Sperimentazione, implementazione e gestione dell'ambiente di codifica automatica della classificazione delle Attività economiche

Esperienze effettuate e prospettive per il Censimento dell'Industria e dei Servizi

2001

Studi e attività svolte nell'ambito del
Gruppo di Lavoro '*sulla codifica automatica di attività economica e natura giuridica*'

Redatto dai partecipanti al G.d.L.

Coordinatore: Stefania Macchia

Membri: **Piero Bretti**
Massimiliano Degortes*
Angelina Ferrillo
Loredana Mazza* (con funzioni di membro e segretario)
Domenico Perrone*
Alberto Socini**
Paola Vicari**
Alberto Valery***

Sommario

Questo documento ha una duplice finalità: descrivere le implicazioni metodologiche e di processo conseguenti l'adozione della codifica automatica in un'indagine statistica, nonché mettere a disposizione dei futuri gestori dell'applicazione di codifica automatica la necessaria documentazione sia tecnica che contenutistica, inerente i criteri classificatori adottati nella costruzione della base informativa. (inerente la classificazione esaminata).
Nella pratica, quindi, è riportata una approfondita descrizione dell'esperienza effettuata nella

disposizione dei futuri gestori dell'applicazione di codifica automatica la necessaria documentazione sia tecnica che contenutistica, inerente i criteri classificatori adottati nella costruzione della base informativa. (inerente la classificazione esaminata).
Nella pratica, quindi, è riportata una approfondita descrizione dell'esperienza effettuata nella costruzione dell'ambiente applicativo della classificazione delle attività economiche 1991, dei risultati ottenuti nell'ambito del censimento intermedio dell'industria (indagine Long Form), delle prospettive per l'adozione di questa tecnica nell'ambito del prossimo censimento e dell'impatto sull'applicazione derivante dalla revisione della classificazione. Sono infine esposti i criteri classificatori, concordati con gli esperti della classificazione, sulla base dei quali è stata implementata la base informativa.

Abstract

This document is aimed at two purposes: to describe methodological and process implications deriving form the adoption of automatic coding in a statistical survey and to support future coders, responsible for the automatic coding application, with all the necessary documentation, concerning both technical aspects and pertaining to content ones, relating to coding criteria adopted in building the data base of the examined classification.
In practice, then, an exhaustive description is reported concerning: the experience made in building the coding application of Industry classification (1991), the results obtained during the intermediate industry census (Long Form survey), the perspectives for the adoption of this technique in next census and the impact on the automated coding application of the classification revision. Finally, coding criteria followed in implementing the data base, defined with classification experts, are reported.

Indice

1 Obiettivi del gruppo di lavoro e problematiche inerenti la codifica nel corso dei censimenti

Il gruppo di lavoro è stato costituito in funzione della decisione *'sull'opportunità di realizzare un dizionario di voci di attività economica e di natura giuridica che consenta l'utilizzo generalizzato di un software di codifica automatica'* (Istat, deliberazione n. 1723/P).

Alcune delle principali motivazioni di tale decisione sono state relative all'imminenza del **censimento dell'industria e del commercio**, nonché alla possibilità di **sperimentare questa metodologia nel corso del censimento intermedio dell'industria (indagine Long Form**).

L'ipotesi, in seguito avvalorata dai fatti (come si vedrà nei paragrafi successivi), era infatti quella che questa innovazione avrebbe comportato una serie di vantaggi sia di carattere organizzativo che di qualità dei risultati.

In corso di indagine, infatti, il ridursi della numerosità dei casi da risolvere manualmente avrebbe consentito la centralizzazione dell'attività di codifica manuale presso il Servizio di statistiche censuarie sulle attività economiche (CUE), evitando il ricorso agli Uffici Comunali che per tale attività abitualmente impiegano alcuni mesi di tempo.

Tale organizzazione, oltre alla drastica riduzione dei tempi, avrebbe garantito un più elevato livello di qualità dei dati in quanto:

- i criteri interpretativi della risposta testuale ai fini dell'attribuzione del codice (estremamente delicati per una classificazione di tale complessità) sarebbero stati univoci: infatti gli stessi criteri adottati per la costruzione del dizionario *elaborabile* si sarebbero concordati con i codificatori esperti del servizio CUE partecipanti al gruppo di lavoro; gli stessi esperti avrebbero fornito assistenza per la codifica manuale dei casi non risolti automaticamente, elaborando metodi ed indicazioni per i casi dubbi o problematici;
- si sarebbe ovviato alla problematica inerente gli errori di registrazione dei codici.

(I verbali di riunione sono disponibili presso la segreteria del gruppo di lavoro).

2 Il sistema di codifica ACTR

Il sistema di codifica adottato è ACTR v3 (Automated Coding by Text Recognition), sviluppato da Statistics Canada, che è stato utilizzato con successo per applicazioni nell'ambito della statistica ufficiale (Tourigny e Moloney, 1995) e

2 Il sistema di codifica ACTR

Il sistema di codifica adottato è ACTR v3 (Automated Coding by Text Recognition), sviluppato da Statistics Canada, che è stato utilizzato con successo per applicazioni nell'ambito della statistica ufficiale (Tourigny e Moloney, 1995) e sarà adottato per i prossimi censimenti non soltanto dal Canada, ma anche dal Regno Unito e dalla Croazia.
Al momento della costituzione del gruppo di lavoro, il sistema era già stato sperimentato in Istat relativamente a tre classificazioni, Professione, Titolo di studio ed Attività Economica (relativamente a quest'ultima variabile, era stato costituito soltanto un ambiente applicativo piuttosto scarno, in quanto non erano disponibili fonti con cui alimentarlo) e testato sui campioni di diverse indagini.
Oltre alle esperienze effettuate da altri paesi, una motivazione forte per preferire questo sistema rispetto ad altri è relativa al fatto che questo presenta la caratteristica di essere assolutamente generalizzato, quindi indipendente dalla classificazione considerata e dalla lingua.
Il sistema gestisce applicazioni di codifica automatica (modalità batch in ambiente Unix o Windows) e con l'ausilio di un'interfaccia grafica in ambiente Windows permette di analizzare interattivamente i risultati della codifica di casi singoli, modificando agevolmente i parametri del sistema *('what if' tools*).
La logica di base del sistema è ispirata alla metodologia sviluppata originariamente presso US Census Bureau (Hellerman, 1983) ed utilizza degli algoritmi di ricerca e *matching* testuale messi a punto successivamente da ricercatori di Statistics Canada (Wenzowski, 1988).
L'attività di codifica vera e propria è preceduta dalla fase di standardizzazione dei dati testuali (definita '*parsing*') che ha lo scopo di rimuovere la variabilità grammaticale o sintattica che differenzia frasi con lo stesso contenuto semantico e che è irrilevante ai fini della codifica. **L'attività di standardizzazione è completamente controllata dall'utente, che ha il compito di adattarla al particolare contesto applicativo** (lingua, classificazione, tipologia di rispondente). Una delle peculiarità di ACTR rispetto ad altri sistemi è la notevole flessibilità e possibilità di personalizzazione del processo di *parsing*, mettendo a disposizione fino a 14 diverse funzioni (mappatura dei caratteri, eliminazione di parole ininfluenti, definizione di sinonimi per parole singole o per gruppi di parole, rimozione di suffissi, prefissi, etc…).
La risposta da codificare, una volta standardizzata, è confrontata con il dizionario di riferimento della classificazione, precedentemente sottoposto allo stesso processo di standardizzazione e caricato nel database di sistema. L'esito di tale confronto, basato sulla ricerca di parole in comune, può essere : *i)* un abbinamento esatto (*direct match*) che dà luogo all'assegnazione di un codice unico *ii)* un abbinamento parziale (*indirect match*), in questo caso il software individua, tramite un algoritmo di misura della somiglianza tra testi (S), il codice o i codici del dizionario con descrizione più simile alla risposta fornita dal rispondente. Tale misura è funzione del numero di parole in comune e del loro grado di 'informatività' all'interno del dizionario di riferimento (*weighting algorithm*).
Le soglie di accettazione per la misura di somiglianza sono:

Smin à soglia minima di accettazione

Smax à soglia massima di accettazione

Δ S à minima distanza tra testo 'vincente' (S_1) e successivo (S_2)

I valori di soglia sono fissati dall'utente in funzione dei suoi obiettivi di qualità e permettono di suddividere i possibili risultati della fase di codifica in:

* codici **Unici** à $S_1 > Smax$ e $(S_1 - S_2) > \Delta S$
* codici **Multipli** à $S_1 > Smax$, $S_2 > Smax$ mentre $(S_1 - S_2) < \Delta S$
* codici **Possibili** à $Smin < S_1 < Smax$
* casi **Falliti** à $S_1 < Smin$

Nel primo caso il codice viene assegnato in modo completamente automatico per effetto di un abbinamento di testi che può essere completo o parziale. I rimanenti casi necessitano della valutazione da parte di codificatori (o di programmi ausiliari) che selezionino il codice corretto tra quelli proposti dal sistema. La fase di addestramento del sistema consiste nell'individuare gli opportuni correttivi al dizionario o alla strategia di *parsing* per ridurre al minimo i casi di fallimento o di codice multiplo.
E' chiaro che valori alti dei parametri soglia elevano l'*accuratezza* (definita come percentuale di codici 'Unici' corretti) a scapito dell'*efficacia* (definita come percentuale di codici 'Unici' assegnati), quindi nella scelta occorre bilanciare questi due aspetti della qualità dei risultati.

N.B. Per un approfondimento tecnico sul funzionamento di ACTR **si rimanda al manuale**.

3 Variabile 'Attività Economica'

3.1 Da dove si è partiti

Come già accennato, presso il Servizio Metodologie per la Produzione Statistica (MPS/C) è stato sperimentato il sistema di codifica automatica ACTR (Automatic Coding by Text Recognition), che ha la caratteristica di essere assolutamente generalizzato, ossia indipendente dalla lingua e dalla classificazione da utilizzare.
Ciò significa che, prima di essere adottato per ciascuna classificazione, è necessario fornirgli la 'conoscenza di base': costruire **l'ambiente applicativo** di partenza.
La sperimentazione effettuata ha riguardato la costruzione del citato **ambiente applicativo** per una serie di variabili, tra le quali l'ATECO.
Per chiarezza, si specifica che per ambiente applicativo si intende:

- dizionari 'elaborabili',
- file ausiliari (per l'adattamento all'italiano, per la definizione di sinonimi, parole o parti di parole ininfluenti, etc.),

Per chiarezza, si specifica che per ambiente applicativo si intende:

- dizionari 'elaborabili',
- file ausiliari (per l'adattamento all'italiano, per la definizione di sinonimi, parole o parti di parole ininfluenti, etc.),
- parametri di sistema che devono essere ottimizzati per ciascuna applicazione (es. soglie).

La costruzione dei dizionari 'elaborabili' (Macchia, 2001) è avvenuta in due fasi:

- **rielaborazione del manuale ufficiale** della classificazione (Istat, Classificazione delle attività economiche, 1991, Metodi e norme. Serie C – n.11) al fine di rendere i testi associati ai codici sintetici, espliciti (assolutamente non ambigui) ed univoci;
- **inserimento nel dizionario di testi precodificati** rilevati nell'ambito di precedenti indagini; per convenzione chiameremo, d'ora in poi, questi testi risposte '*empiriche'*. In questa fase, sono state utilizzate alcune risposte al quesito sull'Ateco fornite nel censimento della popolazione del 1991 e nell'indagine Short Form.

Al termine della fase di sperimentazione, si disponeva, per la variabile ATECO, di un *dizionario elaborabile* di 8860 testi e 1565 sinonimi, a fronte di una classificazione che prevede 874 modalità al massimo dettaglio (5 cifre Ateco).

L'applicazione sperimentale di tale sistema ad una prima tranche di dati dell'indagine Long Form ha dato adito alla codifica automatica del 47% dei testi.

Tabella.1 - Risultati della prima applicazione di codifica automatica sull'indagine Long Form

Testi del dizionario	Sinonimi	Efficacia del sistema
N.	**N.**	**%** di testi codificati automaticamente
8860	1565	47

Tuttavia, dalle esperienze effettuate da altri istituti di statistica, si stima che per variabili complesse, come l'attività economica, per garantire percentuali di codifica del 70-80%, si dovrebbe arrivare a disporre di dizionari di circa 40.000-50.000 testi.

Da qui le attività del gruppo di lavoro finalizzate non soltanto ad applicare la codifica automatica all'indagine Long Form, ma soprattutto a procedere all'arricchimento del dizionario (e più in generale dell'ambiente applicativo), al fine di garantire successivamente migliori performance del sistema e consentirne l'adozione per il prossimo censimento dell'industria e dei servizi.

