Il lessico e i temi della statistica ufficiale in Italia. Un'analisi lessicometrica del Programma Statistico Nazionale degli ultimi dieci anni

Isabella Mingo [1], Cristina Panattoni [2]

[1] Dipartimento di Sociologia e Comunicazione –Università Sapienza Roma- Italia

[2] Segreteria tecnica della Commissione per la Garanzia dell'Informazione Statistica – Presidenza del Consiglio dei Ministri – Italia

Riassunto

Una adeguata valutazione della completezza dell'informazione statistica ufficiale italiana presuppone una conoscenza approfondita e una attenta analisi del principale documento di pianificazione e coordinamento delle statistiche prodotte dagli enti del Sistema statistico nazionale (Sistan): il Programma Statistico Nazionale (PSN). Infatti, il d.lgs. 322/89, istitutivo del Sistan, stabilisce che nel PSN debba essere inserita tutta la produzione statistica di interesse pubblico realizzata in ciascun triennio dagli enti del Sistema.

Nel tempo la struttura e i contenuti del PSN sono cambiati in funzione sia dei mutati fabbisogni informativi del Paese e dell'emergere di nuovi settori tematici di indagine su cui acquisire dati, sia della maggiore attenzione posta alla qualità dell'intero processo di produzione dell'informazione statistica. In tale contesto il presente lavoro propone, mediante un approccio lessicometrico, una analisi longitudinale dei Programmi statistici nazionali – dal triennio 2000-2002 al triennio 2008-2010 – con un duplice obiettivo: porre le basi per la costruzione di un lessico della statistica ufficiale e delineare, sia sul piano lessicale che su quello delle tematiche, i cambiamenti intervenuti nel corso dell'ultimo decennio. La dimensione totale del corpus analizzato è di circa 600 mila occorrenze. Ai fini di un approfondimento tematico sono stati analizzati anche sub-testi estratti dal corpus, costituiti dai lavori progettati per ciascun triennio.

Abstract

An adequate assessment of the completeness of the Italian official statistics requires a knowledge and careful analysis of the main document for planning and coordination of statistics produced by National Statistical System (Sistan): National Statistical Program (NSP). Indeed, the legislative act of the government no. 322/89 (d.lgs. 322/89), establishing the Sistan, states that the NDP should include the statistical production of public interest carried out in each three years by institutions of the system. Over time the structure and contents of the NSP have changed, both the changing information needs of the country and the emergence of new issues of inquiry, both the increased attention to the quality of the production process of statistics. In this context the present work proposes, through a lexicometric approach, a longitudinal analysis of National Statistical Programs – from the period 2000-2002 to the 2008-2010 – with a dual aim: to lay the foundations for a lexicon of official statistics and outline, in terms of content, changes in the issues during the last decade. The total size of the corpus analysis is about 600 thousand hits. A significant gain of thematic deepening was also obtained analyzing sub-tests taken from the corpus, consisting of the works designed for each triennium.

Keywords: official statistics, lexicometric analysis, textual analysis.

1. Introduzione

Il Programma Statistico Nazionale (PSN), quale principale documento di programmazione delle statistiche prodotte dal Sistema statistico nazionale (Sistan), offre lo spunto per una analisi finalizzata alla definizione del lessico specifico della statistica ufficiale. Del resto, la tempistica annuale della stesura del PSN e il suo orizzonte di programmazione, che investe un intero triennio, inducono a ritenere che nel documento stesso siano trattate in maniera esaustiva tutte le questioni e le tematiche toccate dalla statistica ufficiale. Il PSN diviene allora materiale unico e indispensabile per una analisi lessicale e per una esplorazione dei principali argomenti dell'informazione statistica ufficiale.

Nell'ultimo decennio la struttura e i contenuti del PSN sono cambiati. Le aree e i settori tematici di riferimento per la produzione statistica del Sistan sono aumentati e sono stati diversamente organizzati e dettagliati per rispondere ai nuovi fabbisogni informativi del Paese.

Anche la numerosità dei progetti inseriti nel piano triennale e la loro composizione per tipologia (rilevazioni, elaborazioni, studi progettuali) hanno subito delle variazioni (Tab. 1). Nel triennio 2008-2010, inoltre, è stata introdotta una quarta tipologia di lavoro, quella cioè dei "sistemi informativi statistici", per soddisfare l'esigenza via via più sentita di integrare le diverse fonti.

	Triennio di programmazione								
Tipologia di progetto	*2000/ 2002*	*2001/ 2003*	*2002/ 2004*	*2003/ 2005*	*2004/ 2006*	*2005/ 2007*	*2006/ 2008*	*2007/ 2009*	*2008/ 2010*
Rilevazione	417	439	449	467	466	472	464	446	447
Elaborazione	450	470	451	418	437	448	458	444	437
Studi progett.	258	277	196	191	182	179	193	233	236
Sistemi inf. stat.	–	–	–	–	–	–	–	–	26
Totale	*1.125*	*1.187*	*1.096*	*1.076*	*1.085*	*1.099*	*1.115*	*1.123*	*1.146*

Tabella 1: Numero di progetti presenti nel PSN per tipologia e triennio di programmazione

Peraltro, la maggiore attenzione posta dagli enti del Sistema alla qualità, alla completezza e alla accessibilità dell'informazione statistica prodotta, anche a seguito della recente normativa sulla privacy, ha portato alla diversa struttura formale del documento attualmente composto da due parti. Nella prima sono descritti gli obiettivi e i progetti previsti nel triennio; nella seconda sono contenute, per ciascun progetto, dettagliate schede tecniche.

Prendendo in considerazione i Programmi statistici nazionali dal triennio 2000-2002 al triennio 2008-2010, il presente contributo illustra i principali risultati dell'analisi lessicometrica realizzata sul corpus costituito dalla raccolta dei PSN – parte prima – dei diversi trienni considerati. L'obiettivo dello studio è quello di porre le basi per la costruzione di un lessico della statistica ufficiale e delineare secondo un'ottica longitudinale, sia sul piano lessicale sia su quello del contenuto, i cambiamenti intervenuti nell'offerta di informazione statistica pubblica nel corso dell'ultimo decennio.

