	
reagire:	reagál	V	11L	3284	
reale:	valószínűleg	Ag	3L	692	
reality:	valóság	S	10L		#
realizzabile:	megvalósítható	Ag	11L		#
realizzare:	megalkot	V	7L	958	
realtà:		S	7L	365	
reciproco:	kölcsönös	Ag	9L	4868	
recitare:	játszik (szerepet)	V	7L	2572	
reclamizzato:	reklámozott	Ag	8L		#
record:	rekord	S	7L	4395	
redazione:	szerkesztőség	S	10L	5867	
regalare:	ajándékoz	V	7L	581	
regalo:	ajándék	S	0L	1518	
regata:	evezős verseny	S	5L		#
regia:	rendezés	S	7L	2164	#
regina:	királynő	S	6L	5516	
regionale:	helyi, adott tartományra jellemző	Ag	5L	1047	
regione, *la*:	tartomány	S	1L	604	
regista, *il*/*la*:	rendező	S	0LAP	2167	
registrazione:	hangfelvétel	S	8L	5419	
regola:	szabály	S	0L	1013	
regolare:	szabályos	Ag	2L	1876	
regolarmente:	rendszeresen	Avv	11L	2885	
relativamente:	viszonylag	Avv	9L	2494	
relativo (a):	valamire vonatkozó	PzPr	3L		
relazione:	kapcsolat	S	4L	711	
religioso:	vallási	Ag	4L	2602	
rendere:	valamilyenné tesz	V	6L	333	
reparto:	részleg	S	8QA	2027	
repubblica:	köztársaság	S	4L	916	

responsabile:	felelős	S	11QTF	1659	
restante:	maradék	Ag	4L		
restare:	marad	V	2L	506	
resto:	a többi, a maradék	S	3L	594	
rete, *la:*	hálózat, háló; tv-csatorna	S	5L; 10L	1800	
riassumere:	összefoglal	V	3L	2587	
riassunto:	összefoglalás	S	4L	3102	
ricapitolazione:	ismétlés	S	2QTR		#
ricco:	gazdag	Ag	6L	1243	
ricetta:	recept	S	6L	2187	
ricevere:	megkap; fog egy programot	V	3L; 10L	983	
ricevuta:	számla	S	4L	1549	
richiedere:	igényel, igénybe vesz	V	3L	919	
richiesto:	kért	Ag	6L		
riconoscere:	felismer	V	6L	446	
riconoscimento:	elismerés	S	8L	1137	
riconosciuto:	elismert	Ag	11L		
ricordare:	emlékszik	V	1LAV	116	
ricordarsi (di):	emlékszik	V	5L	116	
ricordo:	emlék	S	7L	1954	
ricostruire:	átalakít, újraalkot	V	10L	2102	
ricotta:	túró	S	8L		
ridare:	visszaad	V	10Q	1783	
ridere:	nevet	V	1Q	849	
ridurre:	csökkent	V	2LAP	1133	
rientrare:	újra belép, bemegy	V	4L	1165	
riferire:	beszámol	V	0L	710	
riferirsi:	utal valamire	V	4L	710	
rifiutare:	visszautasít	V	2L	1951	
riflessivo:	visszaható	Ag	4L		#
riflettere:	gondolkodik	V	8L	1567	
riga:	sor; csík	S	7L; 9L	1260	
rigatone, *il*:	olasz tésztaféle	S	6L		#
riguardare:	vonatkozik valamire	V	10QA	222	
rimanere:	marad	V	2LAP	180	
Rinascimento:	reneszánsz	S	6L	6505	#
ringraziamento:	köszönet	S	3L	4288	
ringraziare:	megköszön	V	3L	292	
rinnovamento:	újítás	S	6L	3806	#
rinomato:	híres	Ag	8L		
rinunciare:	lemond	V	5L	1781	
ripartire:	elindul vissza	V	5L	3914	
ripensare:	visszagondol	V	7L	2633	
ripetere:	megismétel	V	0L	379	
ripetuto:	ismétlődő	Ag	7L		
ripieno:	töltött	Ag	5L	6505	
riposarsi:	pihen	V	10Q	3007	
riposato:	kipihent	Ag	9Q	6505	
ripostiglio:	tárolóhelyiség	S	2L		#
riprendere:	újrakezd	V	7Q	660	
riquadro**:**	táblázat, négyzetháló	S	9LAV	6505	#
risata:	nevetés	S	7L	4468	
risolvere:	megold	V	1QTF	507	
risotto:	rizottó	S	6L	5101	#
rispettivo:	odatartozó	Ag	9L	6505	#
rispetto (a):	valamihez képest	PzPr	8L	82	
rispondere:	válaszol	V	0L	335	
risposta:	válasz	S	0L	476	
ristorante, *il:*	étterem	S	0L	1759	
ristorazione:	étkeztetés	S	5L		#
ristretto:	nagyon erős kávé	S	4L		
risultato:	eredmény	S	7L	536	
ritardo:	késés	S	3L	1511	