3.1.1 Costruzione e struttura dell'ambiente applicativo

Al fine di riportare in questo documento tutta la documentazione necessaria ai futuri utenti di questa applicazione di codifica automatica, si entra ora maggiormente in merito circa questa prima fase, inerente alla costruzione del *dizionario elaborabile*, che, come già detto, è stata realizzata in due step:

- **rielaborazione del manuale ufficiale;**
- **inserimento nel dizionario di testi precodificati** .

La rielaborazione del manuale è consistita essenzialmente nelle seguenti operazioni:

- *semplificazione delle descrizioni complesse*, tramite la divisione delle descrizioni *'somma'* di più concetti;
  ad es. '*Commercio all'ingrosso di macchine accessori e utensili agricoli inclusi i trattori'* è stato rielaborato come
  'Commercio all'ingrosso di accessori e utensili agricoli',
  'Commercio all'ingrosso di trattori agricoli',
  'Commercio all'ingrosso di macchine agricole';
- *definizione di sinonimi*, tramite l'esplicitazione di liste di elementi specifici riconducibili ad un'unica categoria generale; ad es. nella '*Produzione agrumicola'* cadono sia la '*Produzione di limoni'* che quella di *'arance'* e '*mandarini*;
- *integrazione con materiale di riferimento*, come ad esempio note esplicative; ad es. '*questa classe non comprende….'* presenti nel manuale stesso oppure con altre classificazioni correlate (ad es. *Prodcom)*

A seguito di queste operazioni, è stato necessario **non riportare nel dizionario informatizzato** alcune attività contenute nella classificazione Ateco 91 che risultavano troppo generiche e potevano determinare *match* non corretti o non univoci. Una di queste è per esempio '*Costruzioni'*, in corrispondenza della quale è stato sostituito il codice '45' con *'n.c.' (Non Codificabile)* (vedi

alcune attività contenute nella classificazione Ateco 91 che risultavano troppo generiche e potevano determinare *match* non corretti o non univoci. Una di queste è per esempio '*Costruzioni'*, in corrispondenza della quale è stato sostituito il codice '45' con *'n.c.' (Non Codificabile)* (vedi paragrafo 5.6).

Altre descrizioni, invece, **non sono state incluse nel dizionario** nella loro formulazione completa; trattasi, per esempio, di: '*28.5 - TRATTAMENTO E RIVESTIMENTO DEI METALLI, LAVORAZIONI DI MECCANICA GENERALE PER CONTO TERZI'* che si ritrova nel dizionario come segue*:*

- *28.51.0 TRATTAMENTO E RIVESTIMENTO DEI METALLI*
- *28.52.0 LAVORAZIONI DI MECCANICA GENERALE PER CONTO TERZI*

Nell'ambito della *classificazione elaborabile* Ateco 91 è stato necessario **rivedere alcune dizioni poco appropriate**, come segue da elenco inviato dal CUE:

- *- (b1) ATTIVITA' DI CARICO E SCARICO DI PASSEGGERI A FERMATE EFFETTUATI MEDIANTE AUTOBUS* è diventata
  60.21.0 - (b1) TRASPORTO DI PASSEGGERI CON AUTOBUS A FERMATE FISSE
- *- (b1) ATTIVITA' DI CARICO E SCARICO DI PASSEGGERI A FERMATE EFFETTUATI MEDIANTE FERROVIA DI SUPERFICE* è diventata
  60.21.0 - (b1) TRASPORTO DI PASSEGGERI MEDIANTE FERROVIA DI SUPERFICIE
- *- (b1) ATTIVITA' DI CARICO E SCARICO DI PASSEGGERI A FERMATE EFFETTUATI MEDIANTE FERROVIA SOPRAELEVATA* è diventata
  60.21.0 - (b1) TRASPORTO DI PASSEGGERI MEDIANTE FERROVIA SOPRAELEVATA
- *- (b1) ATTIVITA' DI CARICO E SCARICO DI PASSEGGERI A FERMATE EFFETTUATI MEDIANTE FILOBUS* è diventata
  60.21.0 - (b1) TRASPORTO DI PASSEGGERI MEDIANTE FILOBUS A FERMATE FISSE
- *- (b1) ATTIVITA' DI CARICO E SCARICO DI PASSEGGERI A FERMATE EFFETTUATI MEDIANTE METROPOLITANA* è diventata
  60.21.0 - (b1) TRASPORTO DI PASSEGGERI MEDIANTE METROPOLITANA
- *- (b1) ATTIVITA' DI CARICO E SCARICO DI PASSEGGERI A FERMATE EFFETTUATI MEDIANTE TRAM* è diventata
  60.21.0 - (b1) TRASPORTO DI PASSEGGERI MEDIANTE TRAM
- *- (b1) ATTIVITA' DI CARICO E SCARICO DI PASSEGGERI A FERMATE GENERALMENTE FISSE (ESCLUSO QUELLO FERROVIARIO)* è diventata
  60.21.0 - (b1) TRASPORTO DI PASSEGGERI MEDIANTE AUTOVEICOLI TERRESTRI CON FERMATE FISSE (ESCLUSO QUELLO FERROVIARIO)

Sulla base del **criterio di prediligere l'attribuzione del codice al massimo dettaglio (5 cifre)** rispetto a quello più generico, alcune dizioni generiche (associate nel manuale a codici con meno di 5 cifre) sono state associate a codici al massimo dettaglio, ritenendo che l'informazione mancante nella dizione generica sottintenda, con una probabilità molto elevata, una casistica piuttosto che un'altra. Segue elenco fornito dagli esperti del CUE:

- *(A*) 18.2 - PRODUZIONE DI ALTRI INDUMENTI ESTERNI* à *18.22.1*
- *(A1) 18.2 - CONFEZIONE DI ALTRI INDUMENTI ESTERNI* à *N 18.22.1*
- *(A1) 18.2 - CONFEZIONI DI ALTRI INDUMENTI ESTERNI* à *18.22.1*
- *(A1) 20.51 - FABBRICAZIONE DI ALTRI PRODOTTI IN LEGNO* à *20.51.1*
- *(A1) 25.2 - FABBRICAZIONE DI ARTICOLI IN MATERIE PLASTICHE* à *25.24.0*
- *(A1) 25.1 - FABBRICAZIONE DI ARTICOLI IN GOMMA* à *25.13.0*
- *(A1) 26.6 - FABBRICAZIONE DI PRODOTTI IN CALCESTRUZZO, CEMENTO O GESSO* à *26.66.0*
- *(A1) 29.3 – FABBRICAZIONE DI MACCHINE PER L'AGRICOLTURA E LA SILVICOLTURA* à *29.32.1*
- *(A1) 29.32 - FABBRICAZIONE DI ALTRE MACCHINE PER L'AGRICOLTURA E LA SILVICOLTURA* à *29.32.1*
- *(A1) 29.22 - FABBRICAZIONE DI MACCHINE E APPARECCHI DI SOLLEVAMENTO E MOVIMENTAZIONE* à *29.22.1*
- *(A1) 33.40 - FABBRICAZIONE DI ATTREZZATURE FOTOGRAFICHE* à *33.40.5*
- *(A1) 36.11 - FABBRICAZIONE DI SEDIE E SEDILI* à *36.11.1*
- *(A1) 45.45 - LAVORI DI COMPLETAMENTO DEGLI EDIFICI* à *45.45.2*
- *(A1) 45.4 –LAVORI DI COMPLETAMENTO DEGLI EDIFICI* à *45.45.2*
- *(A1) 50.40 – COMMERCIO MANUTENZIONE E RIPARAZIONE DI MOTOCICLI, ACCESSORI E PEZZI DI RICAMBIO'* à *50.40.1*
- *(A1) 51.51 - COMMERCIO ALL'INGROSSO DI COMBUSTIBILI SOLIDI, LIQUIDI, GASSOSI E DI PRODOTTI DERIVATI* à *51.51.3*
- *(A1) 51.23 - COMMERCIO ALL'INGROSSO DI ANIMALI VIVI* à *51.23.2*
- *(A1) 51.54 COMMERCIO ALL'INGROSSO DI ARTICOLI IN FERRO, DI APPARECCHI E*

- *(A1) 51.51 - COMMERCIO ALL'INGROSSO DI COMBUSTIBILI SOLIDI, LIQUIDI, GASSOSI E DI PRODOTTI DERIVATI* à *51.51.3*
- *(A1) 51.23 - COMMERCIO ALL'INGROSSO DI ANIMALI VIVI* à *51.23.2*
- *(A1) 51.54 COMMERCIO ALL'INGROSSO DI ARTICOLI IN FERRO, DI APPARECCHI E ACCESSORI PER IMPIANTI IDRAULICI E DI RISCALDAMENTO* à *51.54.4*
- *(A1) 51.64 - COMMERCIO ALL'INGROSSO DI MACCHINE E DI ATTREZZATURE PER UFFICIO* à *51.64.1*
- *(A1) 51.42 - COMMERCIO ALL'INGROSSO DI CAPI DI ABBIGLIAMENTO E DI CALZATURE* à *51.42.5*
- *(A1) 52.24 – COMMERCIO AL DETTAGLIO DI PANE, PASTICCERIA E DOLCIUMI* à *50.40.1*
- *(A1) 52.46 - COMMERCIO AL DETTAGLIO DI FERRAMENTA, COLORI E VERNICI, VETRO*
- *(A1) 71.40 - NOLEGGIO DI BENI PER USO PERSONALE E DOMESTICO N..C..A.* à *71.40.2*
- *(A1) 72.60 - ALTRE ATTIVITA' CONNESSE ALL'INFORMATICA* à *72.60.2*
- *(A1) 74.30 - COLLAUDI E ANALISI TECNICHE* à *74.30.1*
- *(A1) 91.3 - ATTIVITA' DI ALTRE ORGANIZZAZIONI ASSOCIATIVE* à *91.33.0*

Il citato **criterio di prediligere l'attribuzione del codice al massimo dettaglio (5 cifre)** è stato spesso adottato anche in fase di arricchimento del *dizionario elaborabile* di ACTR con risposte *empiriche*, purché emergesse dai dati rilevati nel corso di precedenti indagini che l'informazione mancante sottintendesse con una probabilità molto elevata una casistica piuttosto che un'altra.
Per chiarire con un esempio, una risposta del tipo:

'Attività immobiliare'

che non specifica se esercitata su beni propri o per conto terzi, è stata associata al codice corrispondente alla attività immobiliare per conto terzi (codice 70.31.0). In questo modo è stato possibile assegnare un codice a cinque digit piuttosto che uno a soli due digit (codice 70).

La struttura del dizionario risulta quindi la seguente:

- Colonna 1-2 campo filtro
- Colonna 4-10 codice numerico
- Colonna 15-18 flag identificativo
- Colonna 20-100 descrizione attività economica.

Per il significato del **campo filtro** si rimanda al paragrafo 3.2.2

I **codici numerici** sono associati alle voci della Classificazione ufficiale delle Attività Economiche 1991 e sono cosi suddivisi:

60 voci a due cifre corrispondenti alle **divisioni** della Classificazione,
222 voci a 3 cifre corrispondenti ai **gruppi** della Classificazione,
512 voci a 4 cifre corrispondenti alle **classi** della Classificazione,
874 voci a 5 cifre corrispondenti alle **categorie** della Classificazione .

Ciascuna attività economica viene così codificata generalmente con un numero di cinque cifre, delle quali l'ultima è separata da un punto dalle due precedenti, e queste sono a loro volta separate da un punto dalle prime due.

