Strategie di Text Mining per il controllo e la correzione della codifica dell'attività economica nell'indagine Istat sulle forze di lavoro*

Francesca della Ratta-Rinaldi[1], Mauro Tibaldi[1], Maria Elena Pontecorvo[1]

[1] ISTAT, Istituto Nazionale di Statistica, Dipartimento per le Statistiche Sociali ed Ambientali; dellarat@istat.it; tibaldi@istat.it, mariaelena.pontecorvo@istat.it

Abstract

To improve the quality of data, the Istat Division "Education, Training and Labour" experienced a Text Mining procedure to check and correct the coding of economic activity carried out by interviewers in the Labour Force Survey. The strategy is made possible by the presence in the data file of a text field in which the interviewers wrote the economic activity's features and the occupation as described by respondents, matched by the code of economic activity (and occupation). The procedure has been active since the fourth quarter of 2011, and it's based on a comparison between respondents' vocabulary and the specific dictionary of each official classification division (Ateco 2007 classification, originated from Nace rev. 2). The subset of words used by interviewers and not in the official dictionary is useful for the errors detection. Using the "Research by regular expression", a technique contained in the software Taltac2, you can indicate all records which contain these words: once established that the presence of a particular word or combination of word makes it possible to identify an error code assignment, it's possible to indicate the correct code to put in the dataset. In the case of recurring errors the queries can be saved to be used in future sessions of correction. The process is completed with a thorough examination of data consistency in each session, to validate the corrections made and to assign the definitive proper code.

Riassunto

Per migliorare la qualità dei dati il Servizio Istruzione, formazione e lavoro dell'Istat ha messo in campo una procedura di Text Mining per il controllo e correzione della codifica dell'attività economica effettuata dai rilevatori nell'indagine sulle Forze di lavoro. La strategia è resa possibile grazie alla presenza nel file dati di un campo testuale in cui i rilevatori riportano le caratteristiche dell'attività economica e della professione svolta, così come descritta dagli intervistati, accompagnato dalla codifica relativa all'attività economica. La procedura, attiva dal quarto trimestre 2011, si basa sul confronto tra il vocabolario adottato dai rispondenti e il dizionario specifico di ciascuna divisione, desunto dalla classificazione ufficiale (classificazione Ateco 2007, Nace rev. 2). Ai fini dell'individuazione degli errori è di particolare interesse il sottoinsieme composto dalle parole riportate dai rilevatori e non presenti nella divisione specifica del dizionario ufficiale. A partire da questi termini è possibile rintracciare tutti i record che li contengono attraverso la procedura di Ricerca Entità del software Taltac2: una volta appurato che la presenza di un determinato termine o combinazione di termini permette di individuare un errore di attribuzione del codice, è possibile indicare il codice corretto da aggiungere al file di partenza. In caso di errori ricorrenti, l'istruzione di ricerca viene eseguita automaticamente nelle future sessioni di correzione, in modo da reiterare in automatico parte delle correzioni. Il processo di correzione, tuttavia, si completa attraverso una verifica puntuale della coerenza dei dati; in ogni sessione, difatti, è sempre necessario un intervento manuale per la validazione delle correzioni effettuate e per la definitiva attribuzione del codice corretto.

Keywords: text mining, data quality, data checking, statistical classifications, Labour Force Survey

* Il testo è frutto di un lavoro comune. In particolare Francesca della Ratta ha redatto i paragrafi 2, 6 e 7, Mauro Tibaldi i paragrafi 1, 3 e 4 e Maria Elena Pontecorvo il paragrafo 5.

1. Introduzione

Il processo delle indagini della statistica ufficiale è continuamente investito da interventi mirati a migliorare l'accuratezza del dato, specie nel campo del contenimento degli errori non campionari che, come noto, sono legati alle procedure di misurazione e possono insorgere ad ogni passo del processo di produzione delle informazioni (Istat, 2011). In particolare, gli errori di misurazione assumono un grande rilievo in quanto l'informazione è disponibile ma non è corretta. Pertanto, la fase dei "controlli a caldo" eseguita nel corso del trattamento dei dati statistici può essere intesa come un insieme di azioni predisposte per individuare gli errori che insorgono durante il processo di produzione.

In questo quadro, a partire dal 2011 il Servizio Istruzione, formazione e lavoro dell'Istat ha messo in campo una procedura di Text Mining per il controllo e correzione della codifica dell'attività economica effettuata dai rilevatori nell'indagine sulle Forze di lavoro.

Tale procedura è stata progettata in seguito all'introduzione nel 2011 della nuova classificazione delle attività economiche Ateco 2007 (Nace Rev. 2) nella rilevazione sulle Forze di lavoro (RFL). Considerati i vincoli di tempestività (i dati sono rilasciati a 60 giorni dalla conclusione del trimestre) e l'elevata numerosità campionaria (circa 150 mila record individuali a trimestre) dell'indagine, si è deciso di mettere in piedi una procedura selettiva finalizzata alla correzione dei codici attribuiti alle divisioni di attività economica caratterizzate da un più elevato tasso di errore.

La strategia di controllo è resa possibile grazie alla presenza nel file dati di un campo testuale nel quale i rilevatori riportano le caratteristiche dell'attività economica e della professione svolta, così come descritta dagli intervistati, accompagnato dalla codifica relativa all'attività economica svolta nella sede in cui lavora l'intervistato. La procedura si basa sull'analisi della congruenza tra il linguaggio del dizionario ufficiale della classificazione e il linguaggio dei rispondenti.

