Tecniche di rilevazione: differenze Cawi-Cati per descrivere la professione[1]

Barbara Lorè, Francesca della Ratta*

*Istat – Istituto nazionale di Statistica, Rome, Italy

Abstract

Over the last decade the use of *mixed-mode* in social surveys has become more and more popular, both for reducing time and costs and for enhancing coverage. Meanwhile, internet penetration trough the population has been growing, making respondents more approachable all over the world at any time, so web self-administration mode (CAWI) seems to be a great opportunity to compensate the decreasing CAPI-CATI accessibility to respondents. However, data collected through different modes might be biased by different responding behaviors due to the characteristics of each of them, for instance the presence/absence of an interviewer.
The aim of this paper is to verify if there is a 'mode effect' (CAWI vs CATI) through the use of text mining techniques applied to data coming from the Italian survey on graduates' transition from education to employment, which is based on a sequential *mixed-mode* (CAWI-CATI). We focus our analysis on an open question asking respondents to name their occupation and to describe it in as many details as possible.

Riassunto

Nel corso dell'ultimo decennio le indagini sociali sulle famiglie hanno fatto sempre maggiore ricorso a tecniche miste di rilevazione dei dati, con il duplice obiettivo di contenere tempi e costi e di massimizzare il tasso di risposta. Contemporaneamente, la crescente diffusione di dispositivi informatici di comunicazione all'interno delle famiglie offre sempre di più la possibilità di raggiungere i rispondenti senza vincoli geografici o temporali.
Per questo motivo sta diventando sempre più interessante la prospettiva offerta dalla tecnica CAWI (*Computer assisted web interviewing*), in alternativa al più tradizionale ricorso alla combinazione CAPI-CATI che sempre di più rende difficile il contatto con i rispondenti. Tuttavia, le informazioni raccolte attraverso tecniche di rilevazione diverse rischiano di portare con sé differenze dovute alle tecniche di rilevazione, per esempio la presenza/assenza di un intervistatore.
In questo lavoro vengono utilizzate tecniche di analisi testuale per verificare se tra i dati rilevati in modalità CAWI e CATI vi sia un 'effetto tecnica'. I dati analizzati provengono dall'indagine Istat sull'inserimento professionale dei laureati, un'indagine basata su una tecnica di rilevazione mista sequenziale (CAWI - CATI). I testi analizzati provengono da una domanda aperta in cui si chiede al rispondente di descrivere in modo dettagliato la propria professione.

Key words: mixed mode technique; interviewer effect ; job description; open questions; Text Mining

1. Introduzione

All'inizio degli anni 2000 la possibilità di sviluppare questionari elettronici ha dato avvio a cambiamenti importanti nelle tecniche di rilevazione dei dati anche nell'ambito della statistica ufficiale. Le tecniche di rilevazione tradizionali di tipo *paper and pencil*, con o senza rilevatore, hanno gradualmente lasciato il posto a quelle *computer assisted* (faccia a faccia o tramite telefono) in grado di garantire sia livelli qualitativi più elevati, sia tempi di produzione dei dati più brevi. Nel 2004 l'impianto metodologico dell'indagine sulle Forze di Lavoro, una

[1] Il testo è frutto di un lavoro comune. Barbara Lorè ha redatto i paragrafi 1, 2 e 4.2, Francesca della Ratta i paragrafi 3 e 4.1. Le conclusioni sono a cura di entrambe.

delle indagini più importanti dell'Istituto italiano di statistica (Istat), è stato completamente ridisegnato, segnando l'inizio dell'epoca delle rilevazioni *paperless*. Tra le innovazioni di processo, la più saliente è stata il passaggio all'uso della tecnica mista (CAPI-CATI) per la rilevazione dei dati, in grado di ottimizzare sia i costi che i tempi. Da allora un sempre maggior numero di indagini sulle famiglie ha abbandonato la rilevazione dei dati basata su un'unica tecnica per passare al *mixed-mode*.

A distanza di un decennio, la crescente diffusione di dispositivi informatici di comunicazione all'interno delle famiglie offre sempre di più la possibilità di raggiungere i rispondenti senza vincoli geografici o temporali. La necessità di contrarre i costi e di minimizzare il disturbo statistico sui rispondenti sta progressivamente orientando la statistica ufficiale verso l'uso del web per la rilevazione dei dati. Tuttavia, il ricorso esclusivo della tecnica CAWI (*computer assisted web interviewing*) nell'ambito delle indagini sulle famiglie al momento sembra ragionevole solo in riferimento a target molto specifici[2]. Difatti, quando l'oggetto dell'osservazione è la popolazione generale, non è al momento realistico aspettarsi tassi di risposta equiparabili a quelli che si possono ottenere con l'uso di rilevatori professionali.

Per questo motivo sta diventando sempre più interessante la prospettiva offerta dalla tecnica CAWI in combinazione con una delle tecniche con rilevatore, in particolare la tecnica CATI (Janssen, 2006). L'uso combinato di CAWI e CATI consente al tempo stesso di minimizzare i costi (sono le due tecniche meno costose), di ridurre il disturbo statistico (il rispondente può compilare il questionario quando e dove preferisce se sceglie la modalità web e, in entrambi i casi, non deve accogliere in casa l'operatore), e massimizza il tasso di risposta permettendo di raggiungere anche i rispondenti momentaneamente non reperibili al proprio domicilio.