Il **flag identificativo**, invece, è stato pensato soltanto a fini documentatitivi, per facilitare il lavoro di successivo aggiornamento del dizionario; tale flag indica infatti la fonte dei testi precodificati inseriti. In particolare:

- le descrizioni che presentano come flag (A1) sono le *divisioni, i gruppi, le classi o le categorie della classificazione* (sono in totale 1901);
- le descrizioni che presentano come flag (b1) sono le *Note esplicative delle attività comprese nelle varie classi e categorie di attività economica della classificazione* (sono in totale 5673);
- le *empiriche* con (A*) (b*) (e*) (sono in totale 403) provengono:

⇨ in parte dalla duplicazione della classe 18 (Confezione di articoli di vestiario; preparazione e tintura pellicce) con la sostituzione della parola *produzione* al posto di *confezione,*

⇨ in parte dalla duplicazione delle classi 50.20.4 (Riparazione e sostituzione di pneumatici) con la sostituzione della parola *pneumatici* in *gomme per auto,* e 50.30.0 (commercio di parti ed accessori di autoveicoli) con la sostituzione della parola *gomme* in *pneumatici,*

⇨ alcune (b*) provengono anche dalla duplicazione di alcune voci della classe 01 (Agricoltura caccia e silvicoltura) con la sostituzione della parola *produzion*e al posto di *coltivazione* prevista dalla classificazione,

⇨ alcune (e*) provengono anche dalla duplicazione di alcune voci della classe 55.40.1 (bar e caffè) con la sostituzione della parola *caffè* al posto di *bar;*

- le *empiriche* con (ei) provengono dall'esame da parte degli esperti del CUE dei diversi file

- ⇨ alcune (e*) provengono anche dalla duplicazione di alcune voci della classe 01 (Agricoltura caccia e silvicoltura) con la sostituzione della parola *produzione* al posto di *coltivazione* prevista dalla classificazione,
- ⇨ alcune (e*) provengono anche dalla duplicazione di alcune voci della classe 55.40.1 (bar e caffè) con la sostituzione della parola *caffè* al posto di *bar;*
- le *empiriche* con (ei) provengono dall'esame da parte degli esperti del CUE dei diversi file sottoposti a codifica automatica (sono in totale 15225);
- le *empiriche* con (e0) contengono i codici numerici e la descrizione dell'attività tra parentesi quadra (sono in totale 936);
- le *empiriche* con (ep) provengono dalla fase di validazione del Censimento della Popolazione (CP) (sono in totale 579);
- le *empiriche* con (pi) sono state ricavate dal confronto tra la classificazione Ateco e Prodcom. (sono in totale 1058);
- le *empiriche* con (nc) sono quelle relative a descrizioni troppo generiche che non consentono cioè l'attribuzione di un unico codice, pertanto risultano non codificabili (sono in totale 65);
- le *empiriche* con (na) sono alcune descrizioni che provengono dalla NACE EMENDATA e armonizzata con le altre nomenclature. (sono in totale 25).

(I valori numerici qui riportati sono coerenti con la versione più aggiornata del *reference* file)

Le **descrizioni attività economica** comprendono sia i testi opportunamente rielaborati, come detto sopra, della classificazione ufficiale, che le risposte testuali fornite nelle varie indagini che sono state oggetto di studio nella fase di addestramento.

Infine, le soglie di accettazione per la misura di somiglianza (vedi capitolo 2) adottate sono le seguenti:

Smin à6

Smax à8

Δ S à0,2

3.2 Attività realizzate

La priorità è stata ovviamente data alla **codifica automatica dei dati dell'indagine Long Form,** man mano che questi arrivavano dalla registrazione**.**

Nello stesso tempo **è stato arricchito il *dizionario elaborabile*,** utilizzando come feedback i risultati di ciascun passaggio di codifica automatica.

Scendendo nel dettaglio, per ciascun passaggio di codifica automatica, ACTR produce i seguenti risultati:

- **'Unici'** à testi cui il sistema è riuscito ad attribuire un singolo codice,
- **'Multipli'** à testi per i quali il sistema, non potendo assegnare un singolo codice con un elevato livello di affidabilità, ha individuato una serie di codici (massimo 5) tra i quali potrà essere selezionato quello corretto,
- **'Possibili'** à analoghi ai 'Multipli', ma con un livello di affidabilità ancora inferiore,
- **'Falliti'** à testi per i quali il sistema non riesce ad individuare alcun codice.

Utilizzare questi risultati per arricchire il dizionario e, più in generale, l'ambiente di codifica significa:

- laddove il sistema sia riuscito ad attribuire il codice ('Unici'), verificarne la correttezza e, in caso di errore, aggiornare il dizionario ed i file di sistema;
- in tutti i casi per i quali il sistema non è riuscito ad attribuire un singolo codice ('Multipli', 'Possibili' e 'Falliti') verificare se tale fatto sia dovuto a:
- ⇨ carenza del dizionario à conseguentemente inserire nello stesso la risposta *empirica* o il/i sinonimi mancanti;
- ⇨ non sufficiente 'informatività' della risposta per essere codificata à nessuna attività richiesta.

Le due attività sono state svolte parallelamente, anche se si è pensato di dare particolare impulso all'accrescimento dimensionale del dizionario al fine di elevare in tempi ristretti il tasso di codifica.

Relativamente alla verifica della correttezza degli 'Unici', come già detto, il sistema distingue a seconda che la risposta fornita dall'intervistato sia perfettamente identica ad uno dei testi contenuti nel dizionario (abbinamento esatto) oppure presenti un elevato grado di 'similarità' rispetto ad uno di questi (abbinamento parziale) .
A regime, sarebbe sufficiente verificare la correttezza soltanto di questi ultimi casi, in quanto si presuppone l'assoluta attendibilità del dizionario; tuttavia, trattandosi della prima applicazione del sistema su una variabile così complessa come l'attività economica, da un certo momento in poi si è ritenuto opportuno effettuare la verifica a tappeto di tutti i casi codificati automaticamente.

3.2.1 Assetto organizzativo nell'ambito del gruppo di lavoro

Organizzativamente si è proceduto in tal modo:

- il sistema di codifica è stato installato presso il servizio CUE, dove viene effettuato il vero e

3.2.1 Assetto organizzativo nell'ambito del gruppo di lavoro

Organizzativamente si è proceduto in tal modo:

- il sistema di codifica è stato installato presso il servizio CUE, dove viene effettuato il vero e proprio passaggio di codifica di ciascuna tranche di dati;
- presso lo stesso servizio CUE viene effettuata l'analisi dei risultati di ciascun passaggio di codifica finalizzata ad individuare gli interventi da effettuare sull'ambiente di codifica al fine di elevare la percentuale di successi ed il livello di qualità dei risultati;
- l'unità MPS/C recepisce dal CUE le osservazioni sugli interventi migliorativi, ne effettua un'analisi di congruenza rispetto all'ambiente di codifica ed apporta le modifiche al sistema. Si precisa che l'analisi di congruenza non riguarda la validazione sulle operazioni di codifica effettuate dal CUE (in quanto è lì che risiede la competenza sulla classificazione) bensì la soluzione tecnica per l'effettuazione dell'intervento e la verifica che, a seguito di questo, sia garantito il mantenimento di un ambiente generalizzato rispetto alle possibili applicazioni a diverse indagini;
- l'unità MPS/C mette in linea per il CUE l'ambiente applicativo aggiornato.

3.2.2 Sviluppo di software integrativo

Nel corso delle attività descritte, ci si è resi conto che le performance del sistema sarebbero state ottimizzate se fossero state potenziate due funzioni inerenti:

1. il fatto di consentire al sistema di effettuare la ricerca nel proprio dizionario sulla base di un primo livello di gerarchia, corrispondente alle principali sezioni di attività economica (es. Attività manifatturiere, Produzione di gas, energia, Commercio,….); ciò avrebbe ridotto gli eventuali errori di codifica ed evitato che i casi 'Multipli' contemplassero testi molto simili tra di loro, magari diversi in una singola parola (produzione/commercio) che è però quella discriminante (chiameremo tale funzione *'Attribuzione di filtro'*) ;
2. l'individuazione automatica del codice corretto nell'ambito di un sottoinsieme di 'Multipli' che presentano particolari caratteristiche (chiameremo tale funzione *'Gestione dei Multipli'*).

A tal fine sono stati sviluppati due moduli che realizzano le citate funzioni e che sono gestibili direttamente dall'utente finale tramite apposite interfacce.

Attribuzione di filtro

Il sistema di codifica automatica ACTR consente di effettuare l'abbinamento tra il testo da codificare e quelli del dizionario, limitando la ricerca soltanto in alcuni rami della classificazione.
Questo si realizza prevedendo nei record da codificare un *filtro* (flag) che individui il/i rami nei quali effettuare la ricerca; ovviamente lo stesso flag deve essere previsto in corrispondenza dei record del dizionario.
Il modo ottimale di costruire un filtro sarebbe quello di strutturare i quesiti del modello di rilevazione in modo tale che il filtro possa essere ottenuto dalla risposta ad un singolo quesito. Per chiarire, semplificando:

Q1) In quale dei seguenti rami si svolge la sua attività economica?

1. *Attività manifatturiere*
2. *Produzione di gas ed energia,*
3. *Commercio*
4. *Etc*

Q2) Può descrivere più in dettaglio il suo settore di attività economica?

Nel dizionario utilizzato da ACTR, a tutti i testi inerenti le attività manifatturiere corrisponderà il filtro '1', a quelli della produzione di gas '2', etc..
Dalla risposta al quesito 1 si individuano quindi il/i rami della classificazione nell'ambito dei quali il sistema ricerca il codice da abbinare.

Dal momento che purtroppo però il modello di rilevazione dell'indagine Long Form non prevedeva il quesito sull'ATECO strutturato in questo modo, si è pensato di **effettuare un'analisi delle risposte al quesito a testo libero preliminare rispetto alla fase di codifica, al fine di attribuire il filtro sulla base della presenza/assenza di alcuni termini (e dei relativi sinonimi).**

E' quindi stata progettata e sviluppata un'applicazione che analizza le risposte testuali fornite dal rispondente in modo da individuare eventuali parole discriminanti, sulla base delle quali viene generata la *variabile filtro* da fornire al sistema di codifica.

E' quindi stata progettata e sviluppata un'applicazione che analizza le risposte testuali fornite dal rispondente in modo da individuare eventuali parole discriminanti, sulla base delle quali viene generata la *variabile filtro* da fornire al sistema di codifica.
Si premette che i filtri definiti non sono tali da coprire tutte le classi dell'Ateco, ma sono soltanto quelli che consentono di individuare parole che delimitino la ricerca del testo entro certe classi in maniera molto circostanziata; per meglio comprendere, non è stato possibile attribuire alcun filtro alla classe 45 perché nessuna delle parole usate ricorrentemente in questa classe consentiva di circoscrivere rigidamente la ricerca all'interno di questa stessa (es. costruzioni, edile, edilizia. ecc.).
I filtri impostati e le parole considerate discriminanti sono riportati nella seguente tabella:

Tabella 2 –Filtri impostati per l'indagine Long Form

Filtro	**Tipologia attività**	**Da codice … (compreso)**	**A codice … (compreso)**
10	ESTRAZIONE	10	14.50.3
20	PRODUZIONE	15	36.63.6
20	" "	40	40.3
40	INTERMEDIARI DI COMMERCIO CCCOMMERCORAPPRESENTANTI	51.1	51.19
50	COMMERCIO AMBULANTE	52.6	52.63.5
60	COMMERCIO	50	50.10
60	" "	50.30	50.40.2
60	" "	50.50	50.50.9
60	" "	51	51.00.0
60	" "	51.2	52.50.4
70	ALBERGHI BAR RISTORANTI	55	55.52
80	ASSICURAZIONI (NON OBBLIGATORIE)	66	67.20.2

L'individuazione di tali parole discriminanti non è stata cosa banale, in quanto spessissimo la presenza di altre parole nella frase costituisce un'eccezione alla regola di attribuzione del *filtro*; il caso tipico è quello di risposte che contemplano doppie attività, quali ad esempio 'produzione e commercio'. In tal caso, mentre alla parola 'commercio' verrebbe associato il *filtro* corrispondente alla classe ATECO del commercio, se associata ad attività produttive (individuate dalla parola produzione e dai suoi sinonimi) le deve essere abbinato un *filtro* diverso.
Si è tentato quindi di definire delle *regole di prevalenza* di una certa tipologia di attività rispetto ad altre; tuttavia anche queste regole non possono essere rigide, in quanto, tornando all'esempio della 'produzione' rispetto al 'commercio', attività produttive quali 'la stagionatura del formaggio', se abbinate al commercio, smentiscono la citata *regola di prevalenza* della produzione rispetto al commercio.

E' stato costituito un elenco comprendente le parole discriminanti, con il relativo *filtro*, e le coppie di parole (parola discriminante **+** parola eccezione), cui viene fittiziamente attribuito il *filtro* '00'.
Tale elenco (cfr. elenco in Allegato 2) risulta ad oggi costituito da **339** voci, a conferma della citata complessità dell'applicazione.

Operativamente, la procedura è stata sviluppata utilizzando ACTR stesso.
E' stato definito un ambiente di codifica del quale, l'elenco in Allegato 2 costituisce il *reference* di ACTR ('*reference del filtro*') che attribuisce come codice il *filtro*; tale ambiente condivide le stesse regole di *parsing* dell'Ateco (in tal modo non è stato necessario inserire nel '*reference del filtro*' i sinonimi delle parole considerate), ma utilizza parametri soglia più bassi rispetto all'Ateco, in modo tale che sia spesso sufficiente la presenza di una sola parola nella frase per l'attribuzione del codice (*filtro*).