In questo lavoro viene innanzitutto descritta la procedura seguita per l'individuazione degli errori e la loro correzione (par. 2), che ha messo in luce da un lato un problema legato alla qualità delle descrizioni inserite dai rilevatori (par. 3) e dall'altro la ricorrenza di alcuni errori sistematici nella codifica (par. 4). Un problema più generale emerge poi se si considera la distanza tra il linguaggio degli intervistati e quello delle classificazioni ufficiali, elemento questo che può in parte contribuire al problema dell'errata codifica (par. 5). Infine, viene descritto un test effettuato per verificare la possibilità di utilizzare un software specifico per il controllo e la correzione automatica dei dati. I risultati hanno evidenziato che gli errori, determinati in gran parte da descrizioni non sufficientemente articolate, si concentrano in alcuni settori di attività economica. Alla loro genesi contribuisce probabilmente anche l'inadeguatezza del dizionario ufficiale come unico strumento di codifica, a causa della sua distanza dal linguaggio comune degli intervistati.

2. La procedura di controllo "Dizionari"

Nella fase di avvio della sperimentazione, alcuni controlli a campione sulle codifiche dei rilevatori hanno permesso di individuare nell'Agricoltura, nelle Costruzioni, nell'Amministrazione pubblica, difesa e assicurazione sociale (d'ora in avanti denominata servizi generali della PA), nell'Istruzione e nella Sanità e assistenza sociale i settori affetti dal maggior tasso di errore su cui procedere con il nuovo sistema di controllo.

Appurata l'esistenza di un certo quantitativo di record mal codificati dai rilevatori, per superare la gravosità connessa alla fase di revisione manuale, dal IV trimestre 2011 si è provveduto a implementare una procedura semi automatica per l'individuazione e la correzione degli errori (tuttora in corso), a partire dalla stringa testuale inserita dai rilevatori per la descrizione sia dell'attività economica dell'ente/azienda presso cui viene prestata l'attività lavorativa (d'ora in avanti Ateco), sia della professione svolta dall'intervistato. Nonostante l'oggetto della sperimentazione fosse limitato soltanto alla validazione della codifica Ateco, si è ritenuto indispensabile considerare congiuntamente i testi costituiti dall'insieme delle risposte alle due domande aperte presenti nel questionario[1] per individuare con maggiore precisione la codifica Ateco[2]. Le interviste codificate dai rilevatori in ciascuna delle cinque sezioni sono naturalmente analizzate in distinte sessioni di analisi, in modo da studiare esclusivamente il linguaggio utile a descrivere una specifica attività economica.

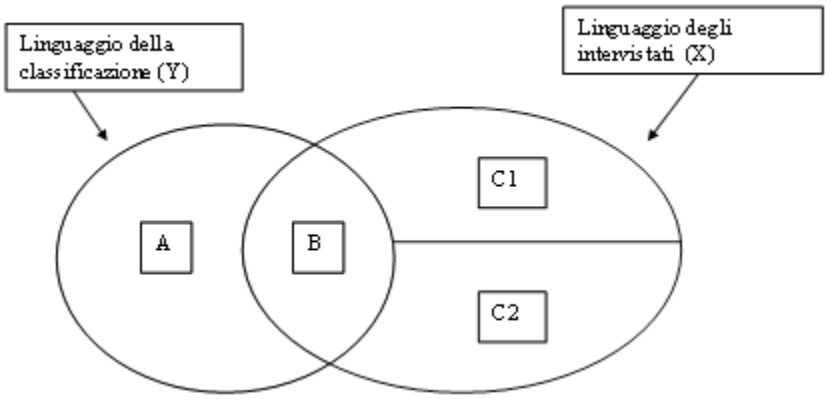

Figura 1. Intersezione tra il linguaggio della classificazione e il linguaggio degli intervistati

La procedura si basa sul confronto tra il vocabolario adottato dai rispondenti (X) e un dizionario specifico, riferito nel nostro caso alla classificazione ufficiale dell'Ateco (Y). Tale dizionario può essere facilmente ottenuto a partire dall'insieme delle denominazioni ufficiali della classificazione Ateco 2007, da cui sono state estrapolate le cinque sezioni in esame in modo da definire un primo dizionario di riferimento per ciascuna sessione di analisi. Si tratta, peraltro, dello stesso dizionario contenuto nel navigatore della classificazione che i rilevatori consultano nel corso dell'intervista per codificare la risposta degli intervistati.

L'intersezione tra i due insiemi determina tre blocchi logici di parole (figura 1). Il sottoinsieme di vocaboli contenuti esclusivamente nel dizionario Ateco (A) non è di interesse in questa fase, in quanto contiene le parole riferite a settori specifici non presenti nel campione di interviste in esame o a una terminologia specialistica non utilizzata dai rispondenti. Il sottoinsieme di termini comuni ai due vocabolari (B) è quello meno problematico, in quanto è costituito da parole specifiche che caratterizzano con elevate

[1] Si tratta delle domande C11 "Può dirmi il nome della sua professione e in che cosa consiste il suo lavoro?" e C15 "Cosa fa l'Ente o l'azienda presso la quale lavora? (Indichi i principali beni e/o servizi prodotti).