In particolare nel paragrafo 2 viene esplorato l'intero corpus, pari a circa 600 mila occorrenze, evidenziandone le caratteristiche lessicali, il linguaggio fondamentale e peculiare, nonché le variazioni intervenute nel tempo riguardo all'attenzione prestata ai requisiti di qualità della produzione statistica. Nel paragrafo 3 viene focalizzata l'attenzione sul sub-testo dei soli titoli dei progetti inseriti nei diversi piani triennali per individuare, attraverso un'analisi delle corri-

spondenze, le dimensioni fondamentali su cui si è imperniata la statistica ufficiale nell'ultimo decennio. Le analisi sopraesposte sono state realizzate mediante l'utilizzo dei *software* Lexico, Taltac e Spad.

2. Verso il lessico della statistica ufficiale: procedure e primi risultati

Il principale obiettivo di uno studio lessicale del principale documento di programmazione delle statistiche ufficiali nazionali è quello di ricostruire il linguaggio utilizzato, al fine di individuare le caratteristiche di un lessico settoriale di cui sarà in seguito possibile studiare l'evoluzione temporale. Un primo obiettivo sarà pertanto quello di individuare l'esistenza di un "linguaggio fondamentale" (Bolasco, 1996), ossia di un insieme di termini più frequentemente utilizzati nel corpus considerato. Si procederà poi a "misurare" la peculiarità del linguaggio della statistica ufficiale, intesa invece come uso di forme non "banali" e dunque non abitualmente riscontrabili nel linguaggio corrente. Un terzo momento sarà infine quello di individuare eventuali specificità lessicali del linguaggio usato nei diversi anni, al fine di introdurre un'analisi tematica dei PSN, utile a ricostruire i cambiamenti nel tempo corrispondenti ai mutati fabbisogni conoscitivi della collettività nazionale.

2.1. Caratteristiche del corpus e dei subcorpora

L'insieme dei testi dei PSN degli ultimi dieci anni, che compongono il corpus analizzato, è costituito, a seguito della normalizzazione effettuata [1], da 596.101occorrenze (N), di cui 19.137 forme grafiche diverse (V): ha pertanto un'ampiezza adeguata per la costruzione di un lessico settoriale, quale può considerarsi quello della "statistica ufficiale" [2]. Gli indici di ricchezza lessicale forniscono alcune indicazioni sulle caratteristiche lessicometriche del corpus analizzato: il type/token ratio (V/N*100) pari a 3,21 potrebbe essere considerato un indicatore della settorialità del linguaggio utilizzato e della scarsa ricchezza lessicale; il coefficiente "a" risulta uguale a 1,34, il numero di *hapaxes* (V_1) sul totale delle forme grafiche diverse è di circa il 32% [3].

Trattandosi di un corpus composto da singoli PSN, l'analisi longitudinale dei subtesti mostra come la loro ampiezza, nell'ultimo decennio, abbia seguito una tendenza all'aumento; i valori estremi si collocano rispettivamente nel 2001-2003 in cui si registra il PSN più breve con poco più di 43mila occorrenze e nel 2005-2007, in cui si rileva invece il corpus più ampio, di circa 81mila occorrenze. Tuttavia è con il PSN del 2003-2005, triennio in cui entra a regime la programmazione su nuove aree tematiche di interesse per la produzione statistica, che si registra un significativo aumento dell'ampiezza del documento di programmazione. Diminuiscono invece gli indici di ricchezza lessicale, in special modo l'incidenza del numero di *hapaxes* sulle forme diverse ed il valore del coefficiente G ($V/\sqrt{N}$) – che indicano una tendenziale diminuzione della varietà linguistica del discorso (Fig. 1).

[1] Prima della normalizzazione il corpus presentava circa 635mila occorrenze di cui oltre 17 mila forme grafiche diverse. La normalizzazione è stata effettuata tramite le liste disponibili in Taltac 2.5.

[2] Nel caso del Lif e del Lip, lessici di frequenza della lingua italiana, l'ordine di grandezza di campioni di testi raccolti è stato intorno alle 500 mila occorrenze; il lessico del discorso programmatico di governo ha invece utilizzato un corpus di circa 700 mila occorrenze (Bolasco, 1996).

[3] Per gli indici lessicometrici utilizzati si veda tra gli altri Bolasco (1999). Valori di a=logN/logV superiore a 1,3 indicano che il vocabolario non è particolarmente ricco (Tuzzi, 2003). L'interpretazione puntuale delle misure di ricchezza lessicale andrebbe compiuta riferendosi ad altri corpora della stessa ampiezza.