ritenere:	tart valamit valamilyennek	V	2LAP	530	
ritirare:	átvesz	V	4L	853	
ritiro:	átvétel	S	5L	4114	
ritmico:	ritmusos	Ag	11L		#
rito:	rítus	S	4L	4741	
ritornare:	visszatér	V	4L	398	
ritorno:	visszatérés	S	5L	1122	
ritrovare:	talál	V	8L	1037	
ritrovarsi:	találkozik	V	9Q	1037	
ritrovo:	találkozóhely	S	4L	6505	
riuscire (a):	sikerül	V	3L	155	
rivedersi:	újra találkozik	V	9L	833	
rivelare:	felfed, elárul	V	9L	3004	
rivestito:	bevont	Ag	9L		
rivista:	magazin	S	2L	1533	
rivivere:	újraéled	V	3Q	6505	#
rivolgersi:	fordul valakihez	V	9L	1036	
romano:	római	Ag	6L	921	
romantico:	romantikus	Ag	5Q	5552	
romanzo:	regény	S	11L	926	
rompere:	tör	V	4LAP	720	
rosa:	rózsaszín	Ag	9L	3152	
rosa:	rózsa	S	6Q	4747	
roseo:	rózsaszínű	Ag	8L	6505	
rosolare:	pirít	V	6L	6505	
rossastro:	vöröses	Ag	6L		
rosso:	vörös	Ag	0L	573	
rotondo:	kerek	Ag	6L	1732	
rovina:	rom	S	11QTGF	5669	
rumore, *il*:	zaj	S	4L	2378	
rumoroso:	zajos	Ag	9Q		
ruolo:	szerep	S	1L	806	
ruota:	kerék	S	10L	3826	
russo:	orosz	Ag	0QTF	5205	
sabato:	szombat	S	2L	681	
sacchetto:	zacskó, szatyor	S	8Q	6505	
sacco:	zsákok	S	4L	771	
sacro:	szent	Ag	4L	2407	
sala:	terem	S	4L	1186	
salame, *il:*	szalámi	S	6L		
salato:	sós	Ag	6L	2506	
salatura:	sózás	S	8L		#
saldo:	árleszállítás	S	9L	6505	
sale, *il:*	só	S	6L	1551	
salire:	felszáll	V	2L	1009	
salita:	felmenetel	S	5L		
salmone, *il:*	lazac	S	6L	6505	
salone, *il*:	szalon	S	3Q	3706	
salotto:	nappali (nagy, elegáns)	S	2L	1705	
saltare:	kihagy	V	6L	1191	
salutare:	üdvözöl	V	1L	653	
saluto:	üdvözlés, köszönés	S	1L	982	
salve:	üdv	I	1L	2281	
sanitario:	egészségügyi	Ag	3L	2204	
sano:	egészséges	Ag	10L	1634	
sapere	tud	V	0L	47	
sapore, *il:*	íz	S	0LAP	3228	
saporito:	ízletes	Ag	6L		#
saracino:	szaracén	Ag	5L		#
satellitare:	műholdas	Ag	10L		#
sbagliare:	elhibáz, eltéveszt	V	2L	493	
sbagliato:	téves, hibás	Ag	8LAV	936	
sbattere:	felver	V	6L	2455	