[2] In realtà si tratta di due classificazioni e due operazioni di codifica ben distinte per i rilevatori; tuttavia nel caso di descrittivi troppo generici, è necessaria la loro integrazione nella fase di correzione per ridurre l'indeterminatezza del contesto produttivo, al fine di codificare con maggior accuratezza l'attività economica. In alcuni casi, ad esempio, la sede presso la quale l'intervistato lavora viene esplicitata nella descrizione della professione invece che nel descrittivo dell'attività economica. Al contrario, in presenza di campi descrittivi valorizzati in maniera appropriata il ricorso alla lettura integrata non è necessario.

probabilità il settore di riferimento, oltre che da quelle strumentali della lingua (articoli, preposizioni ecc.) che non sono di interesse per la presente analisi. Di maggiore interesse è invece il sottoinsieme (C), composto dalle parole riportate dai rilevatori e non presenti nel dizionario Ateco della divisione specifica. Questo sottoinsieme può essere a sua volta scomposto in due blocchi: da un lato la terminologia "pertinente" (C1) che gli intervistati utilizzano per descrivere la propria attività pur con vocaboli non presenti nella classificazione ufficiale (ad esempio laddove la classificazione parla di 'coltivazione di agrumi' gli intervistati potrebbero parlare direttamente di 'limoni' o di 'arance'); dall'altro invece la terminologia "non pertinente" (C2) che probabilmente nasconde la presenza di errori di codifica; ad esempio termini quali 'imbottigliamento', 'confezionamento', 'commercio' nelle interviste classificate dai rilevatori nel settore dell'agricoltura, che possono invece essere ricondotti ad attività industriali o di commercializzazione.

I due sottoinsiemi di parole non presenti nel dizionario Ateco (d'ora in poi denominati 'Unici') costituiscono il punto di partenza della procedura di individuazione degli errori: da un lato i termini da noi attribuiti al sottoinsieme C1 vengono aggiunti di volta in volta al dizionario di partenza dell'Ateco, in modo da ridurre i tempi dell'analisi per le sessioni di controllo dei trimestri successivi (ampliando quindi progressivamente il sottoinsieme B); dall'altro l'insieme di termini del sottoinsieme C2 sono utilizzati per le procedure di controllo e correzione.

Il software di analisi testuale utilizzato per costruire i vocabolari e calcolare le intersezioni, Taltac2 (Bolasco, 2013), consente di rintracciare in modo piuttosto semplice tutti i record che contengono una determinata parola o una loro combinazione, visualizzando sia il contenuto del campo testuale sia l'insieme di variabili a esso associate per la valutazione dell'eventuale errore di codifica.

Si tratta di una funzione particolarmente innovativa presente in Taltac2, la Ricerca Entità (RE), che consente di considerare come unità di analisi non solo la singola parola presente in un testo, come avviene nell'analisi lessicale classica, ma anche l'intero record in cui questa compare. Mediante *query* specifiche, che si avvalgono di operatori logici e booleani e operatori di distanza tra parole, è possibile individuare anche combinazioni di parole (ad. es. 'asilo nido' o 'cooperativa di ristorazione') o insiemi di parole accumunate da una specifica annotazione in uno dei campi del vocabolario, ad esempio la categoria semantica. Sui record che rispondono positivamente alle *query* è possibile effettuare operazioni di etichettatura o conteggio, assegnando una nuova variabile al data set di partenza (nel nostro caso 'Ateco corretta')[3].

Una volta appurato che la presenza di un determinato termine o combinazione di termini permette di individuare un errore di classificazione, è possibile indicare il codice Ateco corretto, aggiungendolo al file di partenza in una nuova variabile personalizzata. Se si tratta di un errore ricorrente, che si presume possa ripresentarsi anche nella correzione del successivo file trimestrale, l'istruzione di ricerca, comprensiva dell'indicazione della codifica Ateco corretta, può essere eseguita automaticamente nelle successive sessioni di correzione, in modo da individuare contemporaneamente sia l'errore sia la codifica pertinente.

In questo modo le sessioni di correzione successive alla prima "apprendono" dalle operazioni effettuate in precedenza, sia perché viene ampliato il dizionario specifico per ciascuna

[3] Per ulteriori dettagli si veda (Bolasco, 2013), soprattutto cap. V, e per alcune applicazioni della Ratta- Rinaldi, (Loré, 2010) e della (Ratta-Rinaldi, 2009).

divisione Ateco - che riduce di volta in volta il numero di vocaboli del sottoinsieme C da controllare - sia perché gli errori di codifica più frequenti possono essere individuati e corretti in automatico. Dopo oltre un anno e mezzo dalla messa a punto della strategia di controllo, è stato dunque possibile semplificare ulteriormente - con una conseguente riduzione dei tempi - la fase di controllo testuale, grazie alla progressiva automatizzazione della stessa e alla possibilità di "apprendere" dalle sessioni di correzione precedenti. Difatti, da un lato le *query* che girano in automatico consentono di individuare circa il 60% dei record da correggere, dall'altro la progressiva riduzione del sottoinsieme di parole da controllare (insieme C2) rende piuttosto agevole etichettare direttamente tutti i record che contengono una parola anomala, controllando poi direttamente nel file finale l'effettiva esattezza della codifica proposta dal rilevatore. In pratica, per ciascuna sessione, con una semplice procedura di *tagging* semantico viene assegnata l'etichetta "VERIFICARE" a tutte le parole anomale del sottoinsieme C2 (insieme a quelle risultate anomale nelle sessioni precedenti). Piuttosto che costruire *query* specifiche per ciascuna di queste parole - spesso presenti con frequenza molto ridotta nel corpus - la funzione di RE consente di assegnare automaticamente un'etichetta (ad esempio "VERIFICARE") a tutti i record che contengono una delle parole anomale presenti nella lista. Nel corso della successiva fase di verifica del risultato finale sarà possibile aggiungere il codice Ateco corretto ai record che risultano effettivamente classificati in maniera erronea.