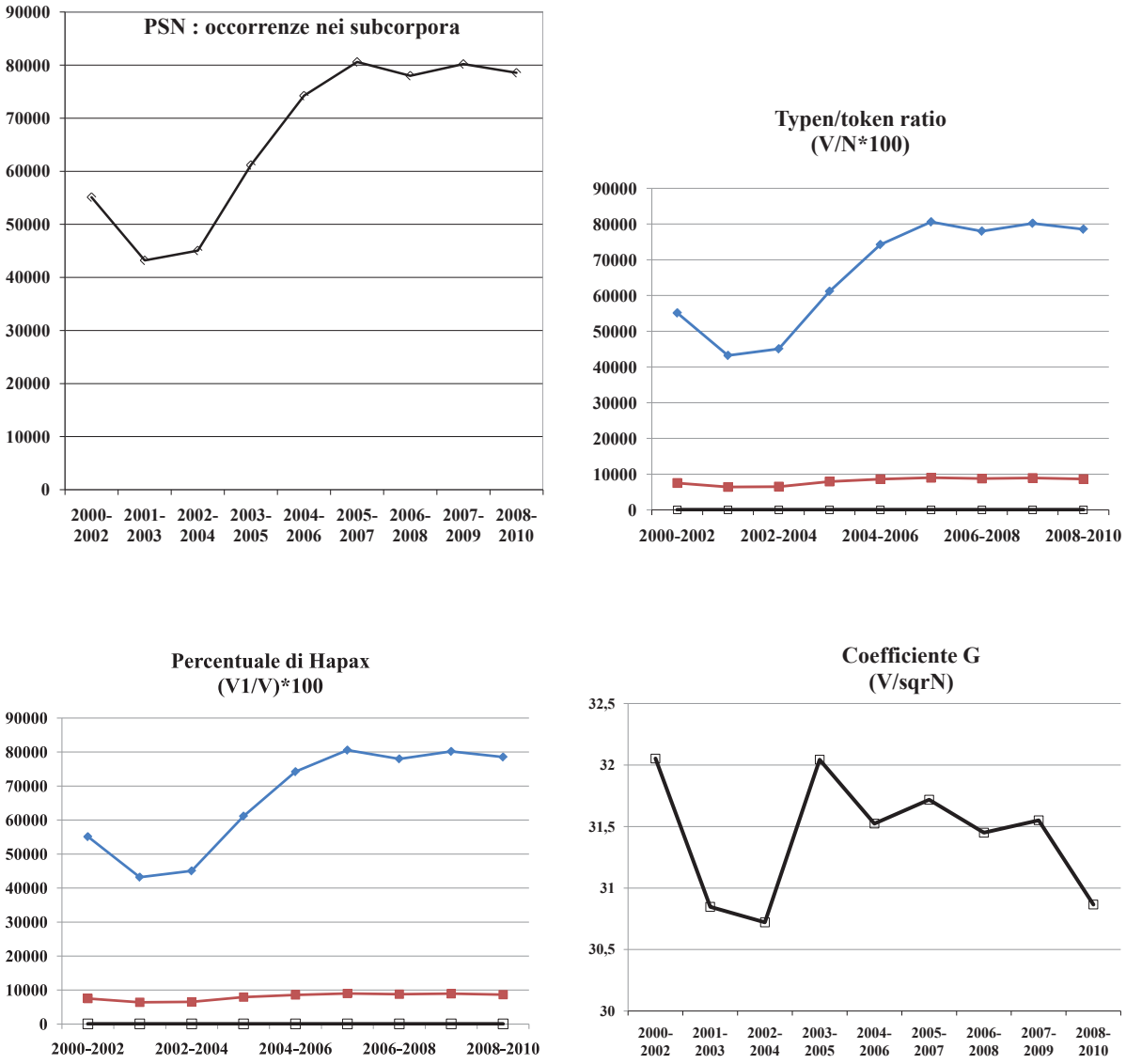

Figura 1: Caratteristiche lessicometriche dei subcorpora

2.2. Il linguaggio fondamentale

Il lessico comprende complessivamente 14.608 lemmi; se si eliminano acronimi (sigle, enti), stranierismi e numerali (circa 3.600), il numero si riduce a 10.954. Per individuare il linguaggio fondamentale, inteso come insieme di vocaboli che ricorrono più spesso all'interno dell'idioma considerato, si è utilizzato il criterio dell'indice d'uso (Bolasco, 1999: 108) calcolato dividendo il corpus nelle 9 sezioni individuate dai diversi anni. In particolare, sono stati considerati i primi 3.000 lemmi a cui è associato l'indice d'uso più elevato: il vocabolario fondamentale, così individuato, comprende termini con frequenza d'uso non inferiore a 11,7 e con numero di occorrenze almeno pari a 13. Limitando l'analisi solo a primi 1.000 lemmi invece la soglia di frequenza d'uso sale a 60,68 e il numero di occorrenze non è inferiore a 64.

Le principali caratteristiche di tale vocabolario si evincono dall'analisi dei principali lemmi più utilizzati, distinguendo tra le categorie grammaticali: verbi, sostantivi e aggettivi (Tab. 2 [4]).

[4] Per brevità si riportano in Tab. 2 solo i primi 35 lemmi di ciascuna categoria.

VERBI			*SOSTANTIVI*			*AGGETTIVI*		
Lemma	*Disp.*	*Uso*	*Lemma*	*Disp.*	*Uso*	*Lemma*	*Disp.*	*Uso*
essere	0,98	6769,97	rilevazione	0,96	3716,84	informativo	0,94	1556,40
avere	0,95	2294,54	attività	0,98	3190,24	nazionale	0,98	1461,49
venire	0,92	1139,96	informazione	0,98	3039,60	statistico	0,95	1344,24
potere	0,97	625,40	sistema	0,95	3029,32	economico	0,97	1066,97
dovere	0,92	570,22	indagine	0,95	2948,00	territoriale	0,95	1044,40
effettuare	0,94	548,24	progetto	0,95	2400,72	nuovo	0,96	933,76
utilizzare	0,94	523,51	impresa	0,97	1937,52	relativo	0,92	816,77
consentire	0,94	463,08	produzione	0,97	1864,99	europeo	0,93	781,42
prevedere	0,92	381,22	settore	0,95	1539,43	pubblico	0,95	775,75
riguardare	0,94	375,42	studio	0,94	1485,90	amministrativo	0,94	739,56
fornire	0,95	349,93	analisi	0,97	1390,16	sociale	0,94	729,32
migliorare	0,92	265,15	lavoro	0,96	1266,75	agricolo	0,91	632,20
rendere	0,95	247,60	servizio	0,96	1265,49	internazionale	0,95	606,90
proseguire	0,91	225,49	sviluppo	0,95	1199,12	progettuale	0,86	575,77
produrre	0,93	214,60	obiettivo	0,97	1144,70	regionale	0,95	560,72
costituire	0,93	201,56	conto	0,94	1073,57	principale	0,97	552,57
completare	0,93	192,95	indicatore	0,94	1023,79	congiunturale	0,96	545,78
realizzare	0,95	174,93	indice	0,94	1023,07	comunitario	0,92	543,88
ottenere	0,96	171,64	elaborazione	0,91	1021,91	locale	0,93	528,18
individuare	0,91	163,34	processo	0,94	990,90	ambientale	0,90	457,53
permettere	0,93	160,19	anno	0,95	962,76	strutturale	0,94	392,78
rappresentare	0,92	154,44	diffusione	0,97	854,99	attuale	0,92	390,53
sviluppare	0,88	146,52	famiglia	0,96	794,76	italiano	0,89	371,91
andare	0,89	145,76	popolazione	0,93	762,04	di_qualità	0,94	370,15
garantire	0,90	138,97	prodotti	0,95	760,10	generale	0,90	367,75
disporre	0,89	138,70	qualità	0,97	738,97	metodologico	0,94	356,10
segnalare	0,74	131,23	fonte	0,94	730,38	maggiore	0,93	346,90
valutare	0,91	128,78	istituto	0,96	718,08	sanitario	0,93	343,97
definire	0,89	124,14	costruzione	0,93	717,59	culturale	0,94	338,69
procedere	0,90	122,71	stima	0,94	717,31	disponibile	0,92	336,13
cadenzare	0,88	119,51	ministero	0,93	715,42	istituzionale	0,96	331,95
soddisfare	0,90	117,10	censimento	0,79	713,48	campionario	0,93	331,14
rispondere	0,92	115,13	risultati	0,97	711,87	comunale	0,93	319,98
continuare	0,88	115,04	triennio	0,96	680,20	ulteriore	0,88	317,81
rilevare	0,86	114,94	realizzazione	0,96	678,46	annuale	0,94	298,93

Tabella 2: Il linguaggio fondamentale: i primi 35 verbi, sostantivi e aggettivi per indice d'uso (lemmi)

Considerando i sostantivi, essi delineano gli strumenti metodologici ed operativi dell'attività statistica (es: <rilevazione>,<indagine>, <analisi>, <studio>, <elaborazione>, <costruzione>, <censimento>, <stima>, <indicatore>, <indice>), gli ambiti di applicazione (es: <impresa>, <lavoro>, <famiglia>, <popolazione>), ma anche le finalità dell'attività stessa (es: <informazione>, <produzione>, <diffusione>, <integrazione>) [5].