La procedura viene utilizzata per due finalità:

- attribuzione del *filtro* al file di input da codificare;
- attribuzione del *filtro* alle nuove *empiriche* da inserire nel dizionario.

La procedura viene utilizzata per due finalità:

- attribuzione del *filtro* al file di input da codificare;
- attribuzione del *filtro* alle nuove *empiriche* da inserire nel dizionario.

Nel caso dell'attribuzione del *filtro* al file di input da codificare la procedura è lineare in quanto consiste semplicemente nell'applicazione di ACTR. Le esperienze finora effettuate hanno dimostrato che il *filtro* viene attribuito in media al 45% dei record.

Nel caso invece delle nuove *empiriche* da inserire nel dizionario, essendo queste composte da testo e codice Ateco, il *filtro* deve essere in realtà attribuito in funzione del codice Ateco corrispondente al testo. Tuttavia è necessario verificare che questo *filtro* sia lo stesso rispetto a quello che sarebbe attribuito in base al testo, altrimenti si avrebbe l'inconveniente che un eventuale record da codificare uguale all'empirica incriminata si vedrebbe attribuito un *filtro* sulla base del testo diverso rispetto all'empirica inserita nel dizionario (il *filtro* diverso impedirebbe quindi l'attribuzione del codice).

La procedura realizza quindi i seguenti passaggi:

- attribuzione del *filtro* all'empirica da inserire nel dizionario sulla base del codice,
- attribuzione del *filtro* all'empirica da inserire nel dizionario sulla base del testo,
- confronto tra i filtri attribuiti secondo i due metodi,
- in caso di divergenza, analisi del caso ed eventuale aggiornamento del '*reference del filtro*' e del relativo ambiente di codifica.

(Per la descrizione tecnica della procedura e le indicazioni per il suo utilizzo, cfr. Allegato 1.).

Rispetto al **prossimo censimento**, è necessario specificare che è stata recepita l'indicazione inerente la strutturazione 'ad albero' del quesito sull'attività economica, in modo tale che il *filtro* possa essere ottenuto dalla risposta fornita ad un apposito quesito.

La variabile sarà infatti rilevata come segue:

1.2 In quale dei seguenti settori l'unità locale svolge l'attività principale?
1 Industria
2 Commercio
3 Altri servizi
Segue il quesito a testo libero.

L'informazione derivante dal quesito I.2 costituirà quindi il *filtro* ai fini dell'attribuzione del codice corretto.

Tabella 3 –Filtri impostati per il censimento

Filtro	**Tipologia attività**	**Da codice … (compreso)**	**A codice… (compreso)**
01	INDUSTRIA	05	45
02	COMMERCIO	50	52
03	ALTRI SERVIZI	55	74
"	"	80	99

Rispetto al *filtro* progettato per la Long Form, come è evidente, questo è meno dettagliato. Tuttavia si ritiene che tale impostazione sia sufficiente per prevenire gli eventuali errori di codifica che potrebbero verificarsi senza *filtro* e nel contempo non comporti un appesantimento del modello di rilevazione.

Nel corso del censimento, dunque, la procedura di attribuzione dei filtri finora descritta, che dovrà recepire l'impostazione corrispondente alla Tabella 3, potrà comunque essere utilizzata nei seguenti casi:

- per codificare i questionari nei quali si è rilevata una mancata risposta al quesito *I.2,*
- per produrre un termine di confronto rispetto ai casi codificati utilizzando come *filtro* la risposta al quesito *I.2*, al fine di effettuare un'analisi di qualità,
- per trattare i casi non codificati automaticamente utilizzando come *filtro* la risposta al quesito *I.2* e verificare se il mancato successo nell'attribuzione del codice possa essere dovuto ad un *filtro* sbagliato, ossia ad un errore di risposta al quesito *I.2*.

e verificare se il mancato successo nell'attribuzione del codice possa essere dovuto ad un *flusso* sbagliato, ossia ad un errore di risposta al quesito *I.2*.

Gestione dei 'Multipli'

Dall'analisi dei risultati della codifica sull'indagine Long Form, è stato rilevato che molto spesso tra i 'Multipli' è presente il codice corretto; ne consegue che questi stessi:

- in fase di indagine possano essere utilizzati dal codificatore manuale per l'attribuzione del codice,
- costituiscano una base informativa molto utile per l'arricchimento del dizionario.

Si è notato inoltre che, in funzione della logica di analisi testuale alla base di ACTR, nonché dell'ampliamento del dizionario, accade ricorrentemente che un testo da codificare finisca tra i 'Multipli' perché il sistema trova nel dizionario un certo numero di testi ai quali corrisponde spesso lo stesso codice, ma che sono molto simili tra di loro, cosicché che non possa essere superata la soglia ΔS (minima distanza tra testo 'vincente' (S_1) e successivo (S_2)), necessaria per individuare il 'vincente'.

E' stata quindi sviluppata una procedura che estrae dai 'Multipli' i casi corrispondenti alle seguenti specifiche (per comodità a ciascuna di queste specifiche è stato attribuito un flag):

- **M55** sta a significare che è stato assegnato il codice a 5 (o meno) digit comune a tutti e 5 i record di output
- **M54** sta a significare che è stato assegnato il codice corrispondente ai primi 4 digit comuni a tutti e 5 i record di output
- **MM** sta a significare che è stato assegnato il codice corrispondente ai casi nei quali ACTR produce soltanto 4 'Multipli' tutti con lo stesso codice, oppure soltanto 3 (si intende sempre codice a 5 digit).

(Per la descrizione tecnica della procedura e per le indicazioni sul suo utilizzo, cfr.. **Allegato 3**.).
In merito all'analisi di qualità dei risultati ottenuti da questa procedura si rimanda al paragrafo 3.3.4

Sono inoltre stati analizzati anche i risultati prodotti da ulteriori specifiche, quali:

- **M45** sta a significare che è stato assegnato il codice a 5 digit presente in 4 dei 5 record di output
- **MM1** sta a significare che è stato assegnato il codice corrispondente ai casi nei quali ACTR produce soltanto 2 'Multipli', entrambi con lo stesso codice.

Poiché i risultati così prodotti implicavano un livello di *accuratezza* inferiore a quello abitualmente ottenuto con ACTR sugli 'Unici', si è deciso di non mettere in produzione tali specifiche.
Come si vedrà in seguito, sono inoltre stati analizzati i risultati di una procedura analoga, ma applicata ai 'Possibili', per valutare l'opportunità di estrarre alcuni di questi e convogliarli negli 'Unici'.

3.2.3 Integrazione del dizionario ATECO con la variabile PRODCOM

Dall'analisi delle risposte fornite nell'indagine Long Form, ci si è inoltre resi conto che spesso queste fanno riferimento al particolare prodotto oggetto dell'attività dell'impresa. Per tale motivo si è ritenuto opportuno integrare il dizionario delle attività economiche con i testi della classificazione dei prodotti (PRODCOM).

A tal fine il gruppo di lavoro è stato integrato con esperti dei servizi **ARC e SSI** che non soltanto hanno fornito consulenze di merito sulle due classificazioni, ma hanno coadiuvato nell'attività di integrazione del dizionario.

Tale attività ha richiesto un intervento di rielaborazione dei testi PRODCOM (in parte manuale ed in parte automatizzato) finalizzato alla ricostruzione di frasi compiute, che comprendano attività e prodotto/i, rendendole il più possibile vicine al linguaggio parlato utilizzato dai rispondenti.

Volendo quantificare il lavoro effettuato, si specifica che ad oggi è stata effettuata un'analisi comparativa ATECO-PRODCOM, a seguito della quale sono stati introdotti nel dizionario ATECO circa 1000 testi.

Nel corso di questo lavoro, inoltre, sono emerse alcune **incoerenze** tra le due classificazioni, delle quali alcune, non di grande importanza (inerenti esclusivamente la quinta cifra ATECO), sono di facile risoluzione, mentre altre evidenziano interpretazioni sostanzialmente diverse tra le due classificazioni e richiedono un intervento che esula dalle competenze del gruppo di lavoro.
(A tale proposito si veda l'Allegato 4).

facile risoluzione, mentre altre evidenziano interpretazioni sostanzialmente diverse tra le due classificazioni e richiedono un intervento che esula dalle competenze del gruppo di lavoro.
(A tale proposito si veda l'Allegato 4).

3.3 Risultati ottenuti

Nei precedenti censimenti le operazioni di codifica delle descrizioni delle attività economiche hanno determinato notevoli aumenti dei tempi necessari per il completamento delle attività censuarie. L'attribuzione manuale del codice ATECO è, infatti, un'operazione molto complessa. Da un lato, l'elevato numero di codici presenti nella classificazione (874 categorie) rende difficoltosa la ricerca nel manuale dei codici corrispondenti alle descrizioni; dall'altro, le informazioni fornite dai rispondenti sono spesso generiche, incomplete o di difficile comprensione.
Proprio per ovviare a tali problemi, si è deciso di sperimentare la codifica automatica nell'ambito del Censimento Intermedio dell'Industria e dei Servizi (Long Form), al fine di adottarla poi nel Censimento Generale dell'Industria e dei Servizi 2001.

L'informatizzazione delle operazioni di codifica presenta numerosi vantaggi:

- minor aggravio di lavoro per i comuni;
- minor costo per la produzione dei dati;
- riduzione dei tempi di elaborazione;
- interpretazione delle descrizioni unica su tutto il territorio nazionale;
- livello più elevato di qualità;
- eliminazione degli errori di registrazione dei codici.

3.3.1 Impatto sull'indagine Long-Form

La sperimentazione della codifica automatica della variabile Ateco, nell'ambito del servizio CUE, è iniziata nel 1998 (Ferrillo A. 1999, 'Sperimentazione della codifica automatica dell'attività economica e della forma giuridica,') ed è stata applicata alla rilevazione Long-Form.
L'indagine Long Form è stata una rilevazione condotta per lista sulla base dell'Archivio statistico imprese attive '96 (ASIA96) e caratterizzata da questionari personalizzati e prestampati per la parte anagrafica e per quella relativa ai principali attributi strutturali dell'impresa.
Nel questionario erano presenti tre descrizioni che necessitavano di una codifica: 'la localizzazione', 'l'attività economica' e 'la forma giuridica'.
L'utilizzo di questionari personalizzati, in base ai quali le imprese hanno variato le descrizioni solo in caso di informazione inesatta o modificata, ha consentito di ridurre notevolmente i tempi di codifica.
Infatti delle 319.000 imprese rispondenti solo il 17% ha indicato una modifica e/o inesattezza alla descrizione dell'attività economica e il 5% ha indicato l'attività economica secondaria per un totale di 70.236 record da codificare (vedi diagramma a torta).

Tabella 4 -- Distribuzione record Long-Form tra attività principale e secondaria

	Valori assoluti	% sul totale record
Attività principale	54.129	17,0
Attività secondaria	16.107	5,0
Totale record	70.236	22,0

I 70.236 record sono stati codificati man mano che venivano inviati dalla registrazione. Ad ogni invio, sulla base dei risultati ottenuti, si è cercato di incrementare il numero di *empiriche* (ovvero descrizioni dell'attività economica) del dizionario ACTR. Nella Tabella 5 è riportata la percentuale dei record codificati per tipo di invio. Si nota un incremento della percentuale dei record codificati per invii successivi. L'alta percentuale del primo invio è determinata dal fatto che tale lotto è stato utilizzato nella fase iniziale di addestramento dell'applicazione di codifica automatica ed arricchimento della relativa base informativa.

La media dei record codificati automaticamente è stata dunque del **58,8%** del totale.