Tuttavia, le informazioni raccolte attraverso tecniche di rilevazione diverse rischiano di portare con sé differenze dovute alle tecniche stesse, prima di tutto la presenza/assenza di un intervistatore (Hope et al., 2014). Diventa, dunque, doveroso per il ricercatore verificare se le differenze contenute nei dati siano in parte addebitabili alle stesse oppure no (Revilla, 2010; Vannieuwenhuyze et al., 2010). Nel caso lo fossero, si renderebbero necessari accorgimenti metodologici in grado di rendere i due questionari *mode specific*, ovvero semanticamente equivalenti (Hox et al. 2015; Nicolaas et al., 2011), in modo da minimizzare il rischio di errore di misurazione e di autoselezione dei rispondenti. Questi due errori, se dovessero presentarsi congiuntamente, renderebbero azzardata l'operazione di integrazione dei dati.

In questo lavoro vengono utilizzate tecniche di analisi testuale per confrontare dati rilevati con entrambe le tecniche di rilevazione CAWI e CATI. L'obiettivo è quello di verificare, attraverso l'analisi delle differenze lessicali riscontrate nella descrizione della professione svolta dai rispondenti, la presenza o meno di un 'effetto tecnica'.

2. L'indagine sull'inserimento lavorativo dei laureati

Il campo di osservazione scelto è l'Indagine sull'inserimento professionale dei laureati condotta dall'Istituto Italiano di Statistica nel 2015. L'Istat realizza questa indagine fin dal 1989; da allora sono state utilizzate diverse tecniche di rilevazione, passando, man mano che nuove tecnologie divenivano disponibili, dall'autocompilazione *paper and pencil* alla tecnica CATI, fino all'autocompilazione via web. Fino all'edizione del 2011 è stata utilizzata

[2] L'Indagine sull'inserimento lavorativo dei dottori di ricerca condotta nel 2014 ha utilizzato come unica tecnica di rilevazione dei dati l'autosomministrazione via web, raggiungendo un tasso di risposta del 72%.

un'unica tecnica di rilevazione (l'autocompilazione cartacea fino al 1998 e la tecnica CATI a partire dal 2001), mentre l'edizione del 2015 ha visto due importanti innovazioni nella rilevazione dei dati: l'utilizzo della tecnica mista e il ricorso al web.

Se è pur vero che l'uso del *mixed-mode* nelle rilevazioni statistiche è iniziato circa un decennio fa, la scelta delle tecniche da utilizzare si è sempre orientata prevalentemente verso l'integrazione tra CAPI e CATI. In particolare, dopo l'esperienza del censimento della popolazione del 2011[3] questo è il primo caso in cui, nell'ambito delle rilevazioni sulle famiglie, si è fatto ricorso alla compilazione via web all'interno di una tecnica mista. In particolare, la buona penetrazione di internet in questa popolazione giovane e altamente scolarizzata ha fatto ritenere che una tecnica di rilevazione mista CAWI-CATI potesse essere la scelta migliore per massimizzare il tasso di risposta. Su un totale di 58.470 interviste, 36.072 (circa il 62%) provengono dalla modalità CAWI e 22.398 da quella CATI. Tra gli occupati il 59,7% ha scelto di rispondere sul web.

Il presente lavoro prende in analisi il quesito aperto in cui si chiede ai rispondenti occupati al momento dell'intervista di specificare il nome della propria professione descrivendo dettagliatamente in cosa consiste il proprio lavoro. Nella modalità CAWI si tratta di un testo "di prima mano", scritto direttamente dal rispondente, mentre nella modalità CATI la presenza del rilevatore, seppur ben formato, rischia di modificare il linguaggio utilizzato dall'intervistato. Il principale obiettivo di questo lavoro è di verificare se esistono differenze significative nei testi provenienti dalle due tecniche di rilevazione e se queste differenze siano imputabili alla tecnica utilizzata.

Riguardo la professione svolta, sembra esserci un'autoselezione dei rispondenti rispetto alle due tecniche di rilevazione (Tab.1). In particolare ciò accade per le professioni della conoscenza (2° e 3° Grande Gruppo della classificazione[4]). Infatti, mentre tra chi svolge professioni intellettuali, scientifiche e di elevata specializzazione (ovvero le professioni del 2° grande gruppo, quelle regolamentate[5] specialistiche, i ricercatori, gli insegnanti, ecc.) si registra una preferenza per la tecnica CAWI (scelta dal 65,2% degli specialisti in confronto al 59,7% del totale degli intervistati), un pattern di scelta contrario si osserva tra coloro che svolgono professioni tecniche (come infermieri, fisioterapisti, geometri, ragionieri e tutte le professioni regolamentate denominate *"junior"*). Negli altri Grandi Gruppi, invece, non si osservano differenze rilevanti.