Il processo di correzione si completa pertanto attraverso una verifica puntuale della coerenza dei dati; in ogni sessione, difatti, è sempre necessario un intervento manuale da parte di esperti per la validazione delle correzioni effettuate e per la definitiva attribuzione del codice Ateco a 6 digit.

3. Il problema a monte: la qualità delle stringhe descrittive

L'analisi delle stringhe testuali ha consentito di far emergere un ulteriore problema nella qualità del testo inserito dai rilevatori, poiché descrizioni troppo generiche non consentono di codificare l'attività economica con il dettaglio richiesto dall'indagine sulle Forze di Lavoro (sesto digit per l'Ateco e quinto per le professioni). Per comprendere la portata di questo problema è necessario considerare le caratteristiche delle classificazioni in uso, che utilizzano variabili di tipo categoriale e sottostanno a criteri di ordinamento astratti che rendono piuttosto complessa l'attività di codifica (Vicari et al., 2009). Mentre la classificazione delle professioni, tra le due, appare relativamente meno problematica poiché richiede di descrivere un insieme di attività lavorative concretamente svolte da ciascun individuo, la classificazione Ateco è di più complesso utilizzo. Quest'ultima, infatti, è una classificazione piatta, non ordinabile e molto analitica (prevede in tutto 1.224 sottocategorie) che rimanda a un livello più astratto, ovvero l'attività svolta dall'ente/impresa presso cui il soggetto presta la propria opera, attività che non sempre gli intervistati conoscono in maniera sufficiente. Un contributo fondamentale al processo di codifica è pertanto fornito dal rilevatore, che attraverso le proprie esperienze e capacità relazionali dovrebbe tendere a ottenere il maggior numero di informazioni riguardo le variabili indagate, in modo da rendere il più possibile pertinente ed esaustiva la codifica.

Per tale ragione, descrizioni troppo generiche non consentono né di assegnare una codifica pertinente né di controllare ex-post l'esattezza del codice Ateco inserito dal rilevatore. Per risolvere quest'aspetto critico, sono state analizzate tutte le stringhe di risposta composte da meno di 13 caratteri inserite dai rilevatori nei primi due trimestri del 2012, al fine di individuare in quali casi una parola troppo corta è insufficiente a descrivere con esattezza una professione o l'attività economica. Ad esempio mentre il termine "Asl" (Azienda sanitaria

locale) consente di descrivere in maniera abbastanza puntuale l'attività economica, lo stesso non si può dire per "commercio", che necessita invece di una descrizione più articolata per poter effettuare una codifica pertinente. A partire dall'insieme delle risposte con meno di 13 caratteri è stato definito un dizionario di circa 800 descrittivi eccessivamente generici per una corretta codifica dell'Ateco, cui è stato aggiunta, quando necessario, l'indicazione dei motivi per cui la descrizione è da ritenersi inadeguata. È stato quindi messo in piedi un sistema di monitoraggio capace di segnalare automaticamente la presenza di un descrittivo generico in un'intervista appena trasmessa, in modo da restituire un riscontro immediato ai rilevatori tramite la ditta esterna che li coordina. Se si prende in considerazione l'insieme dei circa 25 mila record relativi agli occupati codificati dai rilevatori CAPI nel III trimestre 2012, la presenza di un descrittivo generico per l'attività economica ha riguardato circa l'11% dei record. Anche se una descrizione generica non comporta necessariamente un errore di codifica, si ritiene che migliorare la qualità delle stringhe descrittive possa aiutare a migliorare l'accuratezza complessiva del processo di codifica.

4. Le correzioni effettuate e gli errori più ricorrenti

La procedura di correzione adottata si è andata di volta in volta affinando. Per comprendere l'entità delle correzioni effettuate, nella Tabella 1 sono presentati gli interventi eseguiti nel 2012: nel complesso dei quattro trimestri dell'anno sono stati corretti 1.543 record, pari all'1,7% dei circa 90 mila controllati[4]. Le divisioni più critiche risultano i servizi generali della PA (con il 4,1% di errori), l'istruzione (1,6%) e a seguire l'agricoltura e la sanità (1,4%).

Divisioni	Codifiche errate	Record controllati	% su controllati
Agricoltura	145	10.604	1,4
Costruzioni	180	18.739	1,0
PA	669	16.204	4,1
Istruzione	272	17.365	1,6
Sanità	277	19.386	1,4
Totale	**1.543**	**88.846**	**1,7**

Tabella 1. Record errati e totale di record controllati per divisione Ateco – Anno 2012

L'analisi puntuale delle correzioni ha consentito inoltre di ricostruire la casistica degli errori più ricorrenti per ciascuna divisione, che costituisce la base per la progettazione di attività di richiamo formativo presso gli stessi rilevatori.