Per quanto riguarda gli aggettivi, il primo lemma con indice d'uso più elevato risulta <informativo> quasi a sottolineare il principale obiettivo della produzione delle statistiche ufficiali, seguito da <nazionale> e <statistico> che definiscono con precisione gli ambiti dell'attività di informazione. Spiccano inoltre qualificazioni che definiscono i contesti territoriali di riferimento (<territoriale>, <europeo>, <internazionale>, <regionale>, <comunitario, <comunale>); altri

[5] Gli aspetti semantici sono stati verificati mediante un'analisi delle concordanze.

evocano le aree tematiche di interesse (<economico>, <sociale>, <agricolo>, <ambientale>, <sanitario>, <culturale>). Altri attributi si riferiscono invece a caratteristiche tipiche della produzione statistica (es: <congiunturale>, <strutturale>, <generale>, <campionario>). Infine per quanto riguarda i verbi, oltre a quelli con funzione di ausiliare, il primo verbo con indice d'uso più elevato risulta <effettuare> seguito da <utilizzare>, chiaramente connessi con le attività di realizzazione delle indagini e di uso di informazioni già rilevate. Tra gli altri verbi spiccano quelli che esprimono "buone intenzioni" (es: <proseguire>, <completare>, <realizzare>, <sviluppare>, <garantire>, <soddisfare>, <erogare>, <aumentare>); altri invece si riferiscono alla strumentalità della produzione delle fonti statistiche (<consentire>, <prevedere>, <riguardare>, <fornire>, <rendere>, <segnalare>, <definire>, <valutare>).

La dispersione di ciascun lemma, individuata in base alla sua presenza nel corso del decennio considerato, è un primo indizio della presenza di una specificità temporale del lessico: infatti, circa il 56% dei lemmi ha una dispersione d'uso inferiore al 60%, si concentra dunque solo in alcuni dei periodi considerati caratterizzandone il lessico.

2.3. Le forme idiomatiche

L'analisi dei segmenti di forme grafiche che si ripetono nel corpus consente di individuare le principali forme idiomatiche del linguaggio della statistica ufficiale. Alla soglia di 10 occorrenze e selezionando un numero di forme per segmento non superiore a 6, si ottengono 8.309 segmenti. L'estrazione delle forme idiomatiche più significative è stata compiuta utilizzando come criterio l'indice IS relativo (Morrone, 1993) che permette di identificare alcuni poliformi del lessico considerato. Limitandosi in questo contributo ad una prima lettura di alcune fra le forme più ricorrenti tra quelle con indice IS>0,60 (Tab. 3), emergono con chiarezza alcuni gruppi nominali che possono considerarsi tipici del lessico in esame: <studio progettuale>, <studi progettuali>, <serie storiche>, <mancate risposte>, <codifica automatica>, <schede identificative>. Si evidenziano inoltre alcune polirematiche connesse con gli ambiti tematici (es: <procedimenti penali>, <ore lavorate>, <esercizi ricettivi>, <numeri civici>, <comunità montana>, <tronco stradale>, <richiedenti asilo>, <canali distributivi>) nonché alcuni segmenti composti da stranierismi entrati a far parte del linguaggio della statistica ufficiale: <Action Plan>, <Information Society>, <Urban Audit>, <supply and use>.

2.4. Il linguaggio peculiare

L'analisi del linguaggio fondamentale, basato sulla frequenza d'uso dei lemmi, ha evidenziato la prevalenza di parole-tema che sembrano delineare in modo evidente la settorialità del lessico analizzato. Una ulteriore verifica può essere tuttavia compiuta mediante l'individuazione delle parole sovra/sotto utilizzate – in termini di scarti standardizzati delle frequenze relative (Muller, 1977; Bolasco, 1999) – rispetto a lessici di frequenza assunti come modello di riferimento.

Ci si limita in questo contributo a considerare soltanto i lemmi dei verbi sovrautilizzati rispetto a due lessici di riferimento disponibili in Taltac: quello dell'italiano standard (Polif2002) e del linguaggio comune (Rep90). Considerando per brevità solo i primi 20 lemmi con scarto più elevato, emerge che la maggior parte di essi risulta peculiare rispetto a entrambi i lessici di riferimento, anche se le due graduatorie presentano disallineamenti (Tab. 4). I verbi peculiari evocano alcune azioni che caratterizzano l'attività di produzione delle statistiche (<cadenzare>, <cofinanziare>, <effettuare>, <implementare>, <monitorare>), ma anche le intenzioni tipiche di un documento di programmazione, in cui si intende <proseguire>, <migliorare>, <completare>, <soddisfare>, <sviluppare>, <ampliare> <integrare>. È inoltre interessante notare come la maggior parte dei primi venti verbi peculiari sia anche nella lista dei primi verbi più frequenti che caratterizzano il

linguaggio fondamentale (Tab. 2), anche se in posizioni diverse rispetto alla graduatoria ottenuta in base all'indice d 'uso. Ciò contribuisce a evidenziare la settorialità del linguaggio utilizzato.