sbrigarsi:	igyekszik, siet	V	9L	6505	
scadere:	lejár	V	8QA	3411	
scaloppina:	borjúszelet	S	6L	6505	#
scambiarsi:	cserél, vált	V	4L	2161	
scandalo:	botrány	S	11L	4395	
scaricare:	letölt	V	11QTF	2054	
scarpa:	cipő	S	8L	2072	
scatoletta:	konzerv	S	2LAP	1181	
scegliere:	választ	V	0QTF	392	
scelta:	választás	S	0LA	503	
scena:	jelenet	S	4L	1053	
scendere:	lemegy, leszáll	V	0L	747	
scheda:	kártya	S	3L	1336	
schema, *lo:*	vázlat, séma	S	0L	1373	#
schermo:	vetítővászon, képernyő	S	7L	2782	
scherzare:	viccel, viccelődik	V	4L	1429	
scherzo:	tréfa	S	5L	4207	
schiuma:	hab	S	4L		
sciare:	síel	V	5QA	6322	
sciarpa:	sál	S	9L	6505	
sciopero:	sztrájk	S	3L	1592	
scolare:	leszűr	V	6L		
scomparire	elhuny, eltűnik	V	7L	1786	
scompartimento:	vonatfülke	S	5L		
sconto:	kedvezmény	S	4LAV	1550	
scontrino**:**	blokk	S	4L		
scooter, *lo*:	robogó	S	3L	6505	#
scoperta:	felfedezés	S	3L	2615	
scoppiare:	kitör, kirobban	V	7LAV	2481	
scoprire	megtalál, felfedez	V	0L	868	
scorso:	múlt	Ag	4L	454	
scortese:	udvariatlan	Ag	1L	6505	
scritta:	felirat	S	4L	3919	
scritto:	írt	Ag	4L	1660	
scrittore:	író	S	4Q	2371	
scrivania:	íróasztal	S	3L	6505	
scrivere:	leír, ír	V	0L	138	
scuola:	iskola	S	0L	235	
scusare:	megbocsát	V	0L	140	
scusarsi:	elnézést kér	V	10Q	140	
se:	ha; -e	C	1L; 4L	32	
sé:	maga	P	7L		
secco:	száraz	Ag	6L	1817	
secolo:	század	S	6L	1597	
secondo:	második	Nu	1L	213	
secondo:	szerint	Pz	1L	172	
secondo:	második fogás	S	6LA	1569	
sede:	székhely	S	10L	719	
sedere:	leül	V	2LAP	1332	
sedersi:	leül	V	4L	1332	
sedia:	szék	S	3L	2682	
segnare:	beír, bejelöl	V	2L	815	
segretaria:	titkárnő	S	7Q	3816	
segreteria telefonica:	üzenetrögzítő	SPr	5Q	4	
segreto:	titkos	Ag	7L	2903	
segreto:	titok	S	7L	3643	
seguente:	következő	Ag	0Q	2213	
seguire:	követ	V	0L	355	
seguito:	folytatás	S	11L	957	
Self Service:	automata	SPr	5L	x	
sembrare:	tűnik	V	1L	150	
semplice:	egyszerű	Ag	3L	564	
semplicemente:	egyszerűen	Avv	4L	758	

sempre:	mindig	Avv	2L	70	
senso:	értelem	S	11L	165	
sentimentale:	szentimentális, érzelmes	Ag	11QA	4160	
sentire:	hall	V	1L	74	
sentirsi**:**	érzi magát valahogy	V	9L	74	
senza:	nélkül	Pz	1L	144	
separarsi:	különköltözik	V	11Q	1998	
sera:	este	S	2L	276	
serata:	este	S	1L	1356	
serenata:	szerenád	S	11L	6505	
serenità:	derű	S	8QTF	5552	
sereno:	derült	Ag	5L	2195	
seriamente:	komolyan	Avv	9L	2198	
serie, *la*:	sor; sorozat	S	0LAP; 2L	403	
servire (a):	kell, szolgál valamire	V	2LA	250	
servizio:	szolgáltatás; készlet	S	2L;10L	396	
seta:	selyem	S	9L	4591	
sete, *la:*	szomúság	S	4L	5914	
settembre, *il:*	szeptember	S	3L	2019	
settimana bianca:	síszünet	SPr	5L	2	
settimana:	hét	S	0L	289	
settimanale, *il:*	hetilap	S	10L	6505	
settore:	szektor	S	8L	534	
sezione:	kategória	S	4L	1641	
sfilata:	divatbemutató	S	9L	6505	
sfortunato:	balszerencsés	Ag	11L	2932	
sfruttare:	kihasznál	V	10L	956	
shampoo:	sampon	S	8QA		
show, *lo*:	show	S	10L	5328	#
showman:	showman	S	0L		#
sì e no:	körülbelül	AvvPr	2L	2	
sì:	igen	Avv	0L	24	
sia:	mind - mind, is - is	C	4L	317	
siccome:	mivel	C	6L	266	
siciliano:	szicíliai	Ag	5Q	4011	
sicuramente:	bizonyosan	Avv	5Q	296	
sicuro:	biztos	Ag	3L	658	
sigla:	rövidítés, betűszó	S	3L	3611	
significare:	jelent valamit	V	5Q	330	
significato:	jelentés	S	3L	1159	
signora:	hölgy, asszony	S	1LA	188	
signore, il:	úr	S	1LA	192	
signorina:	kisasszony	S	1L	688	
silenzio**:**	csend	S	10Q	1806	
sillaba:	szótag	S	0LAP		
simbolo:	jelkép	S	4L	1976	
simile (a):	hasonló	Ag	1L	1328	
simpatia:	rokonszenv, szimpátia	S	8Q	6505	
simpatico:	rokonszenves	Ag	1L	1086	
sincero:	őszinte	Ag	10L	2633	
singolare, *il*:	egyes szám	S	0L		
singolo:	egyes	Ag	11QTGF		
sinistra:	baloldal	S	2L	467	
sinistro:	bal	Ag	2L	5050	
sinonimo:	szinonoma	S	9L	6505	#
sintesi, *la*:	szintézis, összefoglalás	S	0LAP	1964	#
sistema, *il*:	rendszer	S	8L	419	
sito:	honlap	S	5L	6505	
situato**:**	elhelyezve	Ag	4L	6505	
situazione:	helyzetek, szituációk	S	1L	200	
slogan, *lo*:	szlogen	S	10L	6505	#
smettere (di):	abbahagy	V	5L	903	