Ad esempio, le codifiche erroneamente attribuite all'agricoltura rivelano una confusione tra "prodotto" e "processo", come nel caso delle attività di trasformazione industriale di prodotti agricoli (industria alimentare, macellazione carni, trasformazione) o di commercializzazione dei prodotti (vendita di frutta, piante, mercato ortofrutticolo, ecc.). Allo stesso modo, tra i record erroneamente attribuiti alle costruzioni vi sono alcune attività da ricondurre invece all'industria manifatturiera poiché riferite alla produzione dei materiali utilizzati in edilizia (ad esempio pavimenti, prefabbricati, tettoie).

La divisione nella quale si riscontra il maggior tasso di errore è quella dei Servizi generali della Pubblica amministrazione, che presenta una casistica di errore molto articolata. L'errore più frequente riguarda il personale sanitario (medici, infermieri) o amministrativo (impiegati, operatori) che lavora in strutture della sanità o dell'assistenza, erroneamente classificato nella

[4] Nel corso del 2013 sono stati corretti 1.067 record, pari all'1,5% di quelli controllati.

Pubblica Amministrazione anziché nella sanità. Un altro errore frequente riguarda la codifica delle attività connesse alla gestione dei rifiuti, delle reti idriche o al trasporto pubblico locale, che secondo le regole della classificazione bisogna codificare tra le attività dell'industria o dei trasporti in presenza di unità locali che svolgono tali attività in via principale.

Un altro caso problematico, che si riscontra nei record classificati dell'istruzione e della sanità, riguarda i servizi di ristorazione o di pulizia svolti presso le scuole o gli ospedali, che quando appaltati a ditte esterne devono riportare i codici Ateco delle attività di pulizia e di ristorazione.

5. Linguaggio degli intervistati e linguaggio della classificazione

Il confronto del linguaggio degli intervistati con quello della classificazione ufficiale ha evidenziato inoltre un limite dello strumento di codifica, che utilizza un vocabolario piuttosto astratto e lontano dal modo di esprimersi degli intervistati (della Ratta, Gallo, Loré, 2011). Nel corso della loro attività di codifica i rilevatori hanno a disposizione un motore di ricerca della classificazione (navigatore) che consente di ricercare stringhe di testo per individuare la codifica più appropriata nella classificazione Ateco. Tuttavia, se la parola menzionata dall'intervistato non è presente nel dizionario ufficiale, il rilevatore dovrà ricorrere esclusivamente alle sue nozioni sulla classificazione per individuare il codice giusto, esponendosi a maggiori rischi di errore. Al contrario una classificazione arricchita dalla terminologia impiegata dai rispondenti potrebbe facilitare il lavoro di codifica riducendo gli errori.

Per valutare la distanza tra i due linguaggi si è proceduto all'intersezione tra i due vocabolari (vedi figura 1), quello della classificazione ufficiale e quello delle risposte aperte fornite dai rilevatori in uno specifico trimestre (il II del 2012) relative alla sola codifica Ateco, circoscrivendo l'analisi alle cinque sezioni soggette a procedura di controllo[5].

La misura più utile a valutare il grado di indipendenza lessicale tra i due linguaggi è l'indice di connessione lessicale (Cortelazzo & Tuzzi, 2008). Nel nostro caso l'indice di indipendenza del linguaggio della classificazione (Y) rispetto al linguaggio degli intervistati (X) è dato dal rapporto tra le forme di Y non presenti in X (V_A) rispetto al totale delle forme di Y (V_A+V_B). La maggiore o minore dipendenza[6] tra i due vocabolari fornisce un'indicazione sull'utilità della classificazione ufficiale per codificare l'attività economica. Se si guarda al valore dell'indice (Tab. 2) si nota che le sezioni denotate da maggiore indipendenza sono proprio l'agricoltura e l'istruzione, due settori nei quali il tasso di errore è piuttosto elevato. In questi settori l'indice di connessione lessicale è pari rispettivamente a 0,54 e 0,47. La relazione tra errore e indipendenza è meno evidente nel settore affetto dal maggior livello di errore, i servizi generali della PA, in cui l'indice assume uno dei valori più bassi (0,34). Probabilmente in questo settore gli errori di classificazione sono da ricondurre non tanto all'inadeguatezza dello strumento di codifica, ma piuttosto alla notevole complessità del comparto, che ha fatto emergere un evidente deficit concettuale dei rilevatori, da colmare in sede formativa. Peraltro, è proprio in questo settore che la descrizione della professione (non considerata nell'esperimento di confronto in quanto non presente nel dizionario Ateco) consente di

[5] Per un esempio di confronto tra vocabolari provenienti dalle descrizioni fornite dagli intervistati e vocabolari generati da liste precodificate, si veda anche Tuzzi e Zaccarin (2004).

[6] L'indice raggiunge il valore zero quando i due testi sono uguali (cioè perfettamente «dipendenti» uno dall'altro).

individuare molti errori di codifica. Ad esempio se l'intervistato dichiara di svolgere la professione di "autista di mezzi pubblici" presso il comune, è proprio la descrizione della professione che consente di comprendere che si tratta di attività da classificare nei servizi di trasporto di passeggeri e non nei servizi generali della PA, mentre la semplice stringa "comune" è soltanto generica ma in sé non errata).