Segmento	*Occ.*	*Indice IS relativo*	*Segmento*	*Occ.*	*Indice IS relativo*
studio progettuale	402	0,62	casse edili	29	0,73
situazione attuale	239	0,73	delitti denunciati	26	0,66
prospettive evolutive	170	0,88	supply and use	26	0,63
studi progettuali	165	0,65	fabbricato residenziale	24	0,60
procedimenti penali	115	0,65	sentenza irrevocabile	19	0,75
ore lavorate	104	0,88	tronco stradale	18	0,65
serie storiche	104	0,60	schede identificative	18	0,62
Bilanci consuntivi	89	0,61	di prima accoglienza	16	0,67
mancate risposte	69	0,75	richiedenti asilo	16	0,62
codifica automatica	64	0,60	interruzioni volontarie	15	0,76
medio lungo	61	0,75	Information Society	15	0,71
malattie infettive	56	0,78	canali distributivi	15	0,67
bilanci civilistici	53	0,62	violenza sessuale	15	0,63
portati avanti	52	0,62	evitare sovrapposizioni	15	0,62
Action Plan	49	0,71	rotabili ferroviari	12	0,69
intermediazione finanziaria	48	0,71	a favore degli studenti	12	0,63
posti vacanti	47	0,76	batterie esauste	11	1,00
acqua potabile	39	0,72	rappresentanze diplomatiche	11	0,92
numeri civici	38	0,75	difetti congeniti	11	0,74
esercizi ricettivi	35	0,75	separata legalmente	11	0,74
comunità montane	35	0,69	Urban Audit	11	0,67
incidenti stradali	33	0,67	computer palmari	11	0,63

Tabella 3: Segmenti ripetuti più frequenti con indice IS> 0.60

Italiano standard (Polif2002)		*Linguaggio comune stampa (Rep90)*	
Scarto	*Lemma*	*Scarto*	*Lemma*
133300,00	cadenzare	22101,81	effettuare
40567,83	cofinanziare	17041,34	essere
32344,76	implementare	14119,40	monitorare
24550,11	effettuare	12737,11	fornire
19869,31	monitorare	12405,25	prevedere
19224,74	utilizzare	12337,11	proseguire
18058,62	essere	11658,93	migliorare
16205,49	fornire	11270,67	consentire
15272,75	consentire	11061,94	venire
14102,02	migliorare	11051,74	utilizzare
13713,90	venire	10988,63	completare
12523,90	completare	10801,53	segnalare
11412,23	soddisfare	8675,56	confluire
11116,71	prevedere	8635,62	erogare
10997,95	proseguire	7692,55	avere
10806,72	sviluppare	7479,50	produrre
10709,46	ampliare	7466,94	ampliare
10227,70	integrare	7395,10	stipulare
10150,00	evidenziare	7080,73	implementare
10090,15	riguardare	7063,59	riguardare

Tabella 4: Primi 20 lemmi di verbi sovrautilizzati nel PSN

2.5. Il linguaggio specifico: l'importanza della "qualità"

L'estrazione del linguaggio specifico delle singole partizioni in cui è suddiviso il corpus, consente di individuare le sue "peculiarità interne". In questo contributo, per brevità, ci si limita a riportare quella compiuta al fine di ordinare i diversi sub-corpora – e dunque i diversi trienni di programmazione – secondo l'importanza attribuita [6] a parole chiave che ruotano intorno ai requisiti di "qualità" dell'informazione statistica (Codice delle statistiche europee, COM (2005) 217 def.): rilevanza (rilevan*), trasparenza (traspar*), comparabilità (comparabil*), coerenza (coeren*), accuratezza (accurat*), tempestività (tempest*), completezza (complet*), riservatezza (riservat*) [7]. L'attenzione per la "qualità", così operativizzata, risulta crescente dal triennio 2001-2003 fino al 2005-2007. Nel PSN del 2006-2008 sembra tuttavia registrarsi, rispetto agli altri anni, un momentaneo calo nell'utilizzo di tale terminologia, ripresa nel triennio successivo: è infatti nel PSN del 2008-2010 che l'indice TF-IDF risulta più elevato (Tab. 5). È interessante notare che le prime due forme più rilevanti <comparabilità> e <accuratezza> rinviano a problematiche rilevanti per la qualità sia nella fase di produzione sia in quella di fruizione dei prodotti statistici.

Sub-corpora		*Prime 20 forme per TFID*			
Triennio	*TFIDF*		*TFIDF*		*TFIDF*
2000-2002	2.96	comparabilità	1.77	accessibilità	0.90
2001-2003	2.47	accuratezza	1.68	completati	0.85
2002-2004	2.78	tempestive	1.24	completerà	0.84
2003-2005	3.23	tempestivi	1.09	qualità del servizio	0.80
2004-2006	3.53	comparabile	1.08	qualità/coerenza	0.80
2005-2007	3.72	complete	1.02	completando	0.80
2006-2008	1.70	qualitativamente	1.00	qualitativi	0.76
2007-2009	3.11	qualità	0.98	qualitativa	0.75
2008-2010	3.75	qualitative	0.98	accesso	0.74
		tempestivo	0.96	completano	0.72

Tabella 5: Rilevanza dei requisiti di "qualità"

3. I temi della "statistica ufficiale" nell'ultimo decennio

3.1. Pre-trattamento del testo e scelte operative

Nella seconda parte di questo contributo, l'attenzione è stata focalizzata su una parte del corpus costituito esclusivamente dai titoli dei progetti, afferenti alle diverse aree e settori tematici, presenti nei PSN dei diversi anni, allo scopo di passare dall'analisi lessicale all'analisi dei contenuti, attraverso un approccio, per così dire, "semplificato", trattandosi di una raccolta più snella ma di grande valore semantico.

Tale sub-corpus presenta particolari caratteristiche: è discontinuo da un punto di vista logico-grammaticale, è standardizzato poiché utilizza termini tecnici afferenti ad un ambito disciplinare statistico e conseguentemente è ricco di poliformi e poliremiatiche di contenuto specifico. Pertanto, inizialmente, si è ritenuto opportuno procedere, attraverso l'analisi dei segmenti ri-

[6] Il criterio statistico utilizzato è l'indice TF-IDF (Salton and Buckely 1988) normalizzato per evitare l'effetto della diversa lunghezza dei sub-corpora (cfr. Guida Taltac 2.5).

[7] Gli aspetti semantici sono stati verificati con un'analisi delle concordanze.

petuti, a una lessicalizzazione [8] delle poliremiatiche più frequenti quali, per esempio, <sistema informativo>, <multiscopo sulle famiglie>, <movimento dei procedimenti penali>, <istituti di cura>, <prezzi al consumo>, <aziende agricole>, <rilevazione mensile>, <banca dati>, <settore privato>, <forze di lavoro>, <produzione industriale>, <attività gestionali ed economiche delle Usl>, <studio progettuale>.