soap opera:	szappanopera	SPr	10L	2	
soccorso:	segély	S	3L	6505	
sociale:	társadalmi	Ag	10L	592	
società:	társadalom; társaság	S	7QTF; 9L	391	
soddisfare:	kielégít	V	10L	4837	
soddisfatto:	elégedett	Ag	1LAV	3000	
soffrire:	szenved	V	4L	1315	
soggetto:	alany	S	9L	740	
soggiorno:	nappali; tartózkodás	S	2L; 8L	5253	
sognare:	álmodik	V	3L	3657	
sogno:	álom	S	10L	2263	
solamente:	csupán	Avv	10L	627	
soldo:	pénz	S	3Q	285	
sole, *il*:	nap	S	3Q	1140	
solito:	szokásos	Ag	4L	552	
solitudine, *la*:	magány	S	11L	3652	
solo:	csak	Avv	1L	84	
soltanto:	csak	Avv	5QTR	190	
soluzione:	megoldás	S	1LAV	760	
somiglianza:	hasonlóság	S	3L	6505	
somma:	összeg	S	7L	1570	
sopra:	fölött	Avv	3L	593	
soprattutto:	elsősorban, főleg	Avv	4L	279	
sorella:	lánytestvér	S	0L	1108	
sorellina:	húg	S	6L	1108	
sorpresa:	meglepetés	S	5L	2330	
sorpreso:	meglepett	Ag	6L	6505	
sorriso:	mosoly	S	10QA	5101	
sorta:	fajta, féle	S	6L	1440	
sospettare:	gyanúsít	V	4L	6505	
sosta:	megállás, tartózkodás	S	4L	5669	
sostantivo:	főnév	S	0L	6505	#
sostituire:	helyettesít	V	1L	1049	
sottile:	vékony	Ag	6L	2302	
sotto l'aspetto:	valamilyen szempontból	PzPr	8L		
sotto:	alatt	Pz	3L	352	
sottocultura:	szubkultúra	S	10L		#
sottolineare:	aláhúz	V	1L	1178	
sottostante:	alábbi	Ag	8L		#
souvenir, *il*:	szuvenir	S	5Q		#
spaghetti western:	olasz vadnyugati film	SPr	7L		
spaghetto, *lo:*	spagetti	S	0L	6505	
spagnolo:	spanyol	Ag	0L	5552	
sparire:	eltűnik	V	7L	1309	
spaventoso:	ijesztő	Ag	8L	3670	
spazio:	tér, hely	S	4LAV	562	
specchio:	tükör	S	3L	4089	
speciale:	különleges	Ag	1L	1232	
specialità:	specialitás, különlegesség	S	5L	4541	#
specializzato:	specializálódott	Ag	5QA	1977	
specialmente:	különösen	Avv	6Q	1151	
specie, *la*:	faj	S	0LAP	894	
specificare:	meghatároz, pontosít	V	9L	2969	
spedire:	küld	V	1L	1809	
spegnere:	leolt, kikapcsol	V	2LAP	1841	
spendere:	pénzt költ	V	4L	676	
sperare:	remél	V	2L	298	
spericolato:	vakmerő	Ag	11L		#
spesa:	bevásárlás	S	2L	816	
spesso:	gyakran	Avv	1Q	461	
spessore, *lo*:	vastagság	S	6L	4892	
spettacolare:	látványos	Ag	11L	6505	#