Analizzando le preferenze dei rispondenti del 2° e del 3° Grande Gruppo (quelli peraltro più numerosi) una spiccata propensione a rispondere via web sembra caratterizzare gli specialisti in scienze matematiche, informatiche, chimiche, fisiche e naturali e gli ingegneri e architetti (Tab.2). Tale propensione non si riscontra invece tra coloro che svolgono professioni tecniche

[3] L'ultimo appuntamento censuario ha dato al rispondente la possibilità di scegliere se compilare il questionario in autonomia (on line o su modello cartaceo) o se richiedere l'assistenza alla compilazione presso i centri di raccorlta allestiti presso i comuni. In una successiva sperimentazione, propedeutica al Censimento permanente della popolazione, è stata utilizzata anche la compilazione assistita da operatore telefonico.

[4] I Grandi Gruppi, il primo livello classificatorio della Classificazione delle Professioni CP2011, sono nove: 1 - Legislatori, imprenditori e alta dirigenza, 2 - Professioni intellettuali, scientifiche e di elevata specializzazione, 3-Professioni tecniche, 4 – Professioni esecutive nel lavoro d'ufficio, 5 - Professioni qualificate nelle attività commerciali e nei servizi, 6-Artigiani, operai specializzati e agricoltori, 7 - Conduttori di impianti, operai di macchinari fissi e mobili e conducenti di veicoli, 8 - Professioni non qualificate, 9 - Forze armate.

[5] Le professioni regolamentate sono quelle per le quali esiste una normativa che ne definisce il nome, le mansioni e i criteri di accesso. Un ordine professionale è deputato alla detenzione e aggiornamento di un albo professionale.

in campo scientifico, ingegneristico e della produzione, ovvero quelle figure professionali che condividono lo stesso ambito di competenze, ma alle quali è richiesto un livello di competenza leggermente più basso.

Tab. 1 – Intervistati occupati per tecnica scelta e Grande Gruppo professionale *(valori assoluti e percentuali)*

GRANDI GRUPPI	Valori assoluti			Valori percentuali		
	CAWI	CATI	Totale	CAWI	CATI	Totale
1- Legislatori, imprenditori e alta dirigenza	301	258	559	53,8	46,2	100
2- Professioni intellettuali, scientifiche e di elevata specializzazione	10.700	5.707	16.407	65,2	34,8	100
3- Professioni tecniche	8.495	7.262	15.757	53,9	46,1	100
4- Professioni esecutive nel lavoro d'ufficio	2.988	1.833	4.821	62,0	38,0	100
5- Professioni qualificate nelle attività commerciali e nei servizi	1.638	1.228	2.866	57,2	42,8	100
6- Artigiani, operai specializzati e agricoltori	174	146	320	54,4	45,6	100
7- Conduttori di impianti, operai di macchinari fissi e mobili e conducenti di veicoli	56	63	119	47,1	52,9	100
8- Professioni non qualificate	152	60	212	71,7	28,3	100
9- Forze armate	283	165	448	63,2	36,8	100
Totale	**24.787**	**16.722**	**41.509**	**59,7**	**40,3**	**100**

Una propensione più moderata per la tecnica CAWI si osserva anche tra gli Specialisti in scienze umane, sociali, artistiche e gestionali e tra gli Specialisti della formazione e della ricerca, mentre tutti i Gruppi delle Professioni Tecniche sembrano preferire l'intervista telefonica. Questo risultato sembra suggerire che la scelta della tecnica di rilevazione da parte dei rispondenti non dipenda tanto dall'ambito disciplinare e/o professionale in cui si opera, quanto invece dal livello di competenza richiesto per svolgere la professione e conseguentemente dal tipo di laurea conseguito. Le professioni classificate nel 2° Grande Gruppo, infatti, per essere svolte richiedono una laurea specialistica o una laurea a ciclo unico, mentre le professioni tecniche, classificate nel 3° Grande Gruppo, possono essere svolte da chi ha conseguito un diploma di scuola superiore o una laurea triennale. La relazione tra il tipo di laurea (triennale, specialistica o a ciclo unico) e la tecnica scelta per rispondere al questionario (CAWI - CATI) è confermata dal test del χ^2, che evidenzia un'associazione altamente significativa (Chi^2=89.19482, $p \leq 0.0001$)[6].

[6] Il test del χ^2 è stato eseguito su tutto il campione dei laureati

Tab. 2 – Dettaglio professioni specialistiche e tecniche per tecnica *(valori assoluti e percentuali)*