Successive scelte hanno riguardato la selezione dei termini della matrice "forme grafiche x anno" così ottenuta, con l'obiettivo di individuare e sintetizzare, mediante l'applicazione dell'analisi delle corrispondenze, l'evoluzione dei temi del PSN nel corso degli anni presi in considerazione. A tal fine, utilizzando sia le informazioni derivate dall'analisi delle specificità per anno, sia il criterio della frequenza di soglia consigliata [9], si è ottenuta una matrice "forme x anno", epurata dalle forme vuote e dalle forme idiomatiche di scarso interesse per gli obiettivi dell'analisi.

Infine, poiché l'esplorazione del vocabolario ha evidenziato due principali raggruppamenti semantici – uno riconducibile agli aspetti metodologici, l'altro agli aspetti sostantivi – si è ritenuto opportuno condurre due differenti analisi. La prima ha considerato la matrice contenente i termini (280) riferiti ai metodi e agli strumenti statistici utilizzati, alle fonti, al tipo di informazione statistica prodotta [10]. La seconda si è invece concentrata sui termini (274) riferiti alle tematiche specifiche di volta in volta indagate nel PSN.

3.2. La mappa dei metodi e degli strumenti

L'analisi delle corrispondenze (Lebart and Salem, 1988) condotta sulla prima matrice ha portato all'identificazione di due dimensioni rappresentate rispettivamente sul primo e sul secondo asse del grafico in Fig. 2 e che complessivamente spiegano il 60,23% della variabilità totale [11].

È interessante notare innanzitutto che gli anni si distribuiscono intorno al primo asse fattoriale in ordine cronologico secondo un andamento a "s" così da ritrovare i trienni dal 2000-2002 al 2005-2007 nella parte sinistra del grafico (2000-2002 e 2001-2003 nel III quadrante e 2002-2004, 2003-2005, 2004-2006 e 2005-2007 nel IV) e i restanti trienni in quella destra (2006-2008 nel I quadrante e 2007-2009, 2008-2010 nel II).

La dimensione rappresentata sul primo asse propone l'evoluzione dell'orientamento della statistica ufficiale, che risulta maggiormente rivolto verso "l'impianto della produzione" nella programmazione dei primi anni del decennio di osservazione, e verso "il consolidamento della produzione e l'attenzione alla diffusione" nella programmazione degli anni più recenti. I termini che influiscono nell'attribuzione di senso al primo semiasse negativo e che quindi si trovano localizzati nella parte sinistra del grafico sono: <metodologia>, <metodologico>, <stimatori>, <processi di produzione>, <metodologici>, <quantificazione>, <benchmarking>, <rilevamento>, <produzione statistica>, <telematica dei dati> [12], <metodologie di stima>, <rete di rilevazione>, <metodologica>, <stime degli aggregati>, <controllo della qualità>. I termini

[8] Dimensioni del sub-corpus considerato dopo la lessicalizzazione: 68.696 occorrenze. Indice Type Token ratio=5,88. Frequenza media generale (N/V)=17. Frequenza di soglia consigliata>23.

[9] Le analisi di specificità, che per brevità non vengono qui riportate, sono state condotte con Lexico.

[10] In questa matrice sono stati recuperati anche termini con minor frequenza, ritenuti importanti, sulla base di conoscenze *a priori*, per questa sfera di analisi.

[11] L'attribuzione di senso agli assi è avvenuta sulla base della valutazione complessiva delle informazioni derivanti dall'analisi congiunta dei contributi assoluti, di quelli relativi e della posizione sul grafico delle diverse forme.

[12] Termine che fa riferimento all'acquisizione, alla raccolta e alla trasmissione telematica dei dati dalle diverse fonti.

invece che si trovano localizzati nella parte destra del grafico e che contribuiscono a "battezzare" il primo semiasse positivo sono: <miglioramento>, <ampliamento>, <quadro di riferimento>, <sistema informativo integrato>, <spaziale>, <esigenze informative>, <micro-matching>, <flusso>, <statistica ufficiale>, <contenuti informativi>, <esplorativa>, <microdati>, <sistema di monitoraggio>, <georeferenziazione>, <banche dati>, <dati amministrativi>, <diffusione dei dati>, <programmazione>, <archivi amministrativi>, <rilevazione integrativa>, <base informativa>, <linkare>, <datawarehouse>. Seguendo il *continuum* del primo asse si può quindi osservare come nel tempo la statistica pubblica abbia ampliato i suoi orizzonti passando, dall'attenzione prevalente verso le metodologie e gli strumenti per la produzione, a una maggiore sensibilità verso i problemi della diffusione dei dati e della fruizione da parte di un pubblico sempre più ampio, anche mediante l'impiego delle nuove tecnologie dell'informazione e della comunicazione. In questo passaggio viene data risposta alle esigenze di informazione a più elevato dettaglio territoriale, alla diffusione di microdati, alla predisposizione di banche dati. Peraltro, l'implementazione di datawarehouse testimonia la tendenza degli ultimi anni a utilizzare strumenti di diffusione più completi, flessibili e accessibili per l'utenza. Sono, tra l'altro, proprio gli ultimi due trienni della serie a dare il loro maggiore contributo alla determinazione del primo semiasse positivo.