spettacolo:	előadás	S	4Q	2178	
spettatore, *lo*:	néző	S	11L	4054	#
spiacere:	sajnál	V	8L	3354	
spiegare:	kifejt, elmagyaráz	V	1L	300	
spiegazione:	magyarázat	S	6L	1562	
spingere:	lök, tol	V	4LAP	1691	
spiritoso:	vicces, szellemes	Ag	10L		
spogliarsi:	vetkőzik	V	9L	3081	
sponsor, *lo*:	szponzor, támogató	S	10L		#
sport, *lo:*	sport	S	0L	1754	
sportello:	ablak, kassza	S	5L	4951	
sportivo:	sportos	Ag	2L	2374	
sposare:	feleségül vesz valakit, férjhez megy valakihez	V	5LAV	1070	
sposarsi:	összeházasodik	V	7L	1070	
spostare:	arréb besz	V	10L	679	
spot, *lo*:	reklámfilm	S	10L	2418	#
spremuta:	dzsúsz	S	4L	6505	
spumante, *lo*:	pezsgő	S	5L		
spuntino:	könnyű, gyors étkezés	S	6L		
squillare:	csörög (telefon)	V	7L	3328	
stabile:	állandó	Ag	5L	3900	#
stadio, lo:	stadion	S	2L	1700	
stagionatura:	érlelés	S	8L	6505	#
stagione:	évszak	S	3L	3043	
stamattina:	ma reggel	Avv	4L	736	
stampa:	sajtó	S	7L	751	
stampare:	nyomtat	V	5L	1381	
stancarsi:	elfárad	V	9Q	2721	
stanco:	fáradt	Ag	2L	2444	
standard:	standard, szint	S	5L	2915	#
stanza:	szoba	S	2L	2004	
stare a dieta:	fogyókúrázik	VPr	11L		
stare attento (a):	odafigyel	VPr	8L		
stare insieme:	jár valakivel	VPr	7L		
stare:	van, áll	V	1L	48	
stasera:	ma este	Avv	2L	543	
statale:	állami	Ag	10L	4387	
statistica:	statisztika	S	10L	3376	#
statua:	szobor	S	4Q	3885	
stazione:	állomás	S	2L	1114	
stecchino:	fogpiszkáló	S	7L		#
stella:	csillag	S	0LA	2579	
stereo:	sztereómagnó	S	10L	4024	#
stesso:	ugyanaz	Ag	0L	136	
stesso:	ugyanaz	P	0L	666	
stile, *lo*:	stílus	S	9L	2319	
stilista:	divattervező	S	9L	6505	#
stimare:	nagyra becsül, tisztel	V	7L	6505	
stipendio:	fizetés, kereset	S	1L	1071	
stivale, *lo*:	csizmák	S	0L	6505	
stoffa:	anyag	S	9L	4054	
stomaco:	gyomor	S	0LAP	6505	
storia dell'arte:	művészettörténet	SPr	2QTR		
storia:	történet	S	1L	191	
storico:	történelmi	Ag	5L	617	
storiella:	kis történet	S	7L		
stracciatella:	sztracsatella	S	4L		#
strada:	út, utca	S	0L	492	
stradale:	utcai	Ag	10LAV	4825	
straniero:	külföldi	Ag	0L	1272	
strano:	furcsa, különös	Ag	6L	528	
straordinario:	kivételes	Ag	7L	1209	

strappare:	kitép	V	10L	2823	
stressante:	idegesítő	Ag	9QTF	6505	#
stressato:	feszült	Ag	8L		#
stretto:	szoros	Ag	9L	1343	
strumento:	hangszer	S	11LAV	769	
studente, *lo:*	diák, tanuló	S	0L	706	
studentessa:	tanuló, egyetemi hallgató lány	S	0Q		
studiare:	tanul	V	1L	354	
studio, *lo*:	dolgozószoba	S	2L	342	
stupendo:	csodálatos	Ag	6L	1692	
stupido:	hülye, buta	Ag	10L	3259	
su:	-ra, -re; -on, -en, -ön; -ról, -ről	Pz	2L	46	
subito:	rögtön, azonnal	Avv	2QTR	253	
succedere:	történik	V	4L	240	
successivo:	következő	Ag	4L	1501	
successo**:**	siker	S	6Q	1742	
succo:	gyümölcslé	S	8L	5799	
Sud, *il:*	dél	S	3L	691	
sufficiente:	elegendő	Ag	5L	988	
suggerimento:	javaslat, ötlet	S	9L	2615	
suggerire:	javasol, sugalmaz	V	10QA	2111	
sugo:	szósz	S	6L	5177	
sul serio:	komolyan	AvvPr	8L	2	
suo:	övé	Ag	0L	82	
suocero:	após	S	7L	5732	
suonare:	hangszeren játszik; megszólal	V	4Q; 7L	1126	
superare:	átmegy, leteszi a vizsgát	V	2L	690	
supermercato:	szupermarket	S	1LA	3711	#
supplemento:	pótjegy, helyjegy	S	5L		#
suscitare:	kelt	V	9L	2981	
sveglia:	ébresztőóra	S	11L	5732	
svegliarsi:	felébred	V	4L	2025	
sviluppato:	fejlett	Ag	9L	3274	
svizzero:	svájci	Ag	1L	3736	
svolgere:	zajlik, történik	V	5L	931	
svolgersi:	zajlott	V	8L	931	
tabaccheria:	trafik	S	2L		#
tabacco:	dohány	S	2L		
tabella:	táblázat	S	0L	2028	
tacchino:	pulyka	S	5L		
tacco:	sarok	S	9L	6505	
taglia:	méret	S	9L	4603	
tagliare**:**	felvág, vág	V	6L	804	
tagliatella:	olasz tésztaféle	S	6LA		
tagliere, *il:*	vágódeszka	S	6L		#
tailleur, *il:*	kosztüm	S	9L	6505	#
talento:	tehetség	S	7L	6505	
talk show:	talk show	SPr	8Q		
tango:	tangó	S	7L		#
tant'è vero che:	való igaz	CPr	7L		
tantissimo:	nagyon sok	Ag	4L	162	
tanto:	sok	Ag	1L	162	
tanto:	nagyon, annyira	Avv	1L	173	
tappeto:	szönyeg	S	3L	2120	
tardare:	késik	V	7Q		
tardi:	későn	Avv	1Q	1294	
tardo:	késő	Ag	10L	6505	
tariffa:	ár, tarifa	S	5L	6069	
tassì, *il:*	taxi	S	2L	4011	#
tassista, *il/la*:	taxis, taxisofőr	S	0LAP		#
tastiera:	billentyűzet	S	11L		
tatuaggio:	tetoválás	S	3Q		#
tavola:	asztal	S	5L	2048	