GRUPPI	Valori assoluti			Valori percentuali		
	CAWI	CATI	Totale	CAWI	CATI	Totale
2.1- Specialisti in scienze matematiche, informatiche, chimiche, fisiche e naturali	1.362	437	1.799	75,7	24,3	100
2.2- Ingegneri, architetti e professioni assimilate	2.668	943	3.611	73,9	26,1	100
2.3- Specialisti nelle scienze della vita	590	426	1.016	58,1	41,9	100
2.4- Specialisti della salute	223	233	456	48,9	51,1	100
2.5- Specialisti in scienze umane, sociali, artistiche e gestionali	3.098	2.004	5.102	60,7	39,3	100
2.6- Specialisti della formazione e della ricerca	2.553	1.664	4.217	60,5	39,5	100
3.1- Professioni tecniche in campo scientifico, ingegneristico e della produzione	1.717	1.534	3.251	52,8	47,2	100
3.2- Professioni tecniche nelle scienze della salute e della vita	2.989	2.719	5.708	52,4	47,6	100
3.3- Professioni tecniche nell'organizzazione, amministrazione e nelle attività finanziarie e commerciali	2.274	2.010	4.284	53,1	46,9	100
3.4- Professioni tecniche nei servizi pubblici e alle persone	1.227	999	2.226	55,1	44,9	100
Totale	**18.701**	**12.969**	**31.670**	**59,0**	**41,0**	**100**

Il testo inserito nel campo aperto del quesito sulla professione costituisce la base informativa per la successiva attribuzione di un codice, scelto tra quelli disponibili nella classificazione CP2011. La codifica viene fatta al livello di massimo dettaglio della classificazione (V digit), quello delle Unità Professionali. Nella modalità CAWI viene chiesto al rispondente di eseguire questa operazione, mentre nel CATI è il rilevatore che ha questo compito. Spetta poi al ricercatore valutare l'accuratezza della codifica in entrambi i casi. L'analisi del testo inserito per descrivere la professione consentirà di valutare eventuali differenze linguistiche tra le due tecniche e individuare un eventuale effetto tecnica.

3. Il corpus in analisi

Il corpus che include l'insieme delle descrizioni delle professioni svolte dagli intervistati è composto da 338.378 occorrenze e un vocabolario di 23.499 forme, con una ricchezza lessicale medio bassa (6,9), tipica delle risposte a domanda aperta.

Tra le parole più frequenti incontriamo sia nomi di professioni (*ingegnere, tecnico, consulente, insegnante*) sia termini utilizzati per descrivere il luogo o contesto in cui viene svolta la professione (*presso, azienda, gestione, lavoro*). Tra le parole chiave, ottenute confrontando il vocabolario con il lessico di frequenza dell'italiano standard presente in Taltac2 (Bolasco, 2013), spiccano i nomi di alcune professioni, quali *infermiera, educatrice, consulente, fisioterapista, impiegata, analista, programmatore, commercialista* ecc.

L'elevato carattere descrittivo del testo è peraltro evidente se si analizzano alcune informazioni sulla frequenza delle occorrenze delle singole forme grammaticali, che si ottiene con il programma Taltac2, che fornisce il conteggio di tutte le categorie non ambigue (imprinting grammaticale, Tab. 3): se si analizzano esclusivamente le forme non ambigue, emerge una netta prevalenza di sostantivi e preposizioni (il 26,2 e 20% del totale) e un impiego meno diffuso di verbi e aggettivi. La prevalenza di sostantivi è ancora più accentuata nel testo trascritto dai rilevatori Cati, che probabilmente hanno scritto in modo più stringato mirando alla sola descrizione della professione, con un impiego meno diffuso del linguaggio più strumentale (preposizioni, congiunzioni, avverbi, pronomi).

Tab. 3 Imprinting grammaticale per tecnica e confronto con lessico italiano standard *(scritto-parlato; valori assoluti e percentuali)*

Categoria grammaticale	Totale occorrenze						
	v.a.			%			
	tot. occ.	CAWI	CATI	tot. occ.	CAWI	CATI	% Italiano standard (scritto - parlato)
N	64.620	47.322	17.298	26,2	25,2	29,4	25,3
PREP	51.255	40.530	10.725	20,8	21,6	18,2	14,3
CONG	13.263	10.640	2.623	5,4	5,7	4,5	7,0
A	11.579	9.154	2.425	4,7	4,9	4,1	1,6
V	11.204	8.063	3.141	4,6	4,3	5,3	33,7
AVV	1.703	1.495	208	0,7	0,8	0,4	3,5
PRON	1.489	1.334	155	0,6	0,7	0,3	4,0
altre cat	10.070	8.527	1.543	4,1	4,5	2,6	10,6
tot. Forme non ambigue	**246.223**	**187.438**	**58.785**	**100,0**	**100,0**	**100,0**	**100,0**

Confrontando questa distribuzione con le statistiche ottenute sulle forme non ambigue dell'italiano standard (Bolasco, 2013), la prevalenza di sostantivi e preposizioni rende il testo più vicino a quello dell'italiano scritto, in particolare lo scritto-parlato, tranne che per una incidenza decisamente più bassa di verbi (4,6% in confronto al 33,7). Altre forme poco utilizzate nelle risposte sono gli avverbi e i pronomi.