Oltre all'indice di connessione lessicale è stata utilizzata un'ulteriore misura di similarità, la distanza intertestuale di Labbé (Labbé & Labbé, 2003), che consente di valutare la similarità tra i corpus nella frequenza di impiego delle parole comuni, tenendo conto della differenza di ampiezza dei due linguaggi a confronto (Cortelazzo et al., 2012)[7]. In questo caso le maggiori distanze si osservano nel settore della PA e dell'istruzione, che sono gli stessi settori caratterizzati da una più elevata percentuale di errori di codifica, con valori dell'indice pari a 0,47 e 0,41. Il settore dell'istruzione sembra dunque affetto sia da una maggiore indipendenza tra linguaggio degli intervistati e della classificazione, sia da una maggiore differenza nella frequenza con cui anche le parole comuni sono utilizzate, mentre quello della PA, che pure aveva un grado di dipendenza maggiore tra vocabolario degli intervistati e della classificazione, mostra una distanza più elevata in termini di utilizzo delle forme grafiche comuni.

Per approfondire ulteriormente la natura della distanza tra i due linguaggi è utile analizzare anche per ciascun settore i termini presenti solo nel linguaggio degli intervistati. La tabella 3 riporta in ordine di occorrenza le prime dieci forme grafiche pertinenti[8] per ciascun settore, il cui lemma non è presente nel dizionario della classificazione.

Settore attività economica	Vocabolario intervistati (X)		Dizionario classificazione (Y)		Parole presenti solo in Y (V_A)	Forme comuni $V_{(B)}$	Connessione lessicale (V_A/V_A+V_B)	Distanza Labbé
	Occorrenze (N_X)	Forme grafiche (V_X)	Occorrenze (N_Y)	Forme grafiche (V_Y)				
Agricoltura	7.942	740	2.259	460	247	213	0,54	0,38
Costruzioni	13.558	1.227	1.723	363	138	225	0,38	0,36
Istruzione	9.508	728	779	199	94	105	0,47	0,41
Servizi generali PA	7.710	856	717	182	61	121	0,34	0,47
Sanità	11.734	1.016	707	235	78	157	0,33	0,35

Tabella 2. Linguaggio degli intervistati e della classificazione. Connessione lessicale e distanza intertestuale - II trimestre 2012

[7] Dati i due corpus Y e X con rispettive numerosità $N_Y<N_X$ la distanza intertestuale si calcola sulle parole comuni ai due corpus secondo la formula:

$d(Y,X) = \frac{\sum_{i \in Y \cup x} | f_{iY} - f^*_{iX} |}{2N_y}$ dove f_{iY} sono le frequenze delle parole del corpus più piccolo Y e f^*_{iX} sono pari a f_{iX}*Ny/Nx, ovvero le frequenze del corpus più grande X "ridotte" in ragione della dimensione del testo più piccolo Y. L'indice varia tra 0 e 1, dove 0 indica il massimo della similarità tra i due corpus.

[8] Come si è visto nel paragrafo 2, tra i termini presenti esclusivamente nel linguaggio degli intervistati si incontrano sia i termini non pertinenti indizio di errore (il sottoinsieme C2) sia termini pertinenti che specificano ulteriormente l'attività economica (C1). L'analisi condotta di seguito è naturalmente riferita solo a questi termini.

Ad esempio in agricoltura troviamo forme grafiche che fanno riferimento a specifici oggetti dell'attività agricola (vigneto/i, campo, viti, oliveto ecc.) che non sono compresi nel dizionario Ateco. Nelle costruzioni spiccano parole inerenti il tipo di sede (impresa, ditta, azienda), e vocaboli comuni che connotano l'attività edilizia (ristrutturazione, abitazioni, case, interni). Nell'istruzione gli intervistati utilizzano molto gli aggettivi che specificano il tipo di scuola utilizzando il linguaggio comune al posto di quello istituzionale (materna, tecnico, alberghiero, industriale, agrario ecc.), mentre nella PA le parole più diffuse specificano soprattutto il tipo di amministrazione (comunale, provinciale ecc.). Infine in sanità le parole non presenti nel dizionario Ateco riguardano in particolar modo le Asl, come si evince sia dalla specifica forma grafica, sia dai termini "azienda" e "locale", oppure l'attività svolte dalle cooperative e le cure dentistiche.

Agricoltura		Costruzioni		Istruzione		PA		Sanità	
F.G.	Occ	F.G.	Occ.	F.G.	Occ.	F.G.	Occ.	F.G.	Occ.
vigneto/i	62	impresa	450	materna	312	comunale	160	azienda	191
campo	29	ristrutturazione/i	386	tecnico	172	esercito	89	generale	152
viti	23	ditta	126	inferiore	83	agenzia	88	cooperativa	138
coltiva	23	abitazioni	100	pubblica	59	entrate	61	centro	123
oliveto	22	residenziali	73	statale	45	ente	54	asl	114
olio	19	case	64	comprensivo	44	provinciale	49	dentistico	99
polli	15	termoidraulici	50	alberghiero	44	marina	42	locale	99
agricoltore	15	azienda	49	privata	40	inps	40	privata	58
fieno	12	condizionamento	36	industriale	39	ufficio	38	civile	52
cooperativa	12	interni	35	agrario	20	aeronautica	32	soccorso	21

Tabella 3. Linguaggio degli intervistati: lemmi non presenti nel dizionario Ateco per settore e numero di occorrenze. II trimestre 2012

Questo tipo di risultato suggerisce l'opportunità di arricchire il dizionario ufficiale: verificata l'esistenza di legami tra il tasso di errore e distanza tra i linguaggi, una maggiore aderenza del linguaggio utilizzato dallo strumento di codifica non potrà che facilitare tutto il processo. Un'applicazione di questo esercizio estesa a tutti i settori di attività e costruita su una base di dati più consistente potrebbe costituire il punto di avvio per un processo più ampio, in cui il materiale testuale raccolto dall'indagine possa contribuire ad arricchire il navigatore utilizzato per la codifica con esempi tratti dal modo concreto di esprimersi degli intervistati, come nei casi di indagini sulle famiglie.