tavolino:	kis asztal, dohányzóasztal	S	3L	3641	
tavolo:	asztal	S	3L	611	
taxi, *il*:	taxi	S	2L	4011	
tazza:	csésze	S	4L	2292	
tazzina:	kis csésze	S	4L	2292	
tè, *il*:	tea	S	1QTF	6505	
teatrale:	színházi	Ag	8L	5023	
teatro:	színház	S	2L	605	
tecnico:	műszaki szakember	S	11QTGF	2051	
tecnologia*:*	technológia	S	3L	5357	#
tedesco**:**	német	Ag	1Q	1733	
tegame, *il:*	lábos	S	6L	6505	
telecomando:	távirányító	S	10L	6144	#
telecomunicazione:	távközlés	S	3L	6505	#
telefilm:	tévéfilm	S	1L	6505	
telefonare:	telefonál	V	3L	309	
telefonata:	telefonhívás	S	9L	701	
telefonia:	telefonok, telefonhálózat	S	3L		#
telefonico:	telefon-, telefonos	Ag	3L	1760	
telefonino:	mobiltelefon	S	3L	4700	
telefono:	telefon	S	1Q	339	
telegiornale, *il:*	tévéhíradó	S	8Q	4258	#
telegramma, *il*:	távirat	S	10L	3297	
telespettatore, *il*:	tévénéző	S	10L	5088	#
televendita:	televíziós marketing	S	10L		#
televisione, *la*:	tévé	S	1L	509	
televisivo:	tévés	Ag	7L	2344	#
televisore, *il*:	tévékészülék	S	3L	2345	
tema, *il*:	téma	S	0LAP	414	
tematica:	téma	S	10L	4172	#
temperatura:	hőmérséklet	S	3Q	2202	
tempo fa:	jó ideje	AvvPr	2L		
tempo:	idő	S	1L	111	
temporale, il:	zivatar	S	5L	6505	
tenere compagnia**:**	társaságot nyújt	VPr	2LAP		
tenere:	tart	V	2LAP	145	
tenero:	lágy, puha	Ag	8L	4762	
tenersi:	zajlik	V	11L	145	
tenore:	tenor	S	11L	6505	
tentativo:	kísérlet, próbálkozás	S	11L	1422	
teoria:	elképzelés, teória	S	6L	1206	
terme, *le*:	fürdő	S	11Q	6505	#
terminare:	végződik	V	0L	2440	
termine, *il*:	szó, kifejezés	S	3L	424	
terra:	föld	S	0L	402	
terrestre:	földi	Ag	10L	6505	
territorio:	terület	S	5L	1339	
terzo mondo:	harmadik világ	SPr	11L	14	
terzo:	harmadik	Nu	0LAP	360	
terzultimo:	hátulról a harmadik	Ag	0LAP		#
tesi, *la*:	dolgozat, esszé	S	0L	842	
tesoro*:*	drágám	S	0L	1366	
tessuto:	szövet	S	9L	3159	
test finale:	záróteszt	SPr	0L		
test, *il:*	teszt	S	0Q	2895	#
testa:	fej	S	0L	709	
testata:	újság	S	10L	4664	#
testimone, *il*:	tanú	S	11L	2640	
testo**:**	szöveg	S	1L	512	
thriller, *il*:	thriller	S	10L		#
tigre, *la*:	tiigris	S	7L	6505	
timbrare:	bélyegez	V	2L		#
timido:	félénk	Ag	7L	3922	