L'estrazione dei segmenti ripetuti, vale a dire le combinazioni di parole che si presentano con la stessa sequenza nel testo, rende evidente la necessità di procedere all'unificazione in un'unica forma grafica dei nomi di professioni composte da più vocaboli (es. *ingegnere meccanico* e *ingegnere informatico*, o *insegnante di scuola media* e *insegnante di scuola secondaria*), attraverso procedura di lessicalizzazione. Tale procedura è stata condotta utilizzando sia la lista delle voci professionali[7] del dizionario delle professioni, sia una lista estratta dai segmenti ripetuti contenente la descrizione di specifiche professioni o il

[7] Il programma Taltac2 consente di "lessicalizzare" in un'unica forma grafica più parole espresse in sequenza attraverso l'aggiunta del simbolo "_". La lessicalizzazione può essere ottenuta automaticamente a partire da liste di parole in sequenza predisposte dall'utente. E' possibile quindi trasformare in un'unica forma sia una selezione di segmenti ripetuti effettivamente rinvenuti nel testo e selezionati dall'utente, sia sottoporre al programma una lista di parole già predisposte, come nel caso dell'elenco delle voci professionali, un elenco che contiene circa 6.000 varianti linguistiche con cui vengono denominate le professioni, inserito nella classificazione delle professioni (cfr. Istat, 2013).

riferimento al contesto lavorativo (Tab. 4). Tale operazione, unificando in un'unica entrata più termini, riduce naturalmente il numero di occorrenze (scese a 313.276) e fa incrementare il numero di forme grafiche (24.456) e la ricchezza lessicale (7,8).

Tabella 4 - Segmenti ripetuti più frequenti utilizzati per la lessicalizzazione del corpus

Voci professionali		Contesto	
Segmento	Occ.	Segmento	Occ.
infermiera professionale	423	scuola primaria	484
ingegnere civile	384	studio legale	257
impiegata amministrativa	329	scuola secondaria superiore	223
ingegnere meccanico	280	call center	214
insegnante di sostegno	250	controllo qualità	208
tecnico di laboratorio	241	educazione fisica	175
ingegnere edile	223	back office	158
dottore commercialista	209	scuola dell' infanzia	128
ingegnere informatico	196	cooperativa sociale	114
educatore professionale	194	asilo nido	113
assistente sociale	193	azienda ospedaliera	113
analista programmatore	190	laboratorio biomedico	113
impiegato amministrativo	182	sicurezza sul lavoro	110
tecnico sanitario	181	ufficio acquisti	93
project manager	166	studio professionale	86
personal trainer	159	scuola secondaria di secondo grado	84
insegnante scuola primaria	149	consulenza informatica	83
assegnista di ricerca	148	assistenza clienti	72
ingegnere elettronico	147	recupero crediti	70
praticante avvocato	136	gestione del personale	68
commessa di negozio	134	università degli studi	68
tecnico sanitario di radiologia medica	131	polizia locale	67
ingegnere gestionale	128	web marketing	67

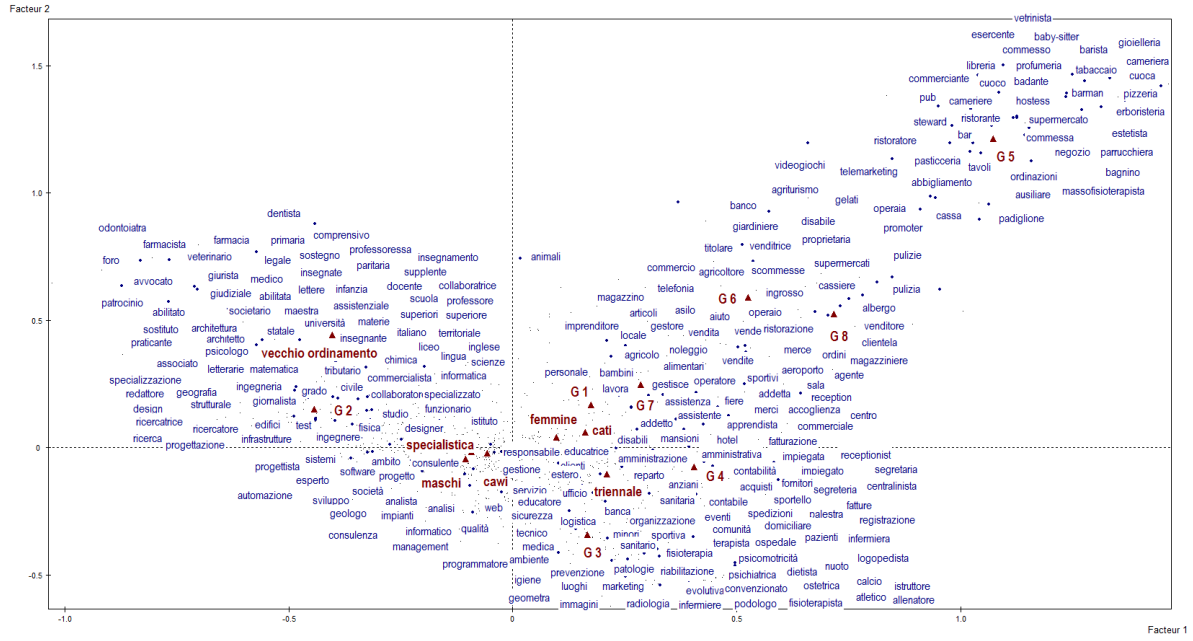
Una rappresentazione sintetica dei contenuti del testo può essere ottenuta con l'analisi delle corrispondenze semplici su tabella lessicale, considerando la distribuzione del vocabolario[8] per Grande Gruppo professionale, il tipo di laurea (distinta tra vecchio ordinamento, specialistica o triennale), il sesso degli intervistati e la tecnica scelta per rispondere.