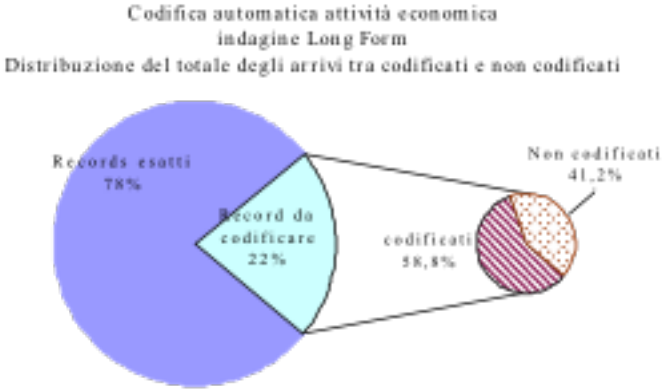
Tabella 5 -- Situazione per numero di invio dalla registrazione

INVIO	Tot. RK da codificare	*Codificati*	*Non codificati*	*%*

Tabella 5 -- Situazione per numero di invio dalla registrazione

INVIO	Tot. RK da codificare	*Codificati*	*Non codificati*
		%	
T1	1.882	*65,1*	*34,9*
T2	9.835	*54,2*	*45,8*
T3	7.319	*57,6*	*42,4*
T4	3.825	*58,2*	*41,8*
T5	7.677	*53,3*	*46,7*
T6	3.370	*49,8*	*50,2*
TB	326	*57,4*	*42,6*
T7	1.460	*60,0*	*40,0*
T8	1.462	*57,5*	*42,5*
T9	5.522	*59,7*	*40,3*
T10	8.023	*64,1*	*35,9*
T11	5.826	*62,7*	*37,3*
T12	4.156	*61,5*	*38,5*
T13	3.122	*61,2*	*38,8*
T14	2.709	*63,3*	*36,7*
T15	1.470	*62,6*	*37,4*
Tn	1.127	*57,1*	*42,9*
Bz	1.125	*67,6*	*32,4*
Totale	**70.236**	***58,8***	***41,2***

Dall'analisi sopra descritta emerge che migliori performance di ACTR si possono ottenere solo lavorando ed integrando il dizionario con descrizioni mancanti e frasi corrispondenti al linguaggio comunemente usato dai rispondenti nel corso delle rilevazioni statistiche.



3.3.2 Attività finalizzate all'arricchimento della base informativa

Come evidenziato nel precedente paragrafato, miglioramenti del tasso di codifica, sono possibili principalmente incrementando il numero di descrizioni presenti nel dizionario ed ottimizzando la strategia di *parsing*.
La sperimentazione del software ACTR si è mossa in tale direzione. Al fine di individuare nuove descrizioni di attività economiche, conformi al linguaggio utilizzato dai rispondenti, si sono analizzati i seguenti file:

- 54.129 descrizioni dell'attività economica principale dell'indagine Long Form;
- un campione di 64.111 descrizioni dell'attività economica delle unità locali Long Form;
- 843.325 descrizioni provenienti da un'estrazione per codice contributivo statistico dal file fornito dall'INPS per l'aggiornamento dell'archivio ASIA (Archivio Statistico Imprese Attive).

Sono stati estratti da quest'ultimo file i codici inerenti le seguenti attività dei servizi:

- **112** studi e laboratori fotografici;
- **115** trasporti;
- **118** servizi di pulizia;
- **ramo 6** credito, assicurazioni, servizi tributari appaltati;
- **ramo 7** commercio, servizi, professioni ed arti.

- **118** servizi di pulizia;
- **ramo 6** credito, assicurazioni, servizi tributari appaltati;
- **ramo 7** commercio, servizi, professioni ed arti.

Nella Tabella 6 è riportata la distribuzione percentuale dei risultati ACTR a diverse date corrispondenti a successivi incrementi del dizionario.

Tabella 6 -- Evoluzione dei risultati della codifica Ateco sui dati Long-Form senza filtro

	Anno 2001								
	09/02	26/02	13/03	26/03	24/04	30/05	13/06	12/07	06/08
FALLITI	6,6	6,8	6,3	6,1	6,0	5,7	5,2	5,1	5,1
MULTIPLI	11,8	11,6	11,7	11,4	10,9	10,2	10,2	10,3	9,5
MULTIPLI con 5 cod Ateco uguali	2,0	0,6	0,7	0,6	0,7	0,5	0,5	0,6	0,6
POSSIBILI	12,1	11,6	10,6	10,2	9,5	8,1	7,9	7,7	7,6
POSSIBILI con 5 cod Ateco uguali	1,0	0,5	0,5	0,5	0,6	0,4	0,4	0,5	0,5
UNICI con cod ATECO < 5 digit e 'n.c.'	2,0	1,9	2,1	2,2	2,4	2,4	2,5	2,4	2,5
UNICI	64,6	67,0	68,1	69,0	70,0	72,6	73,3	73,4	74,2
Totale testi del dizionario	**19.436**	**20.026**	**20.760**	**21.265**	**22.256**	**23.098**	**23.707**	**24.421 6**	**24.95**

Come si può osservare, gli output prodotti da ACTR sono stati ulteriormente elaborati; dai 'Multipli' e dai 'Possibili' sono stati, infatti, scorporati i 'Multipli' e 'Possibili' con cinque codici Ateco uguali (M55, P55) e dagli 'Unici' quelli con codici < di cinque digit e 'n.c.'.

La riga 'Unici' della Tabella 6 evidenzia un notevole incremento della percentuale degli 'Unici' all'aumentare delle *empiriche* inserite nel dizionario.

Analisi di qualità dei risultati

L'analisi di qualità dei risultati della codifica automatica (il livello di *accuratezza*) dovrebbe essere effettuata sottoponendo all'analisi di codificatori esperti i casi codificati prodotti dal sistema a seguito di un *indirect match* , ossia agli 'Unici' con punteggio minore di 10, in quanto, dando per scontato la correttezza del dizionario, si presume che gli 'Unici' con punteggio uguale a 10 siano corretti.

Tale operazione è stata effettuata nell'ambito del gruppo di lavoro ogni qual volta l'ambiente applicativo veniva aggiornato con modifiche ed integrazioni sostanziali; non si dispone quindi di un valore univoco che esprima questo parametro.

Tuttavia, l'*accuratezza* è stata stimata indirettamente, mettendo a confronto il codice assegnato da ACTR e quello ritrovato nei dati definitivi Long Form, consci del fatto che su questi ultimi era stato già effettuato il passaggio di controllo e correzione che non si basa sull'analisi testuale, ma sulla coerenza del codice con altre variabili rilevate.

In particolare, sono stati considerati non soltanto gli 'Unici', ma anche i 'Possibili' e 'Multipli' con 5 codici uguali, per testare ulteriormente la possibilità di inserire tali casi tra gli 'Unici'.

Nella Tabella 7 sono riportati i confronti tra il codice assegnato da ACTR ed il codice Long Form. Per il 24% dei casi il codice ATECO risulta differente e per l'8% dei casi non corrisponde neanche un digit. Sui 3276 record che non corrispondono neanche ad un digit è stata effettuata una verifica manuale per stabilire il codice esatto. I risultati sono riportati nella Tabella 8. Si ha che solo nel 9% dei casi ACTR assegna un codice errato. Nella Tabella 9, infine, è stata riportata la distribuzione per tipo di errore.

Tabella 7 -- Confronto Ateco ACTR con Ateco Long-Form

Unici

	Corrispondenze	*%*	*Non corrispondenze*	*%*	
ATECO 5 DIGIT	30514	76,0	9662	24,0	100
ATECO 4 DIGIT	31617	78,7	8559	21,3	100
ATECO 3 DIGIT	33189	82,6	6987	17,4	100

ATECO 5 DIGIT	30514	76,0	9662	24,0	100
ATECO 4 DIGIT	31617	78,7	8559	21,3	100
ATECO 3 DIGIT	33189	82,6	6987	17,4	100
ATECO 2 DIGIT	35300	87,9	4876	12,1	100
ATECO 1 DIGIT	36900	91,8	3276	8,2	100

Tabella 8 – Verifica delle non corrispondenze

	valori assoluti	%
Esatti ACTR	2981	91,0
Errati ACTR	295	9,0
Totale	3276	100,0

Tabella 9 -- Distribuzione errati per tipo di errore

	valori assoluti	%sul totale
Errati per testo generico	164	5,0
Per dizionario ancora carente	98	3,0
Errori di ortografia	33	1,0
Totale	295	9,0

3.3.3 Impatto su altri campioni (INPS)

Nella Tabella 10 è riportata la distribuzione percentuale di ACTR per il file I.N.P.S. a diverse date. Come si può osservare le percentuali di codifica degli 'Unici' del file I.N.P.S., risultano più basse rispetto a quelle ottenute dalla Long-Form (69,3%, 74,2%); ciò è dovuto principalmente alle descrizioni contenute nel file non sempre chiare; spesso, oltre alla descrizione dell'attività svolta, il testo contiene, per il settore *alberghi e ristoranti* il nome del gestore (es: *albergo Serena; ristorante da Peppino ecc.),* oppure riferimenti legislativi inerenti quella particolare categoria di lavoratori (*posizione per iscritti enpals, iscrizione provvisoria, personale optante inpdap, legge n°… del …, ecc.);* o ancora l'ubicazione dell'impresa *(bar via Nazionale 157, ecc.),* e il settore del *commercio* spesso risulta senza la specifica *ingrosso e/o dettaglio*.

Inoltre si rileva una diminuzione di codifica degli 'Unici' dell'1,1% tra il 13/6 e il 12/7. Con l'applicazione utilizzata a giugno, infatti, si era rilevato che rientravano tra gli 'Unici' alcuni casi che non dovevano essere codificati; sono quindi state effettuate alcune modifiche al dizionario ed al *parsing*, tali da far rientrare questi casi nei non codificati, elevando quindi il livello di *accuratezza* del sistema. Ci si riferisce, per esempio, ai record, contenenti la descrizione ' *prop. Fabbr.' (proprietario fabbricato*), che non consentiva l'attribuzione di un codice Ateco corretto; poiché l'abbreviazione '*fabbr*' veniva erroneamente trasformata in '*fabbricazione*', determinando un incremento degli 'Unici', ma con codice errato.

Problema analogo si verificava per le descrizioni inerenti alcuni prodotti, che erano presenti nel dizionario soltanto nell'ambito delle divisioni della produzione e non anche del commercio all'ingrosso e/o dettaglio.

Tabella 10 -- Evoluzione dei risultati della codifica ateco I.N.P.S. con l'applicazione del filtro

	24/04/01	13/06/01	12/07/01	06/08/01
	Valori assoluti			
FALLITI	80.377	74.263	79.274	76.122
MULTIPLI	103.279	101.187	106.261	83.967
MULTIPLI con 5 cod Ateco	3.120	4.656	5.060	1.409

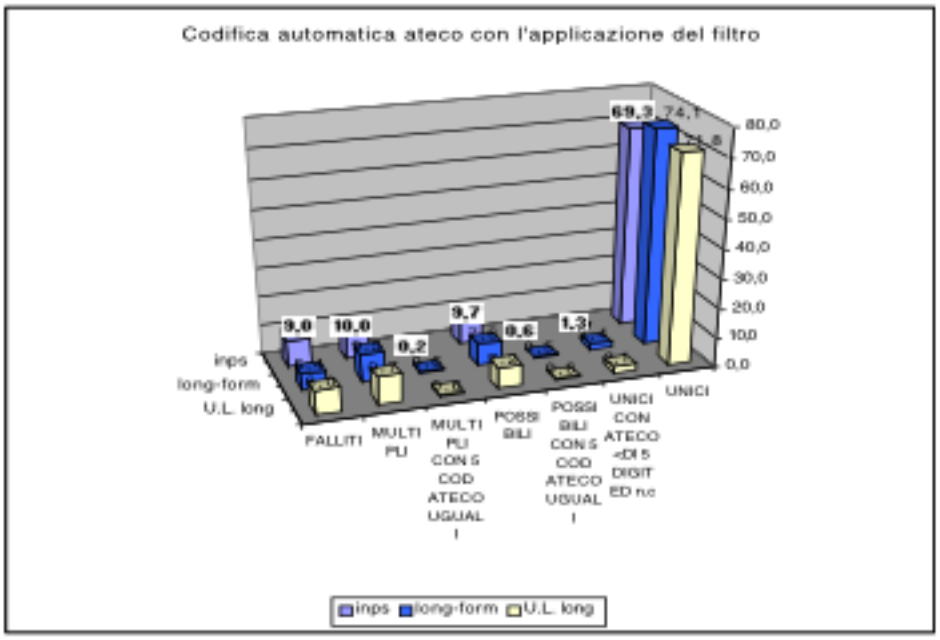
		Valori assoluti		
FALLITI	80.377	74.263	79.274	76.122
MULTIPLI	103.279	101.187	106.261	83.967
MULTIPLI con 5 cod Ateco uguali	3.120	4.656	5.060	1.409
POSSIBILI	112.003	84.162	83.491	81.585
POSSIBILI con 5 cod Ateco uguali	5.226	4.846	5.060	5.228
UNICI con Ateco < 5 digit e 'n.c.'	10.038	10.432	9.277	10.680
UNICI	529.302	563.799	554.921	584.354
	843.345	843.345	843.345	843.345
		%		
FALLITI	9,5	8,8	9,4	9,0
MULTIPLI	12,2	12,0	12,6	10,0
MULTIPLI con 5 cod Ateco uguali	0,4	0,6	0,6	0,2
POSSIBILI	13,3	10,0	9,9	9,7
POSSIBILI con 5 cod Ateco uguali	0,6	0,6	0,6	0,6
UNICI con Ateco < 5 digit e 'n.c,'	1,2	1,2	1,1	1,3
UNICI	62,8	66,9	65,8	69,3