tipico:	tipikus	Ag	4L	1804	
tipo:	fajta, típus	S	1L	147	
tiramisù:	tiramisu	S	4L		#
tirare:	fúj (a szél)	V	5L	444	
toccare:	megérint	V	5Q	587	
togliere:	levesz, lehúz	V	2LAP	477	
tonnellata:	tonna	S	4L	6505	
tonno:	tonhal	S	4Q	5732	
tornare:	visszamegy, visszatér	V	1L	254	
torre, la:	torony	S	0L	4317	
torta:	torta	S	4L	2528	
tortellino:	töltött tésztaféle	S	6L		#
toscano**:**	toszkán	S	5QTF	2615	
totocalcio:	totó	S	6Q	6505	#
tournée, *la*:	turné	S	2L		#
tovaglia:	terítő	S	6L	3770	
tovagliolo:	szalvéta	S	6L	4269	
tra:	között	Pz	3L	121	
tradire:	megcsal	V	10L	5623	
tradizionale:	hagyományos	Ag	5L	1735	
tradizione:	hagyomány	S	8L	1427	
tradurre:	fordít	V	2LAP	1312	
traduzione:	fordítás	S	8LA	1383	
traffico:	forgalom, közlekedés	S	2L	1937	
tragedia:	tragédia	S	7LA	2673	
tram, *il*:	villamos	S	2L	6505	
trama:	cselekmény	S	7L	5357	
tramezzino:	háromszög alakú szendvics	S	4L		#
tramite:	keresztül	Pz	9L	1111	#
tranquillità:	nyugalom	S	8QTF	2021	
tranquillo:	nyugodt	Ag	4L	823	
transitivo:	tárgyas	Ag	4L		#
trapassato prossimo:	trapassato prossimo (igeidő)	SPr	7L		
trarre:	húz, von	V	2LAP	695	
trascorrere:	eltölt	V	5L	4728	
trascorso:	eltöltött	Ag	4L		
trasformare***:***	átalakít	V	1Q	1289	
trasformarsi:	átalakul	V	11L	1289	
trasmissione:	közvetítés, műsor	S	10L	1079	
trasporto:	közlekedés	S	2L	1577	
trattare:	bán valakivel	V	8L	337	
trattarsi (di):	szó van valamiről	V	4L	337	
trattoria:	kisvendéglő	S	6L	6144	#
traversa:	keresztút	S	10LA		
tremendo:	rettenetes, hatalmas	Ag	3L	1905	
treno ad alta velocità:	nagysebességű vonat	SPr	5L		
treno:	vonat	S	0L	1330	
trentina:	mintegy harminc	Nu	8Q	6505	
tricolore:	háromszínű	Ag	6L		
trio:	trió, hármas	S	7L	6505	#
trionfare:	ünnepel	V	4L		#
triste:	szomorú	Ag	1L	3324	
troppo:	túl sok	Ag	3L	1406	
troppo:	nagyon, túlságosan	Avv	1L	329	
trovare:	talál	V	2L	83	
trovarsi:	található valahol	V	2L	83	
t-shirt, *la:*	póló	S	9L		#
tu:	te	P	0L	34	
tubetto:	tubus	S	8L		
tuo:	tied	Ag	0L	209	
turco:	török	S	7L	6505	
turista, *il*/*la*:	turista	S	0L	5253	
turno:	turnus; műszak, munkaidő	S	1L; 5L	3764	

tutelare:	óv, védelmez	V	8L	1380	
tutto:	az egész	Ag	1L	49	
tutto:	minden, mindenki	P	1L	63	
tv, *la*:	tv	S	3L	3650	
uccidere:	megöl	V	4LAP	1221	
uffa:	ejnye	I	3L	6505	
ufficiale:	hivatalos	Ag	10L	1156	
ufficio postale:	postahivatal	SPr	1L		
ufficio, *l'*:	iroda	S	1L	372	
uguale:	ugyanolyan, egyforma	Ag	10L	487	
ultimamente:	az utóbbi időben	Avv	4L	1987	
ultimo:	utolsó	Ag	0L	198	
umido:	nedves	Ag	8L	3691	
umore, *l' (m.)*:	hangulat	S	8L	4342	
un po':	egy kicsit	AvvPr	5L	x	
un sacco (di):	egy csomó	AgPr	4L	2	
ungherese:	magyar	Ag	0L	6505	
unico:	egyetlen	Ag	8L	5357	
unione:	unió	S	6Q	2083	
unire:	egyesít, összeköt	V	1L	1398	
unità:	lecke; egység	S	0LAP; 5L	920	
università:	egyetem	S	0LAP	749	
uno:	egy	Art	0L	4	
uno:	egy	Nu	0L	109	
uomo:	férfi	S	1L	204	
uovo:	tojás	S	4Q	2861	
urbano:	városi	Ag	2L	2466	
urlare:	kiabál, üvölt	V	10L	2452	
usare:	használ	V	0L	316	
usato:	használt, igénybe vett	Ag	2L		
uscire:	kimegy, elmegy	V	0L	233	
uscita:	kijárat	S	0LAP	2060	
uso:	használat	S	2L	1608	
utensile, *l' (m.)*:	eszköz	S	6L	6505	#
utile:	hasznos	Ag	3L	2257	
utilizzare:	alkalmaz, használ	V	3L	608	
utilizzato:		Ag	8L	6505	
uva:	szőlő	S	8L	6505	
va bene:	rendben	Ipr	5L	1004	
vacanza:	nyaralás	S	0L	3302	
vacanze studio:	tanulmányút	SPr	10L		
valere la pena:	megéri	VPr	2L	7	
valigia:	bőrönd	S	5L	6505	
valutare:	értékel	V	9QTF	1259	
vantaggio:	előny	S	5L	1626	
variabile:	változékony	Ag	5L	4998	#
varietà, *il*:	varieté, show-műsor	S	10L	5357	
varietà, *la*:	választék	S	6L		
vario:	különböző	Ag	2L	278	
vasetto:	befőttesüveg, kis üveg	S	8L		
vaso:	váza	S	3L	4888	
vecchietta:	anyóka	S	5L	6505	
vecchio:	öreg, régi	Ag	1L	1736	
vedente:	látó	Ag	11L		
vedere:	lát	V	2L	42	
veloce:	gyors	Ag	5L	1713	
velocemente:	gyorsan	Avv	4Q	2752	
velocità:	gyorsaság	S	4L	4455	
vendere:	elad	V	2L	351	
vendita:	eladás	S	7L	1379	
venerdì:	péntek	S	2L	1270	
veneziano:	valencei	Ag	5QTF		