Le variabili più influenti nell'analisi sono il Grande Gruppo professionale (soprattutto il 2°, il 3° e il 5°) e il tipo di laurea, mentre il tipo di tecnica diviene rilevante soltanto nel quinto fattore (inerzia riprodotta=9,5%). Il primo fattore (inerzia riprodotta=20,7%) riproduce soprattutto l'opposizione tra le professioni specialistiche del 2° Grande Gruppo (associate soprattutto a una laurea del vecchio ordinamento) e quelle del 5°, mentre il secondo fattore

[8] Per l'analisi delle corrispondenze semplici è stata utilizzata una tabella lessicale forme*testi, con in riga una selezione del vocabolario composto dalle parole chiave con scarto standardizzato>4, cui sono state aggiunte tutte le forme lessicalizzate, altamente caratterizzanti il testo e in colonna la ripartizione delle stesse parole per Grande gruppo professionale, tipo di laurea, tipo di tecnica e sesso degli intervistati. Tra i grandi gruppi professionali, quello delle forze armate è stato escluso dall'analisi per la quantità troppo bassa di unità di contesto facenti parte di quel gruppo.

(inerzia=14,4%) si caratterizza per l'opposizione tra le professioni del 5° e quelle del 3° Grande Gruppo, associate soprattutto alla laurea triennale. Il grafico fattoriale offre una sintesi dei diversi termini utilizzati per descrivere le professioni, con una distinzione netta tra la terminologia utilizzata per descrivere le professioni qualificate nei servizi alle persone (5° Grande Gruppo, semiasse positivo) e quelle più specialistiche del 2° Grande Gruppo (semiasse negativo). La laurea specialistica si colloca in posizione intermedia tra quelle meno diffuse del vecchio ordinamento e quelle triennali, con una maggiore associazione con le professioni specialistiche del 2° Grande Gruppo e con la modalità di compilazione CAWI. Si osserva inoltre una associazione tra sesso e tecnica, con una preferenza dei maschi per il CAWI.

Figura 1 – Piano fattoriale: distribuzione del vocabolario per Grande Gruppo, tipo di laurea, sesso e tecnica di rilevazione



Oltre che descrivere il contenuto del corpus e le differenze di linguaggio utili a caratterizzare le diverse professioni, l'analisi delle corrispondenze evidenzia che l'impatto del tipo di tecnica sul testo in analisi è piuttosto ridotto, e comunque subordinato ad altre variabili. Per valutare l'impatto della tecnica di risposta sulle caratteristiche del testo, nel paragrafo successivo verrà effettuato il calcolo delle parole caratteristiche (Tuzzi, 2003) per individuare le differenze di linguaggio tra i rilevatori e coloro che hanno compilato direttamente il questionario sul web.

4. Analisi delle specificità

4.1 Differenze di linguaggio per tecnica

Obiettivo primario del presente lavoro è quello di studiare le differenze di linguaggio per tecnica di rilevazione. Per questo obiettivo sono state estratte le parole caratteristiche, vale a dire quelle che sono utilizzate soprattutto in una specifica parte del testo. La prima differenza che emerge è che i testi inseriti direttamente dagli intervistati sono più ricchi di termini

strumentali (*presso, sono, mi occupo, attualmente, lavoro, consiste, attività*) utilizzati per articolare il discorso, mentre i rilevatori inseriscono direttamente la descrizione della professione, facendo meno attenzione all'articolazione della frase. Come emerso già dall'analisi dell'imprinting, sembra quindi che chi compila autonomamente il testo sul web cerchi maggiormente di rispettare la "leggibilità" del testo. Si tratta di una differenza interessante, che tuttavia non impatta in modo significativo sulla qualità dei dati inseriti, fornendo una semplice notazione stilistica tra i due modi di inserire le informazioni, anche in considerazione della specifica formazione ricevuta dai rilevatori, addestrati a inserire le informazioni utili a una corretta codifica della professione.

Tale differenza si accentua se si analizzano le parole caratteristiche ottenute sul corpus lessicalizzato: i termini caratteristici di chi ha riempito direttamente il campo aperto sono soprattutto quelli utili a descrivere la propria attività (*lavoro presso, in azienda, mi occupo, attualmente, ruolo, consiste*) e il contesto in cui si svolge (*Milano, Roma, sede, studio, società multinazionale, Londra, università degli studi, azienda, scuola secondaria di primo grado, cooperativa sociale,* ecc.). Tra le professioni citate da chi ha compilato il questionario in rete incontriamo soprattutto *libero professionista*, *engineer, docente, analyst, business analyst, project manager, dottore commercialista*. Di contro, i testi inseriti dai rilevatori sono ricchi di riferimenti a denominazioni di professioni della classificazione, in particolare *infermiera, insegnante, avvocato, fisioterapista, contabile, educatrice, commercialista, commessa, avvocato civile, impiegata amministrativa*.