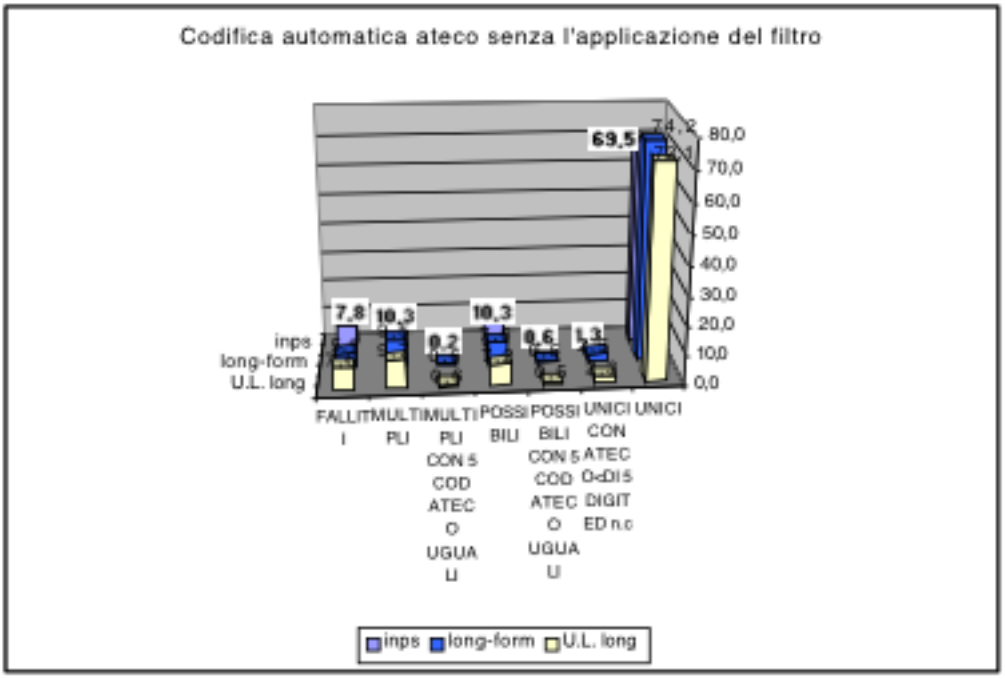
3.3.4 Analisi dei risultati della procedura 'Attribuzione di filtri'

A tutti i record dei file presi in esame (Long Form, U.L. long, I.N.P.S.) sono stati assegnati i codici *filtro* tramite l'apposita procedura di '*Attribuzione dei filtri*'. Detti file sono stati sottoposti ad ACTR utilizzando la procedura con *filtro* e senza *filtro*.

Tabella 11-- Confronto con e senza filtro dei file presi in esame

	Filtro			***No filtro***		
	I.N.P.S.	*LongForm*	*U.L. long*	*I.N.P.S.*	*LongForm*	*U.L. long*
	Valori assoluti			Valori assoluti		
FALLITI	76.122	2.901	5.109	65.675	2.752	4.865
MULTIPLI	83.967	5.142	6.236	86.891	5.149	6.188
MULTIPLI con 5 cod Ateco uguali	1.409	311	366	1.389	310	364
POSSIBILI	81.585	4.126	4.639	86.778	4.125	4.646
POSSIBILI con 5 cod Ateco uguali	5.228	279	366	5.130	267	352
UNICI con Ateco < 5 digite ' n.c.'	10.680	1.273	1.356	11.319	1.350	1.490
UNICI	584.354	40.097	46.039	586.163	40.176	46.206
	843.345	54.129	64.111	843.345	54.129	64.111
		%			%	
FALLITI	9,0	5,4	8,0	7,8	5,1	7,6
MULTIPLI	10,0	9,5	9,7	10,3	9,5	9,7
MULTIPLI con 5 cod Ateco uguali	0,2	0,6	0,6	0,2	0,6	0,6
POSSIBILI	9,7	7,6	7,2	10,3	7,6	7,2
POSSIBILI con 5 cod Ateco uguali	0,6	0,5	0,6	0,6	0,5	0,5
UNICI con Ateco < 5 digit e 'n.c.'	1,3	2,4	2,1	1,3	2,5	2,3
UNICI	69,3	74,1	71,8	69,5	74,2	72,1





Dalla Tabella 11 si può osservare che utilizzando la procedura con *filtro*, si riscontra una riduzione di codifica 'Unici' di circa 0,2%.
Dal punto di vista della qualità dei dati sono stati analizzati i record risultanti 'Unici' in una applicazione e non 'Unici' nell'altra e gli 'Unici' in entrambe le applicazioni, ma con codici diversi (Tabella 12).

Tabella 12

	Unici con filtro	Unici senza filtro
Rk assegnati a file di output diversi	39	140
Rk con codici diversi	23	23

I 39 rk 'Unici' con *filtro*, senza l'applicazione del *filtro* vengono smistati per la maggior parte nei 'Multipli'; i 140 rk 'Unici' senza *filtro*, con l'applicazione del *filtro* vengono smistati nei 'Falliti' e 'Possibili'. Nella Tabella 13 è riportata la distribuzione degli 'Unici' errati con codice diverso, gli errati con *filtro* pari a 60,9% sono dovuti principalmente ad un'attribuzione *filtro* errato, quelli senza *filtro*, invece per testo generico (es: commercio di un determinato prodotto senza la specifica di ingrosso e/o dettaglio).

Tabella 13 - Analisi 'Unici' con e senza filtro con codici diversi

	totale	errati	Esatti

senza la specifica di ingrosso e/o dettaglio).

Tabella 13 - Analisi 'Unici' con e senza filtro con codici diversi

	totale	errati	Esatti
Unici	*valori assoluti*		
Con filtro	23	14	9
Senza filtro	23	13	10
	%		
con filtro	100	60,9	39,1
Senza filtro	100	56,5	43,5

Dall'analisi sopra descritta risulta evidente la fragilità dell'applicazione dell'attribuzione dei filtri, nonché emerge la necessità di integrare il dizionario con le descrizioni mancanti e di rivedere il dizionario aggiungendo al testo sempre la parola '*produzione*', '*commercio ingrosso', 'commercio dettaglio'* (es: '*porte e finestre in legno*' con codice della produzione deve essere corretto in '*produzione porte e finestre in legno'*), onde evitare l'attribuzione del codice Ateco errato per le descrizioni '*commercio porte e finestre in legno*' (descrizione senza la specifica di ingrosso e/o dettaglio)

Infine è riportata un'analisi degli 'Unici' sulla base del punteggio assegnato da ACTR. Il punteggio assegnato agli 'Unici' è compreso tra 8 e 10; assegna dieci quando il testo da codificare, dopo le eventuali trasformazioni (*parsing*), risulta uguale a quello contenuto nel dizionario informatizzato.

Tabella 14 – Frequenza punteggio 'Unici' con filtro e senza filtro dei tre data-set presi in esame

Frequenza punteggio Unici Long-Form

	Con filtro		*senza filtro*	
Punteggio	*Valori assoluti*	%	*Valori assoluti*	%
8	2991	7,5	3000	7,5
9	5385	13,4	5388	13,4
10	31721	79,1	31788	79,1
Totale	40097	100,0	40176	100,0

Frequenza punteggio Unici I.N.P.S.

8	49036	8,4	49663	8,5
9	54418	9,3	55368	9,4
10	480900	82,3	481132	82,1
Totale	584354	100,0	586163	100,0

Frequenza punteggio Unici U.I. Long-Form

8	2825	6,1	2877	6,2
9	5616	12,2	5653	12,2
10	37598	81,7	37676	81,5
Totale	46039	100,0	46206	100,0

Come si può osservare dalla Tabella 14, la differenza di assegnazione punteggio, con *filtro* e senza *filtro*, non è molto rilevante, e la percentuale di 'Unici' con punteggio 10 per i tre file presi in esame va da 79,1% a 82,3%.

3.3.5 Analisi dei risultati della procedura 'Gestione dei Multipli'

Analisi sulle specifiche M45, M54 e M55 (primo test)

Il primo test di questa procedura è stato effettuato sui **dati Long-Form**, implementando le specifiche contrassegnate **dai flag M45, M54 e M55**.

Analisi sulle specifiche M45, M54 e M55 (primo test)

Il primo test di questa procedura è stato effettuato sui **dati Long-Form**, implementando le specifiche contrassegnate **dai flag M45, M54 e M55**.

I risultati ottenuti sono stati analizzati dagli esperti del CUE per verificarne i livelli di qualità ed, in funzione di questi stessi, confermare le specifiche di base o renderle più stringenti.

Prima di descrivere quanto emerso dall'analisi effettuata, al fine di avere un'idea sulla mole di dati trattata, si riportano nella seguente tabella gli output del passaggio di codifica automatica, con cui sono stati prodotti i 'Multipli' tra i quali estrarre quelli da convogliare tra gli 'Unici'.

Tabella 15 –Efficacia del sistema di codifica al 19/07/01

Risultati della codifica automatica	N.	%
Unici	40851	75
Multipli	5918	11
Possibili	4437	8
Falliti	2923	6
Totale	54129	

A seguito di questa analisi sull'output della procedura '*Gestione dei Multipli*', dunque, è emerso quanto segue:

- La procedura ha comportato l'estrazione di 1209 record. Gli esperti del CUE, a seguito di un'analisi puntuale, hanno associato a ciascun record appositi flag, come riportato nella Tabella 16.

Tabella 16 –Analisi 'Multipli' estratti dalla procedura 'Gestione dei Multipli'

N.ro record estratti	**%**	**Flag**	**Legenda**
170		**E**	Errore, codice **Ateco sbagliato**
56		*	Ateco incompleta, quando la definizione sarebbe stata sufficiente per essere codificata a 5 digit (**Carenza di dizionario**)
168		**?**	**Definizione generica** od equivoca
815		-	Codice **Ateco corretto**
Tot. 1209			

Questi risultati sono stati analizzati a seconda della specifica in base alla quale i singoli record erano stati estratti. I risultati sono riportati nella Tabella 17.

Tabella 17 –Analisi 'Multipli' estratti dalla procedura 'Gestione dei Multipli' a seconda della specifica che li ha prodotti

Flag	**E (sbagliati)**	*** (Carenza di dizionario)**	**? (Definizione generica)**	**Vuoti (corretti)**	**Totale Record**
M45	154	-	145	487	786
M54	-	56	5	23	84
M55	16	-	18	305	339
Totale	170	56	168	815	1209

La Tabella 18 riporta gli stessi dati della Tabella 17, secondo le percentuali di riga.

Tabella 18 –Analisi 'Multipli' estratti dalla procedura 'Gestione dei Multipli' a seconda della specifica che li ha prodotti, in percentuale

Flag	**e (sbagliati)**	*** (Carenza di dizionario)**	**? (Definizione generica)**	**Vuoti (corretti)**	**Totale**	**Totale in valore assoluto**
M45	20%	0%	18%	62%	100	786

	(sbagliati)	(Carenza di dizionario)	(Definizione generica)	(corretti)		assoluto
M45	20%	0%	18%	62%	100	786
M54	0%	67%	6%	27%	100	84
M55	5%	0%	5%	90%	100	339

- Analisi dei record con il **codice Ateco sbagliato**: come si vede dalle tabelle, dei 1209 estratti dai 'Multipli', a 170 (pari a circa il 14%) è stato associato un codice Ateco non corretto. La maggior parte (totale record 154) di codici Ateco sbagliati proviene dai record che presentano il **flag M45** (ovvero è stato assegnato il codice a 5 digit presente in 4 dei 5 record di output).
  I record con **flag M55** (ovvero è stato assegnato il codice comune a tutti e 5 i record di output), invece, sono soltanto 16, molti dei quali contengono errori d'ortografia e pertanto non possono essere riconducibili ad errori dell'applicazione; i restanti, come per esempio 'AEREO FOTOGRAMMATRIA' e 'FABBRICAZIONE MACCHINE INDUSTRIALI', segnalano invece la necessità di addestrare ulteriormente ACTR. Non sono invece presenti record con **flag M54** (assegnazione del codice corrispondente ai primi 4 digit comuni a tutti e 5 i record di output).