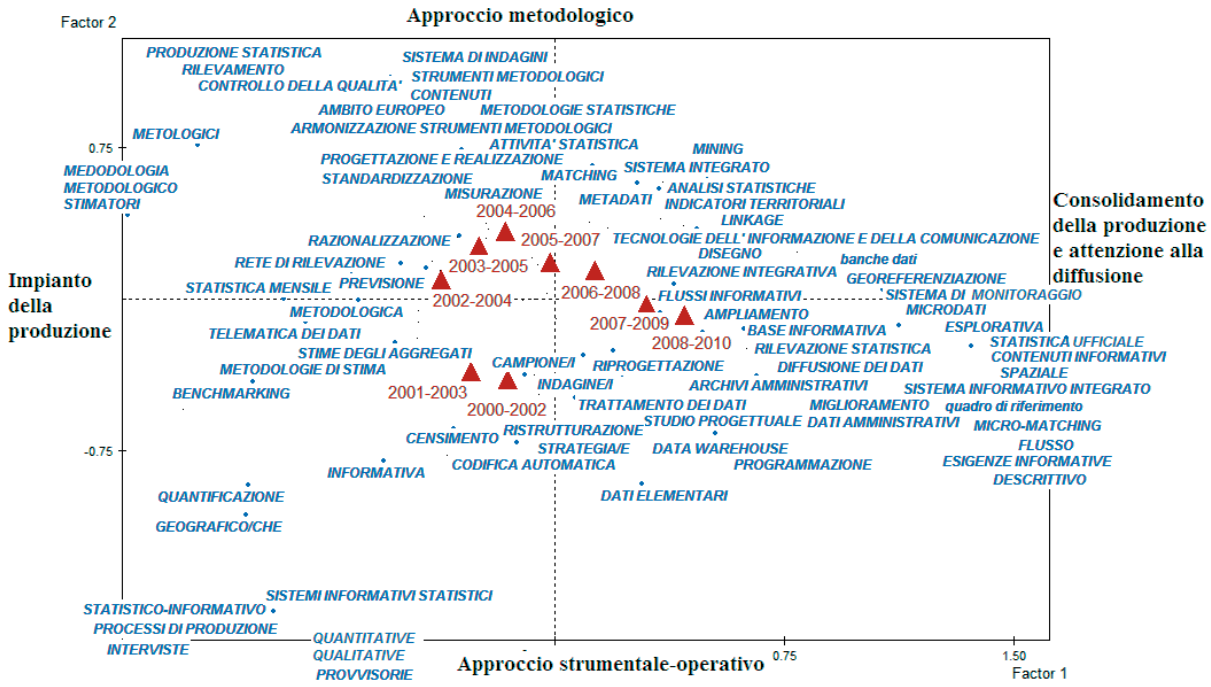

Figura 2: La mappa dei metodi e degli strumenti della statistica ufficiale dal 2000-2002 al 2008-2010

Il secondo asse descrive, lungo un *continuum,* quella che può essere definita "la strategia della costruzione del dato". Su di esso, infatti, si collocano da un parte (semiasse positivo) termini che fanno riferimento alle scelte metodologiche e, in particolare, alle esigenze di armonizzazione e di standardizzazione, divenute sempre più stringenti per garantire la comparabilità dei dati, specie in ambito europeo; dall'altra (semiasse negativo) si individuano invece quelle forme che richiamano le fasi operative della costruzione del dato. I termini che caratterizzano maggiormente il semiasse positivo sono: <sistema di indagini>, <contenuti>, <strumenti metodologici>, <armonizzazione strumenti metodologici>, <ambito europeo>, <metodologie statistiche>, <attività statistica>, <progettazione e realizzazione>, <misurazione>, <standardiz-

zazione>, <mining>, <sistema integrato>, <metadati>. I termini che contribuiscono a dare un senso al semiasse negativo sono invece: <indagine/i>, <campione/i>, <sistemi informativi statistici>, <interviste>, <processi di produzione>, <studio progettuale>, <geografico/che>, <dati elementari>, <quantificazione>, <codifica automatica>, <trattamento dei dati>, <strategia/e>. I trienni che presentano i maggiori contributi assoluti per questa dimensione sono, da un lato, il 2000-2002 e il 2001-2003 e, dall'altro, il 2004-2006 e il 2003-2005.

3.3. La mappa delle tematiche

L'analisi delle corrispondenze applicata alla matrice dei termini riferiti alle tematiche presenti nei PSN ha guidato l'esplorazione attraverso i temi trattati dalla statistica ufficiale nel corso dell'ultimo decennio.

In questo caso (Fig. 3), i trienni dal 2000-2002 al 2004-2006 si trovano posizionati nella parte sinistra del grafico (2000-2002 e 2001-2003 nel III quadrante e 2002-2004, 2003-2005 e 2004-2006 nel IV), mentre i restanti trienni sono localizzati in quella destra (2005-2007 e 2006-2008 nel I quadrante, mentre 2007-2009 e 2008-2010 nel II). Il triennio 2005-2007 mostra, rispetto alla precedente analisi, una maggiore similarità con i profili degli anni successivi piuttosto che con quelli degli anni precedenti.

In particolare sono stati individuati due fattori rappresentati in Fig. 3, che spiegano il 68,19% della variabilità complessiva.