Tale differenza persiste se si effettua l'analisi della specificità sui sub corpus distinti che raccolgono solo le interviste codificate nel secondo o terzo Grande Gruppo: nella tabella 5 sono presentate le parole caratteristiche utilizzate per descrivere le professioni intellettuali e tecniche, distinte per tecnica di rilevazione CATI-CAWI, ordinate per valori crescenti del p-value e con l'indicazione delle occorrenze normalizzate nel testo complessivo e nella specifica parte in esame, in modo da apprezzare la misura della sovra-rappresentazione alla base del modello ipergeometrico utilizzato per il calcolo delle specificità (Tuzzi, 2003). Tra le risposte di chi svolge una professione specialistica e ha compilato il questionario sul web si nota, oltre alla prevalenza dei termini strumentali già riscontrata in precedenza, un maggiore impiego di termini inglesi e di parole che contestualizzano la professione, come *business analyst* o *project manager.* Di contro nel testo inserito dai rilevatori sono più diffuse le forme composte da noi lessicalizzate, e termini quali i*nfermiera professionale, insegnante di scuola primaria, insegnante di sostegno.*

In generale, i rispondenti che hanno scelto la tecnica CAWI fanno uso di un linguaggio più tecnico-specialistico, con una frequenza maggiore del ricorso a descrizioni e nomi ufficiali delle professioni (per es. "dottore commercialista" invece che semplicemente "commercialista").

Tab.5 Parole caratteristiche per tecnica di rilevazione e Grande Gruppo professionale *(ordinate per valori del p-value crescenti)*

	CAWI			CATI		
Grande Gruppo	Forma grafica	Occorrenze normalizzate (*10000)	Sub-occorrenze normalizzate (*10000)	Forma grafica	Occorrenze normalizzate (*10000)	Sub-occorrenze normalizzate (*10000)
Professioni intellettuali, scientifiche e di elevata specializzazione	presso	193.4	227.6	insegnante	67.9	198.4
	in_azienda	33.1	40.4	avvocato	56.6	147.1
	svolgo	19.9	24.4	farmacista	29.6	70.2
	libero_professionista	18.5	22.6	commercialista	14.9	48.9
	engineer	17.8	21.5	insegna	2.4	12.7
	business_analyst	4.3	5.3	educatrice	11.7	33.7
	dottore_commercialista	15.0	16.8	gestisce	3.9	16.4
	project_manager	15.5	17.2	veterinario	3.9	16.0
Professioni tecniche	presso	170.0	249.8	infermiera	75.8	176.3
	svolgo	14.4	22.2	lavora	5.5	18.4
	in_azienda	17.5	26.1	fisioterapista	51.3	93.7
	milano	8.5	13.2	consulente	42.6	80.2
	libero_professionista	6.4	9.8	contabile	20.0	43.5
	consiste	6.9	10.3	educatrice	17.0	34.9
	engineer	3.6	5.5	ingegnere	15.4	31.2
	automotive	2.0	3.2	consulente_finanziario	7.8	18.7

4.2 Differenze di età, sesso ed area disciplinare

Interessanti anche le differenze di linguaggio tra rispondenti di età diverse. Nel linguaggio dei più giovani si fa più spesso riferimento a professioni per svolgere le quali è richiesta la laurea triennale (*fisioterapista, ostetrica, infermiere*) o a posizioni professionali tipiche di chi ha da poco iniziato la propria carriera (*apprendista, stagista*). All'aumentare dell'età si osserva anche un cambiamento nel linguaggio che probabilmente riflette anche lo sviluppo della carriera dei rispondenti. A partire dai 30 anni diventano più frequenti termini che fanno riferimento a professioni che richiedono una laurea quinquennale (*architetto, ingegnere civile, medico veterinario*) o a ruoli direttivi (*direzione lavori*). Solo dopo i 45 anni il linguaggio dei rispondenti testimonia il raggiungimento di posizioni dirigenziali consolidate (*funzionario, Ufficiale*, ecc.). Questa evoluzione del linguaggio nelle diverse fasce di età si presenta in modo pressoché identico nelle due tecniche di rilevazione.

Rispetto all'area disciplinare della laurea conseguita, le differenze linguistiche sono coerenti con i temi e gli argomenti caratteristici del titolo di studio. Così, per esempio, nel lessico di chi ha conseguito una laurea del gruppo agrario sono presenti chiari riferimenti al mondo animale e alle coltivazioni (*medico veterinario, dottore agronomo, allevamento, azienda agricola*), mentre chi ha conseguito una laurea del gruppo dell'insegnamento fa spesso riferimento al contesto scolastico (*scuola primaria, insegnante di sostegno, asilo nido*). Analizzato rispetto al sesso dei rispondenti, il testo riflette sia le differenze di genere (*infermiere/infermiera*, *imprenditore/imprenditrice*), sia la diversa distribuzione delle

professioni tra i due sessi, con una prevalenza di termini riferiti all'insegnamento tra le donne e una maggiore frequenza di parole riferite all'ingegneria e all'ICT tra gli uomini.