- Analisi dei record con il **codice Ateco incompleto.** Risulta che 56 dei 1209 record (pari a circa il 5%) hanno un codice Ateco incompleto, pur essendo la risposta testuale sufficiente per l'attribuzione del codice a 5 digit, ed hanno tutti **flag M54** e corrispondono quindi **a carenze del dizionario**. L'analisi dettagliata di questi record fa infatti notare la necessità di addestrare ulteriormente ACTR, visto che certe definizioni, presenti nei record di input, come per esempio 'TAPPEZZERIA' e 'FINISSAGGIO CALZATURE' sarebbero state sicuramente codificate al massimo dettaglio (5 digit) se ci fossero state le *empiriche* corrispondenti nel *dizionario elaborabile* Ateco.

- Analisi dei record con **codici Ateco dubbi**, ossia **definizioni generiche od equivoche**. In totale assommano a 168 record (pari a circa il 14%), dei quali 145 hanno **flag M45**. L'analisi da fare in questo caso è più complessa perché riguarda non solo la fase di addestramento di ACTR, ma soprattutto la necessità di evitare l'assegnazione di un codice equivoco.
  Per la fase di addestramento, per esempio, resta da stabilire quali siano i codici da associare ad alcune delle definizioni qui riscontrate, che rientrano in una casistica piuttosto frequente e che risultano essere generiche, in quanto non viene precisata in maniera completa il tipo di attività svolta. Per esempio:
  LAVANDERIA
  RESTAURATORE
  ELETTROMECCANICA
  PELLETTERIA
  VERNICIATURA INDUSTRIALE
  RADIOFONICA

- Analisi dei record **con Ateco corretta**. Sono in totale 815 (pari a circa il 67%), dei quali:
  305 con il flag M55 (37,4% dei corretti)
  23 con il flag M54 (2,8% dei corretti)
  487 con G il flag M45 (59,8% dei corretti)

Sulla base di questi risultati, è stato deciso di:
- mantenere come valide le specifiche caratterizzate dai **flag M55 e M54**; la prima, infatti, produce risultati caratterizzati da un livello di *accuratezza* (90%) coerente con quello che, sulla base delle sperimentazioni finora effettuate (si veda par. 3.3.2) caratterizza gli 'Unici';. per la seconda specifica, invece, deve essere fatto un discorso particolare; del totale dei record prodotti da questa specifica, soltanto il 27% sono stati giudicati corretti, ma in realtà il 67% è stato giudicato incompleto, nonostante una risposta testuale dettagliata;.quindi, a parte il fatto che saranno necessarie integrazioni del dizionario, il farli rientrare tra gli 'Unici' non produce un vero e proprio errore, ma implica la necessità di portarli al massimo dettaglio nel corso delle operazioni censuarie, probabilmente tramite apposite procedure di imputazione;
- la specifica corrispondente al **flag M45**, invece, produce risultati con un tasso di errore superiore a quello giudicato tollerabile, quindi non viene implementata.

Approfondimento dell'analisi (secondo test)

Gli esperti del CUE hanno poi approfondito l'analisi sui dati Long Form e unità locali Long Form, esaminando, oltre ai casi nei quali ACTR produce 5 'Multipli' che presentano 5 codici tutti uguali, anche quei casi nei quali ACTR produce soltanto 4 'Multipli' tutti con lo stesso codice, oppure soltanto 3, oppure 2 (limitandosi sempre a codici a 5 digit).
I risultati sono riportati nella Tabella 19. La percentuale di record esatti varia da un minimo di 89,3% per i record che presentano solo due codici uguali proposti da ACTR, ad un massimo di 97,1% per i record che presentano quattro codici uguali proposti da ACTR. In media gli esatti sono pari a 92,7% per Long Form e 94,1% per U.L. Long Form.

casi nei quali ACTR produce 3 'Multipli' che presentano 3 codici tutti uguali, anche quei casi nei quali ACTR produce soltanto 4 'Multipli' tutti con lo stesso codice, oppure soltanto 3, oppure 2 (limitandosi sempre a codici a 5 digit).
I risultati sono riportati nella Tabella 19. La percentuale di record esatti varia da un minimo di 89,3% per i record che presentano solo due codici uguali proposti da ACTR, ad un massimo di 97,1% per i record che presentano quattro codici uguali proposti da ACTR. In media gli esatti sono pari a 92,7% per Long Form e 94,1% per U.L. Long Form.
Nella Tabella 20 si nota che gli errori sono principalmente determinati da una descrizione generica.

Tabella 19- Analisi dei 'Multipli' con solo 2,3,4,5 codici Ateco uguali Long Form e U.L. long

	long-form			U.L. long		
	totale	Esatti	errati	totale	esatti	errati
	Valori assoluti					
Solo due codici ateco	224	200	24	289	271	18
Solo tre codici ateco	140	133	7	134	124	10
Solo quattro codici ateco	103	100	3	93	90	3
Solo cinque codici ateco	310	287	23	364	343	21
Totale	777	720	57	880	828	52
			%			
Solo due codici ateco	28,8	89,3	10,7	32,8	93,8	6,2
Solo tre codici ateco	18,0	95,0	5,0	15,2	92,5	7,5
Solo quattro codici ateco	13,3	97,1	2,9	10,6	96,8	3,2
Solo cinque codici ateco	39,9	92,6	7,4	41,4	94,2	5,8
Totale	100	92,7	7,3	100	94,1	5,9

Tabella 20 – Distribuzione errati per tipo di errore

	long-form		U.L.long	
	valori assoluti	*%sul totale*	*Valori assoluti*	*% sul totale*
Errati per testo generico	34	59,7	40	76,9
Per dizionario ancora carente	17	29,8	11	21,2
Errori di ortografia	6	10,5	1	1,9
Totale	57	100	52	100

Nella Tabella 21 sono riportate le analisi sui 'Multipli'/output in cui i primi quattro codici sono uguali (flag M45). La percentuale di record esatti (74,2%) risulta più bassa rispetto ai casi sopra descritti (confermando il risultato già ottenuto nel test precedente).

Nella Tabella 22 si nota che gli errori sono principalmente determinati da un dizionario ancora carente.

Tabella 21--Analisi dei 'Multipli' con i primi quattro codici Ateco uguali proposti da ACTR (4 su 5)

Esatti ACTR	213	74,2
Errati ACTR	74	25,8
Totale	287	100

Tabella 22 – Distribuzione errati per tipo di errore

	valori assoluti	*%sul totale*
Errati per testo generico	22	29,7
Per dizionario ancora carente	46	62,1

Errati per testo generico	22	29,7
Per dizionario ancora carente	46	62,1
Errori di ortografia	6	8,1
Totale	74	100

Sulla base di questi risultati è stata implementata anche la specifica contrassegnata con il flag **MM**, di cui al paragrafo 3.2.2.

Ulteriore analisi sui 'Possibili'

I 'Possibili' con cinque codici uguali sono stati analizzati per verificare la percentuale d'errore di attribuzione codice e quindi l'eventuale inserimento di questi tra gli 'Unici'.
Sono stati estratti ed analizzati i casi 'Possibili' per i quali ACTR assegna 5 codici uguali, o solo quattro, tre, due uguali; l'analisi è stata effettuata su Long Form (limitandosi sempre a codici a 5 digit). I risultati ottenuti sono riportati nella Tabella 23. La percentuale di record esatti varia da un minimo di 83% per i record che presentano solo due codici uguali proposti da ACTR, ad un massimo di 95,5% per i record che presentano tre codici uguali proposti da ACTR. In media gli esatti sono pari a 87%. Nella Tabella 24 si nota che gli errori sono principalmente determinati da dizionario ancora carente.

Tabella 23 - Analisi dei 'Possibili' con solo 2,3,4,5 codici ateco uguali long-form

	Totale	esatti	errati
	valori assoluti		
solo due codici ateco	294	244	50
solo tre codici ateco	111	106	5
solo quattro codici ateco	59	55	4
solo cinque codici ateco	267	231	36
Totale	731	636	95
		%	
solo due codici ateco	40,2	83,0	17,0
solo tre codici ateco	15,2	95,5	4,5
solo quattro codici ateco	8,1	93,2	6,8
solo cinque codici ateco	36,5	86,5	13,5
Totale	100,0	87,0	13,0

Tabella 24 - Distribuzione errati per tipo di errore

	Valori assoluti	%sul totale
Errati per testo generico	31	32,6
per dizionario ancora carente	55	57,9
Errori di ortografia	9	9,5
Totale	95	100,0

3.3.6 L'impatto degli errori di ortografia

Dall'analisi degli output prodotti dal sistema di codifica automatica ACTR si è riscontrato che una parte di descrizioni non è codificata correttamente, poiché il testo contiene degli errori ortografici o abbreviazioni non univoche che non si è

Dall'analisi degli output prodotti dal sistema di codifica automatica ACTR si è riscontrato che una parte di descrizioni non è codificata correttamente, poiché il testo contiene degli errori ortografici o abbreviazioni non univoche che non si è potuto risolvere nei file di *parsing*.
ACTR pertanto, non riconoscendo le parole errate, codifica quel testo con codice non Unico o Unico ma con codice errato. Per ovviare in parte a tale problema è stato sperimentato un programma S.A.S. di correzione delle parole abbreviate o errate più frequenti (cfr. .Allegato 5.).
Il testo da codificare è stato elaborato considerando una frase massima di 10 parole, analizzando le frequenze della 1°. 2°, 3°,.....10° parola, con il fine di individuare e correggere gli errori o abbreviazioni più comuni.
La suddivisione in 1°-- 10° parola è stata necessaria dal momento che la correzione dipende dalla posizione della parola nella frase. Esempio: *'prod*.' può essere corretta in '*produzione*' solo sulla prima parola del testo, sulla seconda può essere invece corretta con '*prodotto*'.
Se si volesse codificare, per esempio, la seguente attività economica:
'PROD. CONFEZIONI UOMO DONNA BAMBINO'
poiché nell'attuale applicazione di codifica automatica l'abbreviazione *'prod'* non viene risolta in quanto ambigua, si genererebbe un match Unico ma con codice Ateco errato.
Tramite il citato passaggio di correzione ortografica in SAS, invece, '*prod'* diventa *'produzione',* in quanto, essendo la 1° parola della frase, l'ambiguità può essere risolta; ACTR genera quindi un match Unico con codice Ateco corretto.
Un altro esempio per spiegare meglio l'utilità del passaggio di correzione potrebbe essere il seguente:
'COMM. INGROSSO PROD. SEMIL. LEGNO'
nell'applicazione di codifica automatica viene generato un match corretto ma non Unico, mentre con il secondo passaggio di codifica, a seguito della correzione ortografica che trasforma la frase in '*commercio ingrosso prodotti semilavorati legno'* viene attribuito un codice Unico corretto a cinque digit con punteggio uguale a 10.
La gestione delle abbreviazioni è quindi, in generale, molto complessa. La gestione poi delle abbreviazioni con il punto e senza si complica ulteriormente perché può determinare anche un'attribuzione di *filtro* sbagliata, considerando che il contesto di codifica e il contesto che attribuisce il *filtro* lavorano con gli stessi file di *parsing*.

Prendiamo in esame per esempio *'costruzione fabbrica.'* che può essere considerata come abbreviazione di *'costruzione fabbricato/i',* quindi sinonimo di *'edificio'*.
La stessa cosa non avviene con la stessa frase senza punto, *'costruzione fabbrica',* in quanto *'fabbrica'* è considerata una parola completa e sinonimo di *'produzione'*.

Per quanto riguarda l'attribuzione del *filtro*, quindi, alla frase *'costruzione fabbrica.'* non viene attribuito alcun *filtro*, mentre a *'costruzione fabbrica'* viene attribuito il *filtro* della produzione (20).

Anche nel passaggio di codifica vera e propria, poi, sia che si lavori con o senza *filtro*, i risultati sono diversi a seconda della presenza/assenza del punto, infatti:
'costruzione fabbrica.'-à Unico *'costruzione di edifici'* codice corretto
'costruzione fabbrica' -à Unico *'costruzione produzione'* codice errato

Anche le altre abbreviazioni di *'fabbricati/o'* con punto e senza presentano problematiche derivanti da problemi di ambiguità e producono risultati diversi, infatti:

*'costruzione fabbricat.'*à *Unico 'costruzione di edifici'* codice corretto
*'costruzione fabbricat'*à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabbric.' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabbric' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabbri.' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabbri' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabbr.' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabbr' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabb.' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fabb' à *Unico 'costruzione di edifici'* codice corretto
'costruzione fab.' à *Unico 'costruzione' N.C.*
'costruzione fab' à *Unico 'costruzione' N.C.*

Le ultime due frasi, infatti, non vengono risolte perché *'fab.'* e *'fab'* sono troppo ambigue per essere ricondotte ad un singolo sinonimo (con *'fab.'* e *'fab'* si potrebbe anche intendere *'fabbro'*, mentre si è ritenuto poco probabile che tale parola sia abbreviata con *'fabb'*).