Si tratta di una procedura già implementata per il navigatore delle professioni, che attualmente è arricchito con esempi di nuove professioni rilevate in sede di indagine, mentre il navigatore delle attività economiche si basa esclusivamente sull'elenco delle attività ufficiali contenute nel dizionario della classificazione (Gallo & Loré, 2012).

6. E gli altri settori? Un controllo con ACTR[9]

Al fine di ottimizzare l'automazione della procedura di individuazione e correzione degli errori è stato effettuato un test utilizzando ACTR (*Automatic Coding by Text Recognition*), un

[9] Le elaborazioni con ACTR sono state effettuate da Filomena De Filippo, che ha collaborato con gli autori per la stesura del paragrafo.

sistema in uso nell'Istat dalla fine degli anni '90, che consente nella sua versione aggiornata di individuare le possibili codifiche per l'Ateco 2007 a fronte di una descrizione testuale. L'obiettivo del test era quello di valutare le potenzialità dello strumento per le operazioni di controllo e correzione, in modo da velocizzarne i tempi di esecuzione ed estendere l'attività di correzione a tutte o almeno ad altre divisioni dell'Ateco. Come il nostro test ha mostrato, un limite del sistema consiste sia nell'impiego di una terminologia distante dal linguaggio corrente degli intervistati, sia nell'impossibilità di coniugare le informazioni tra attività economica e professione (classificazione non implementata da ACTR nell'ultima versione della CP2011 attualmente in uso), laddove, soprattutto per le divisioni più critiche, è emerso che proprio la combinazione tra Ateco e professione costituisce la chiave per risolvere le descrizioni più ambigue.

Il test effettuato ha riguardato due sezioni finora escluse dal processo di controllo, l'industria in senso stretto e il commercio. Il software utilizza un apposito algoritmo per misurare la similarità tra i testi presenti nel proprio dizionario e le descrizioni libere inserite dai rilevatori[10].

In base al livello di similarità riscontrato, i possibili abbinamenti del sistema ACTR si possono suddividere in unici, possibili, multipli e falliti. In particolare i record codificati come "unici" sono quelli di migliore qualità, quelli cioè per i quali ACTR individua un'unica possibile codifica dell'Ateco. L'incidenza dei record unici è piuttosto elevata, pari al 57,5% nell'industria in senso stretto e al 70,8% nel commercio e costruzioni. Di contro, ACTR non riesce in nessun modo a codificare il 9,6% dei record dell'industria e il 4,9% di quelli del commercio e delle costruzioni.

Passando alla corrispondenza tra la codifica indicata dal rilevatore e quella suggerita da ACTR, è stato invece riscontrato un elevato numero di discordanze al primo digit, vale a dire record che ACTR e rilevatore attribuiscono a divisioni differenti. Si tratta del 18,4% del totale dei record dell'industria in senso stretto e del 14,7% di quelli attribuiti al commercio e alle costruzioni, un risultato che all'apparenza sembra costituire un serio indizio di scarsa qualità delle codifiche. In realtà, a una verifica puntuale dei casi discordanti si è appurato che in circa 9 casi su 10 la codifica esatta è quella inserita dal rilevatore. Il software, infatti, non sembra possedere la flessibilità necessaria a interpretare le stringhe spesso eccessivamente sintetiche e/o generiche inserite dai rilevatori. In particolare, al termine dei controlli sono risultati errati appena 107 record attribuiti all'industria in senso stretto (pari all'1,1% del totale dei record esaminati) e 104 record al commercio (pari all'1,4%).

Il test, se da un lato ha evidenziato la presenza di un certo tasso di errore (seppur inferiore ai livelli inizialmente riscontrati nelle divisioni sottoposte a correzione sistematica), ha mostrato al contempo le difficoltà che comporterebbe l'impiego di ACTR come strumento di controllo e correzione utilizzando i dizionari attualmente implementati. Questi, infatti, risultano adatti soprattutto per descrizioni testuali precise e tecniche, come quelle che è possibile ottenere dalle indagini sulle imprese, in cui la descrizione dell'attività viene fornita da una persona ben

[10] Su ACTR si vedano (De Angelis, Macchia, Mazza, 2000; Macchia, Murgia, Talucci, 2008; Vicari, 2009). Senza entrare in dettagli tecnici, l'attività di codifica è preceduta da una fase di standardizzazione dei testi chiamata *parsing*, finalizzata a rimuovere tutte le varianti grammaticali e sintattiche, in modo da rendere uguali due descrizioni con lo stesso contenuto semantico originariamente diverse. Oltre che sulla risposta da codificare, il *parsing* viene effettuato anche sulle descrizioni del dizionario (*reference file*). I testi così trattati vengono confrontati tra di loro: se si realizza un match perfetto (*direct match*) viene assegnato un unico codice, altrimenti il software tramite un algoritmo individua nel dizionario i testi più simili a quello da codificare (*indirect match*).

informata sul settore in cui opera l'impresa stessa. D'altra parte, invece, il software non è sufficientemente flessibile per recepire il linguaggio degli intervistati perché utilizza esclusivamente il dizionario della classificazione ufficiale. Pertanto, l'impiego di questo software non sembra di immediata applicabilità, anche perché richiede tempi elevati per il controllo dei risultati.