venire:	jön	V	2L	66	
ventina:	kb. húsz	Nu	5Q	3460	#
vento:	szél	S	5L	705	
veramente:	igazán, valóban	Avv	2L	156	
verbale:	igei	Ag	8L	3637	
verbo:	igen	S	0L	2704	#
verde:	zöld	Ag	1L	723	
verdura:	zöldség	S	6L	2757	
vergine:		S	5L	2549	
verificare:	ellenőriz, bizonyít	V	0L	199	
verità:	igazság	S	2L	565	
vero e proprio:	igazi, hamisítatlan	AgPr	9L		
vero:	igaz	Ag	1L	143	
verso:	felé	Pz	1L	275	
verso:	szöveg, vers	S	11L	1582	
verticale:	függőleges	Ag	2LAV	6505	
vestire:	öltöztet	V	9L	1807	
vestirsi:	öltözködik	V	9L	1807	
vestito:	ruha	S	0L	1327	
vetrina:	kirakat	S	9L	5158	
via:	út, utca	S	1L	336	
viabilità:	járhatóság	S	3L		#
viaggiare:	utazik	V	2L	5470	
viaggio:	utazás	S	1L	781	
vicino (a):	valamihez közel	PzPr	1L	x	
vicino:	közel	Ag	1L	832	
vicino:	szomszéd	S	3Q	2632	
videogioco:	videojáték	S	2L		#
vigile del fuoco:	tűzoltó	SPr	3L	x	
vigile urbano, *il*:	rendőr	SPr	9Q	12	
vigile:	őr	S	7L	3061	
vignetta:	rajz	S	7L	6505	
villa:	villa	S	1QTF	1917	
vincere:	nyer	V	2L	763	
vino:	bor	S	2QTR	1757	
virtù, *la*:	erény	S	0LAP	6505	
viscosa:	viszkóz	S	9L	6505	#
visita:	látogatás	S	6Q	1413	
visitare:	meglátogat	V	1L	2195	
viso:	arc	S	1L	2936	
vista:	látás	S	9Q	338	
visto che:	mivel	CPr	9QA	14	
vita:	élet	S	2L	227	
vitamina:	vitamin	S	8LAV	6505	
vitello:	borjú	S	6LA	4566	
vivace:	élénk	Ag	9L	5101	
vivamente:	nagyon, szívből	Avv	8L	4350	
vivere:	él	V	1Q	344	
vocabolario:	szótár	S	11Q	2154	
voce, *la:*	hang	S	11L	421	
volare:	repül	V	11L	1665	
volentieri:	szívesen	Avv	2L	1319	
volere bene:	szeret	VPr	10Q	7	
volere:	akar	V	2L	35	
volgere:	áttesz, átalakít	V	4Q	2806	
volo:	repülés, repülőjárat	S	5QA	1755	
volta:	alkalom	S	1L	89	
volto:	arc	S	7L	6505	
volume:	hangerő	S	8Q		
vostro:	tietek	Ag	1L	303	
vuoto:	üres	Ag	8L	1241	
yogurt, *lo*:	joghurt	S	0LAP	6505	#
zabaione, *lo*:	borsodóhoz hasonló édesség	S	4L	6505	#

zaino, *lo:*	hátizsák	S	0L	5205	#
zero*:*	nulla	Nu	0L	359	
zia:	nagynéni	S	0L	1884	
zio, *lo*:	nagybácsi	S	0L	1924	
zona:	városrész	S	6LA	322	
zucchero:	cukor	S	3L	4207	

Lista B

In questa piccola lista sono riuniti i c.d. fonosimboli, privi di un corrispondente ungherese. Per ciascun lemma si indica l'unità in cui compare per la prima volta.

beh:	3L
mah:	1L
ehi:	1L

Lista C

Questa lista contiene i nomi propri estratti dal glossario.

Adriatico
Alpi, *le*
America
Ancona
Annibale
Aprilia
Arezzo
Argentina
Ariston
Asti
Australia
Babbo Natale
Barcellona
Bari
Befana
Belgio
Belpaese
Boccaccio
Bologna
Brasile, *il*
Budda
Canarie, isole
Cannes
Cartagine
Cesare
Cina
Cleopatra
Colosseo
Como
Cristo