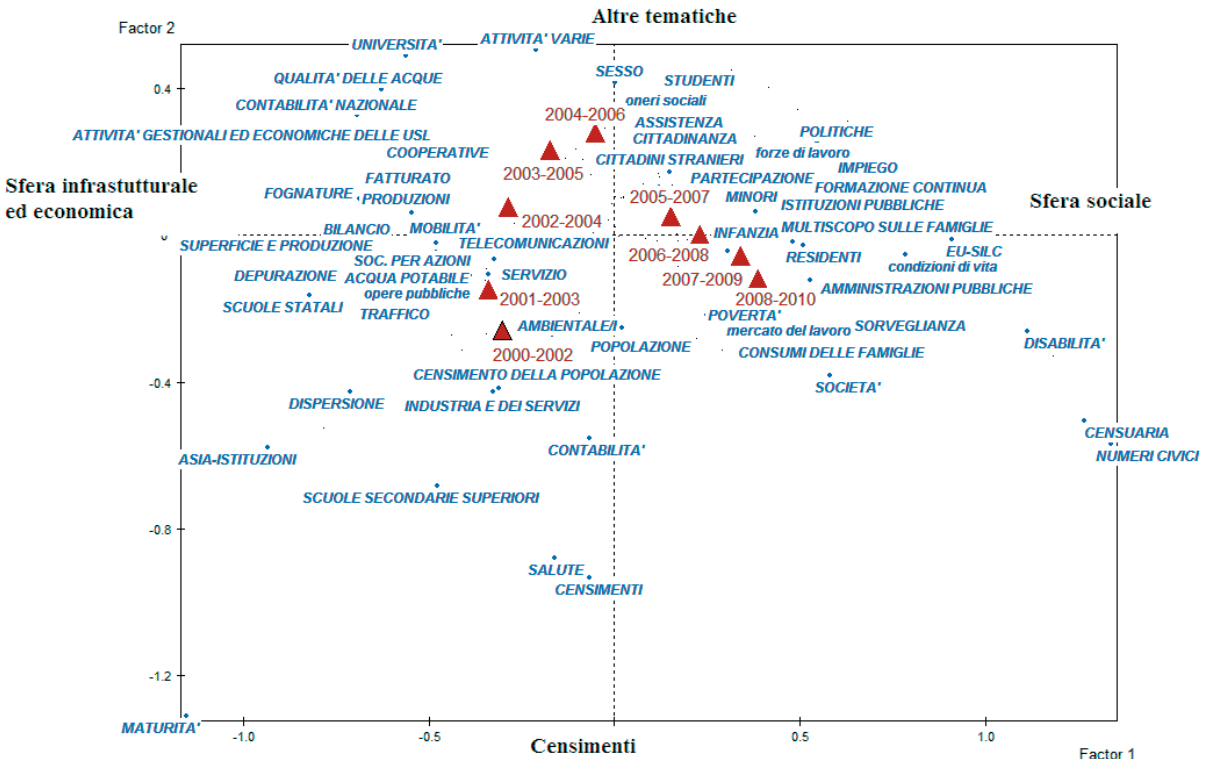

Figura 3: La mappa delle tematiche della statistica ufficiale dal 2000-2002 al 2008-2010

Il primo asse descrive l'"evoluzione storica delle tematiche" inserite nei PSN dell'ultimo decennio. Si osserva chiaramente il progressivo avvicinamento da tematiche afferenti alla sfera infrastrutturale dei servizi alla collettività ed economica (semiasse negativo) verso tematiche tipiche dell'ambito sociale e della partecipazione sociale (semiasse positivo). Dal 2000-2002 al

2004-2006 infatti il *focus* della statistica ufficiale è incentrato su <maturità>, <Asia-istituzioni>, <attività gestionali ed economiche delle Usl>, <scuole statali>, <depurazione>, <superficie e produzione>, <dispersione> [13], <contabilità nazionale>, <fognature>, <bilancio>, <qualità delle acque>, <traffico>, <acqua potabile>, <fatturato>, <università>, <produzioni>, <società per azioni>, <mobilità>. Dal 2005-2007 al 2008-2010 i temi emergenti invece sono evocati dalle forme <numeri civici> e <censuaria> connesse alla progettazione dei Censimenti del 2010-2011; ma anche da termini e poliremiatiche che richiamano tematiche inequivocabilmente sociali: <disabilità>, <Eu-Silc>, <condizioni di vita>, <sorveglianza>, <impiego>, <politiche>, <amministrazioni pubbliche>, <residenti>, <multiscopo sulle famiglie>, <istituzioni pubbliche>, <formazione continua>, <forze di lavoro>, <consumi delle famiglie>, <minori>, <povertà>, <diritto>, <infanzia>. La sfera sociale sembra dunque caratterizzare nell'ultimo quinquennio, più che nel passato, i temi della statistica ufficiale.

Il secondo asse risulta invece polarizzare da un parte la programmazione delle attività censuarie attuata nei trienni 2000-2002 e 2008-2010 (semiasse negativo), dall'altra le altre tematiche presenti nel PSN.

4. Conclusioni

L'approccio lessicometrico adottato in questo contributo ha consentito di esplorare un corpus di 600 mila occorrenze, riferito a un arco temporale di un decennio, che può considerarsi rappresentativo del lessico utilizzato dalle fonti statistiche ufficiali in Italia.

Sul piano lessicale, l'analisi qui presentata, che dà conto soltanto dei primi risultati prodotti, ha evidenziato la settorialità del linguaggio, che emerge dalle forme più frequenti, da quelle peculiari, nonché dalle più ricorrenti forme idiomatiche che hanno fatto da tramite per l'analisi di tipo testuale.

Sul piano testuale, l'estrazione di forme semplici e composte incentrate sui "requisiti di qualità" dell'informazione statistica ha permesso – mediante un ordinamento di sub-corpora – di individuare l'importanza ad essi attribuita nel tempo dalle fonti statistiche ufficiali. Infine, un'analisi multidimensionale esplorativa compiuta su una parte del corpus costituito esclusivamente dai titoli dei progetti, ha consentito di delineare l'evoluzione temporale dell'orientamento della statistica ufficiale, sia sul piano metodologico che sul piano delle tematiche affrontate.

Questi primi risultati incoraggiano ulteriori e più approfondite analisi – anche su eventuali estrazioni di frammenti di più specifico interesse – finalizzate alla messa a punto di un lessico di frequenza della statistica ufficiale, utile a monitorare nel tempo, attraverso i cambiamenti del linguaggio e l'analisi del ciclo di vita delle parole (Bolasco and Canzonetti, 2005), le variazioni nei contenuti che rispondono alla domanda informativa proveniente dalla collettività.

Riferimenti

Bolasco S. (1996). Il lessico del discorso programmatico di governo in Zuliani, A. e Villone, M., editors, *L'attività dei governi della repubblica italiana (1948-1994)*, Bologna: il Mulino, pp. 163-349.

Bolasco S. (1999). *L'analisi multidimensionale dei dati*. Roma: Carocci.

Bolasco S. and Canzonetti A. (2005). Some insights into the evolution of 1990s' standard Italian using Text mining techniques and automatic categorization. In Vichi, M., Monari, P., Mignani, S. and

[13] Intesa come dispersione scolastica.

Montanari, A., editors, *New developments in classification and data analysis*, Berlin: Springer, pp. 293-302.

Commissione delle Comunità Europee (2005) *Indipendenza, integrità e responsabilità delle autorità statistiche nazionali e dell'autorità statistica comunitaria.* Bruxelles, 25.5.2005, COM(2005) 217 definitivo, http://www.istat.it/istat/codice_europeo_maggio2005.pdf.

Lebart L. and Salem A. (1988). *Analyse statistique des données textuelle*s. Paris : Dunod.

Morrone A. (1993). Alcuni criteri di valutazione della significatività dei segmenti ripetuti. In Anastex S.J., editor, *JADT93*, ENST-Telecom, Paris, pp. 445-53.

Muller C. (1977). *Principles et méthodes de statistique lexicale*. Paris: Hachette (ristampa 1992, Paris, Champion).

Salton G. and Buckely C. (1988). Term-weighting approach in automatic text retrieval. *Information Processing & Management*, 24(5): 513-523.

SISTAN-ISTAT (dal 2000-2002 al 2008-2010). *Programma Statistico Nazionale*, parti prima e seconda, Roma (www.sistan.it).

Tuzzi A. (2003). *L'analisi del contenuto. Introduzione ai metodi e alle tecniche di ricerca*. Roma: Carocci.