7. Conclusioni

La procedura fin qui descritta presenta il vantaggio di concorrere al miglioramento progressivo della qualità dei dati dell'indagine sulle forze di lavoro, eliminando un numero significativo di record mal classificati.

Si tratta di un'attività che ormai viene svolta sistematicamente in ogni trimestre in modo da assicurare sia la qualità sia la coerenza della dinamica occupazionale nei diversi settori di attività economica. Rimane invece aperta e da valutare con attenzione l'eventualità di estendere il campo di correzione ad altri settori. Va considerato, difatti, che avviare un'operazione di controllo sistematico su tutti i dati richiederebbe un tempo consistente. Da un lato occorrerebbe un investimento iniziale per costruire dizionari specifici per ogni sezione/divisione di attività economica, attraverso cui confrontare la coerenza delle stringhe descrittive, dall'altro le procedure non possono essere interamente automatizzate perché si rende comunque necessario una verifica finale di congruità da parte di esperti in materia. Non da ultimo è poi necessario tener conto dei tempi stringenti di diffusione dei dati dell'indagine RFL.

Una strada più compatibile con le risorse attuali è pertanto quella di intervenire sempre più in sede di prevenzione dell'errore. La strategia di Text mining adottata ha consentito di individuare alcune delle cause più frequenti dell'errata classificazione, definendo specifiche strategie di prevenzione dell'errore non campionario. Da un lato la ricostruzione della casistica di errori (successivamente ampliata tramite controlli a campione realizzati su tutte le sezioni di attività economica) consente la progettazione di interventi formativi che possano contribuire a prevenire gli errori riducendoli alla fonte piuttosto che correggerli ex-post. Dall'altro un ragionamento più ampio sull'aderenza del navigatore dell'Ateco al linguaggio degli intervistati potrà consentire la definizione di uno strumento più efficace per il lavoro di codifica, che possa essere applicato anche in altre indagini sulle famiglie o sperimentato in indagini auto compilate sul web. Su entrambi questi fronti la possibilità di esplorare il linguaggio concreto degli intervistati offre interessanti prospettive di sviluppo.

Riferimenti bibliografici

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

Cortelazzo M. et Tuzzi A. (2008). *Metodi statistici applicati all'italiano.* Zanichelli, Bologna.

Cortelazzo M. A., Tuzzi A. et Nadalutti P. (2012). Una versione iterativa della distanza intertestuale applicata a un corpus di opere della letteratura italiana contemporanea. In A. Diester, D. Longrée, G. Purnelle (eds), *Actes des 11es Journées internationales d'Analyse statistique des Données Textuelles.* Université de Liège, Facultés Universitaires Saint-Louis, Bruxelles, pp. 295-307.

De Angelis R., Macchia S. et Mazza L. (2000). Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale. Istat, *Quaderni di ricerca - Rivista di statistica Ufficiale*, n.1/2000, pp.29-54.

della Ratta Rinaldi F. (2009). Il trattamento dei dati. In Gallo F., Scalisi P., Scarnera C. *L'indagine sulle professioni. Anno 2007, Contenuti, metodologia e organizzazione.* Collana Metodi e Norme, n. 42, cap. 7, pp. 73-89, Roma, Istat.

della Ratta-Rinaldi F. et Loré B. (2010). Il lavoro e i suoi contenuti. Un'applicazione di Text Mining per categorizzare le attività dettagliate di lavoro nell'indagine campionaria sulle professioni Istat. In Bolasco S., Chiari I., Giuliano L. (a cura di). *Statistical Analysis of Textual Data. Proceedings of 10th International Conference 9-10 June 2010,* pp.195-202, 917-928 e 929-937, Roma.

della Ratta-Rinaldi F., Gallo F. et 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 *Cladag 2011, 8th Scientific Meeting of the Data Analysis Group of the Italian Statistical Society.* Pavia, settembre 2011, Pavia University press.

Gallo F. et Loré B. (2012). *Training on the new occupational classification: the Italian experience.* Istat Working paper n. 12.

Istat (Aa. Vv.). (2011). Linee guida per la qualità dei processi statistici, Roma. (http://www.istat.it/it/files/2010/09/Linee-Guida-Qualit%C3%A0-_v.1.1_IT.pdf).

Labbé C. and Labbé D. (2003). La distance intertextuelle. *Corpus*, 2:95-118.

Macchia S., Murgia M. et Talucci V. (2008). Coding the spoken language through the integration of different approaches of textual analysis. In *JADT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles.* Lyon, 12-14 mars 2008.

Tuzzi. A. et Zaccarin S. (2004). Il lavoro raccontato dai laureati: analisi lessico-testuale delle professioni. In E. Aureli Cutillo. *Strategie metodologiche per lo studio della transizione Università lavoro,* pp. 357-373, Padova, Cleup.

Vicari P., Ferrillo A. et Valery A. (a cura di), (2009). *Classificazione delle attività economiche - Ateco 2007.* Metodi e norme n. 40, Roma, Istat.

Vicari P. (a cura di). (2009). *L'ambiente di codifica automatica dell'Ateco 2007.* Metodi e norme : 41, Roma, Istat.