Le differenze sin qui analizzate consentono in parte di ricostruire lo stile linguistico con cui si descrive la professione svolta (della Ratta et al., 2011)[9]: questo naturalmente dipende anche dal Grande Gruppo in cui è classificata la professione. In particolare, chi svolge una professione classificata nel 2° o nel 3° Grande Gruppo più spesso fa riferimento al nome della professione. Questa preferenza si spiega in parte con la presenza all'interno di questi due Grandi Gruppi della maggior parte delle professioni regolamentate, ovvero quelle occupazioni il cui nome è definito dalle leggi dello Stato. Una preferenza per lo stile linguistico che descrive il ruolo ricoperto all'interno dell'organizzazione si ritrova principalmente in coloro che svolgono una professione del 1° o del 9° Grande Gruppo, i quali fanno riferimento alla titolarità dell'impresa nel primo caso e al grado gerarchico nel secondo. I rispondenti del 5° e del 6° Grande Gruppo, infine, mostrano una preferenza sia per la citazione del nome della professione sia per la descrizione delle mansioni e delle attività svolte. Anche in questo caso l'aver scelto di rispondere al questionario via web o per telefono non ha introdotto differenze rilevanti nella descrizione della professione.

5. Conclusioni

Le analisi condotte nell'ambito di questo lavoro hanno evidenziato una ragionevole equivalenza tra le tecniche CAWI e CATI rispetto al modo di rispondere al quesito aperto sulla professione. Le differenze individuate riguardano principalmente le varianti stilistiche, mentre dal punto di vista del contenuto sono altre le variabili (età, Grande Gruppo professionale, area disciplinare, sesso) che entrano in gioco nel determinare delle differenze, che quindi non sembrano legate tanto ad errori di misurazione, e quindi alle tecniche di rilevazione utilizzate, quanto invece alla realtà dei fenomeni, in particolare all'età dei rispondenti.

Se risultati simili emergessero anche dall'analisi di variabili quantitative, sarebbe ragionevole un ripensamento sulle tecniche miste attualmente utilizzate nelle indagini sulle famiglie, con l'assegnazione di un ruolo maggiore alla tecnica CAWI in disegni misti sequenziali (CAWI – CATI – CAPI), nell'obiettivo di massimizzare il tasso di copertura e minimizzare i costi.

Inoltre, come già evidenziato dall'Indagine sull'inserimento lavorativo dei dottori di ricerca, l'uso del web per la rilevazione dei dati in target giovani e così altamente scolarizzati sembra essere un'opzione che andrebbe ancor più valorizzata, concentrando maggiormente le risorse in azioni di sollecito volte a favorire la risposta online.

[9] Gli stili linguistici identificati nel linguaggio dei lavoratori intervistati sono cinque: 1 - stile linguistico che fa riferimento al nome della professione, 2 – stile linguistico che descrive attività o mansioni, 3 – stile linguistico che fa riferimento a strumenti, macchinari o apparecchiature, 4 – stile linguistico che fa riferimento al ruolo ricoperto dall'individuo all'interno del processo produttivo, 5 – stile linguistico che fa riferimento al datore di lavoro, luogo o all'organizzazione in cui si lavora.

Riferimenti bibliografici

Buelens et al. (2012). Disentangling mode-specific selection and measurement bias in social surveys, *Discussion paper* (201211), Statistics Netherlands.

Bolasco S. (2013). *L'analisi automatica dei testi. Fare ricerca con il text mining*. Roma, Carocci.

de Leeuw E.(2005). To Mix or Not to Mix Data Collection Modes in Surveys, *Journal of Official Statistics*, Vol. 21 (No. 2): 233–255.

della Ratta Rinaldi F., Gallo F., Loré B. (2011). How do you name your occupation? A text mining application on the language used by workers and by the standard occupational classification. In: *Meeting of the Classification and Data Analysis Group of the Italian Statistical Society 2011.*

Hope S. et al. (2014). The Role of the Interviewer in Producing Mode Effects: Results from a Mixed Modes Experiment Comparing Face-to-Face, Telephone and Web Administration. Understanding Society Working Paper Series, No. 20141-20, ESRC, UK.

Hox et al., 2015; Measurement equivalence in mixed mode surveys, *Frontiers in Psychology*, Vol.6: 87.

Istat (2013). *Classificazione delle professioni*, Roma.

Janssen (2006). Web data collection in a mixed mode approach: an experiment, *Proceedings of Q2006 European Conference on Quality in Survey Statistics*

Nicolaas, G. et al. (2011). Is it a good idea to optimise question format for mode of data collection? Results from a mixed modes experiment. *ISER Working Paper Series*, No. 2011-31, ISER

Revilla. (2010). Quality in Unimode and *Mixed-mode* designs: A Multitrait-Multimethod approach, *Survey Research Methods*, Vol.4 (No.3): 151-164.

Vannieuwenhuyze et al.,2010; A method for evaluating mode effects in *mixed-mode* surveys, *Public Opinion Quarterly*, Vol. 74 (No. 5): 1027–1045.

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

Carocci.