Da quanto sopra illustrato emerge pertanto l'impossibilità di risolvere nell'applicazione di codifica molte abbreviazioni non univoche. In alcuni casi si è dimostrato che ciò è invece possibile tramite il programma di correzione ortografica in SAS.

Da quanto sopra illustrato emerge pertanto l'impossibilità di risolvere nell'applicazione di codifica molte abbreviazioni non univoche. In alcuni casi si è dimostrato che ciò è invece possibile tramite il programma di correzione ortografica in SAS.

Nella sperimentazione, il programma di correzione ortografica è stato applicato solo sui record non codificati correttamente, escludendo gli 'Unici' a cinque digit e i 'Multipli' con cinque codici Ateco uguali del primo passaggio.
I record così corretti sono stati quindi nuovamente sottoposti a codifica automatica tramite ACTR (chiameremo questo '*secondo passaggio'*). Il risultato raggiunto con il *secondo passaggio* è illustrato nelle tabelle sottostanti

Dalla Tabella 25 si evince che la percentuale di codifica degli 'Unici' completi, dopo la correzione di alcune parole errate o di abbreviazioni non riconosciute da ACTR, varia da un massimo del 2,3% per il dataset Long Form ad un minimo di 1% per I.N.P.S, e le variazioni con *filtro* e senza *filtro* non sono rilevanti.

Tabella 25 – Confronto con e senza filtro 2° passaggio

	filtro			**no filtro**		
	I.N.P.S.	Long Form	U.L. Long	I.N.P.S.	Long Form	U.L. long
			valori assoluti			
FALLITI	74688	2975	5036	64294	2824	4792
MULTIPLI	84132	4867	6119	87043	4879	6069
MULTIPLI con 5 cod Ateco uguali	28	1	5	28	1	5
POSSIBILI	80593	4056	4536	85750	4046	4543
POSSIBILI con 5 cod Ateco uguali	4904	267	332	4809	255	319
UNICI con Ateco < 5 digit e 'n.c.'	10729	1234	1379	11368	1308	1514
UNICI	2508	321	299	2501	330	299
Totale	257582	13721	17706	255793	13643	17541
			%			
FALLITI	29,0	21,7	28,4	25,1	20,7	27,3
MULTIPLI	32,7	35,5	34,6	34,0	35,8	34,6
MULTIPLI con 5 cod Ateco Uguali	0,0	0,0	0,0	0,0	0,0	0,0
POSSIBILI	31,3	29,6	25,6	33,5	29,7	25,9
POSSIBILI con 5 cod Ateco uguali	1,9	1,9	1,9	1,9	1,9	1,8
UNICI con Ateco < 5 digit e 'n.c.'	4,2	9,0	7,8	4,4	9,6	8,6
UNICI	1,0	2,3	1,7	1,0	2,4	1,7
Totale	100,0	100,0	100,0	100,0	100,0	100,0

Su 321 record 'Unici' Long Form con *filtro* è stata effettuata una verifica manuale per stabilire il codice esatto al fine di verificare la bontà della codifica ACTR. I risultati sono riportati nella Tabella 26. Si ha che nel 14,6% dei casi ACTR assegna un codice errato. Nella Tabella 27 è riportata la distribuzione per tipo di errore.
La presente analisi mette in evidenza un numero elevato di errori nell'attribuzione del codice ateco, dovuto ad errori di ortografia ancora presenti nel testo da codificare, e ad un numero elevato di descrizioni troppo generiche da parte dei rispondenti.

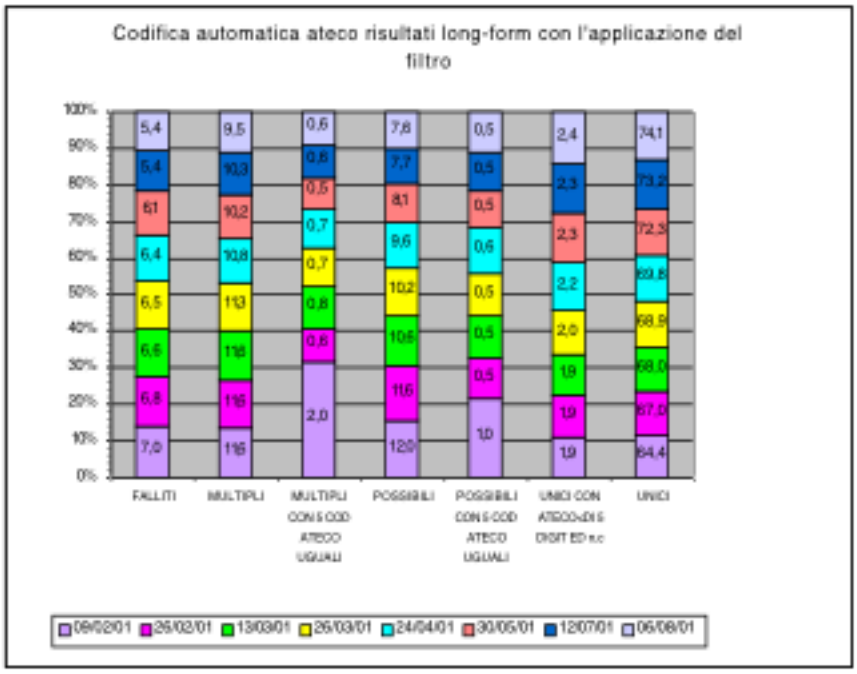
Tabella 26 – Analisi 'Unici' 2° passaggio ACTR

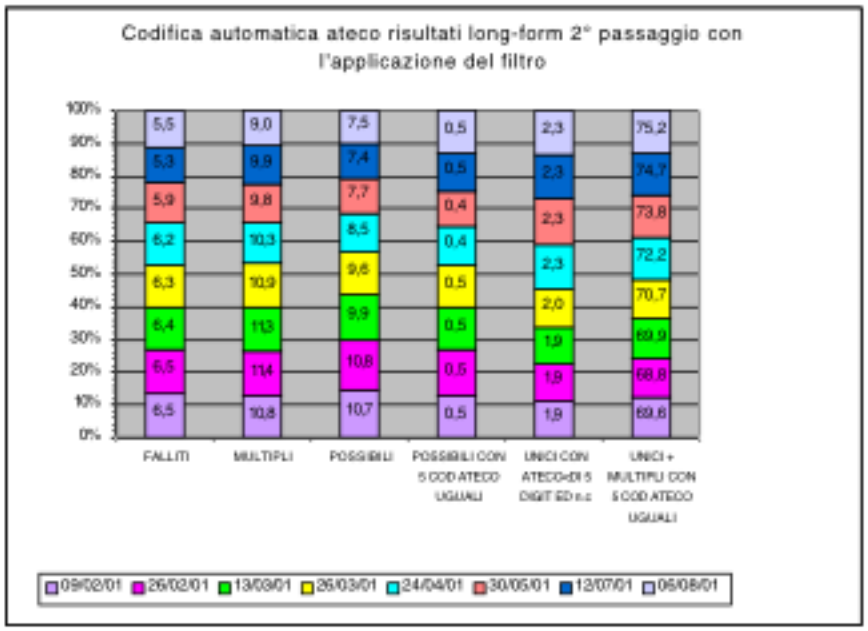
	Valori assoluti	%
Esatti ACTR	274	85,4

	Valori assoluti	%
Esatti ACTR	274	85,4
Errati ACTR	47	14,6
Totale	321	100

Tabella 27 – Distribuzione errati per tipo di errore

Errati per testo generico	22	46,8
Attribuzione *filtro* errato o dizionario ancora carente	10	21,2
Ortografia o abbreviazioni	15	32,0
Totale	47	100





4 Prospettive e problematiche per il prossimo censimento

In base a quanto descritto nei paragrafi precedenti, l'adozione della codifica automatica nel prossimo censimento dell'industria comporterà indubbiamente miglioramenti in termini di processo e di qualità dei dati.
Tuttavia, ulteriori miglioramenti sono tuttora apportabili.

4.1 Arricchimento della base informativa

Innanzi tutto**, l'attività di arricchimento della base informativa** del sistema di codifica non è stata interrotta ed è opportuno farla continuare utilizzando tutte le fonti nelle quali il quesito sull'attività economica viene rilevato a testo libero e registrato.
E' importante sottolineare comunque che, come appare ormai chiaro da quanto riportato in questo documento, i dizionari utilizzati per la codifica automatica non possono essere considerati e trattati come archivi statici, ossia da aggiornare soltanto a seguito delle revisioni delle classificazioni ufficiali. Il successo delle applicazioni di codifica dipende infatti dal continuo aggiornamento di questi ultimi al fine di

- estenderne la variabilità linguistica,
- allinearli alle evoluzioni del linguaggio,
- allinearli ai cambiamenti della società inerenti la variabile di riferimento.

4.2 Gestione degli errori ortografici

Un secondo elemento da considerare è inerente gli **errori ortografici**. Come si evince dalla descrizione della logica di *matching* dei testi adottata in ACTR, due parole contenutisticamente uguali, di cui una affetta da errore ortografico, sono considerate assolutamente diverse, quindi non abbinate. Questo può inficiare la possibilità di successo nell'attribuzione del codice, soprattutto se il testo da codificare è molto breve (costituito da poche parole) e la parola errata avrebbe avuto un peso significativo, se fosse stata riconosciuta.
Non è stato possibile quantificare l'impatto degli errori ortografici sull'*efficacia* del sistema (tasso di codifica); tuttavia, come descritto nel paragrafo sui risultati ottenuti nel corso dell'Indagine Long Form, analizzando il livello di *accuratezza* (percentuale di codici corretti assegnati automaticamente) si è rilevato che, data una percentuale di errore stimata del 9%, soltanto l'1% era attribuibile a errori di ortografia.

Come già descritto, è stato sperimentato un programma in SAS di correzione ortografica ed espansione della abbreviazioni, che tiene conto sia della frequenza delle parole che della posizione delle stesse nella frase. Come si evince dalle tabelle riportate in questo documento, tale programma comporta un miglioramento dell'*efficacia* del sistema che varia dall'1% al 2,3%, pur non risolvendo completamente tale problema.
Si ritiene che sarebbe opportuno ottimizzare tale fase e premetterla al passaggio di codifica automatica.

Nel contempo, nell'implementazione dell'ambiente applicativo di codifica automatica, si è provveduto a non prescindere completamente da questo problema. Sono stati infatti considerati gli errori ortografici sulle parole più frequenti e su alcune abbreviazioni, trattandoli come sinonimi delle parole corrette (tecnicamente ciò è stato realizzato sia a livello di '*replacemnt words'*, che a livello di *'double words'*, prendendo cioè in considerazione coppie di parole che permettessero di risolvere le eventuali ambiguità nella definizione dei sinonimi).

Le tabelle che seguono mostrano alcuni esempi di tali trasformazioni

Tabella 28 – Errori ortografici contemplati nell'ambiente di codifica automatica a *livello di RWRD*

Parola affetta da errore ortografico o abbreviazioni	***Parola corretta cui è ricondotta***
Ltro, latro	Altro
Ffitto, fitto, ocazione, locaz	Locazione
Asciugaturta	Asciugatura
Apparecciatura	Apparecchio
Autodesivo	Adesivo
Autoespurgo	Autospurgo
Aurto	Auto
Attrazzatura	Attrezzatura
Attivià	Attività
Ambilante	Ambulante
Comm, ommercio, comme, commer, commerc, comemrcio, commecio, commeercio, commerfcio, commerciio, commercioal, commercioo, commerco, commerxio, commerico, commertcio, commwercio, ommerio, ommer, omm, coomm	Commercio
Calziturificio	Produzione Calzatura
Carne	Carne

commerxio, commerico, commertcio, commwercio, ommerio, ommer, omm, coomm	
Calziturificio	Produzione Calzatura
Carme	Carne
Costr, costruz, costruzi, ostruzione, costurzione/i, ostr, ostruz	Costruzione
Ucitura	Cucitura
Dettalgio, aldettaglio, dettahglio, ddettaglio	Dettaglio
Eleborazione	Elaborazione
Elateria	Gelateria
Eletronico	Elettronico
Ervizio	Servizio
Elettrocustico	Elettroacustico
Idustriale	Industria
Gioelleria, giolleria	Gioielleria
Laboratoorio	Laboratorio
Iimpianto, mpianto	Impianto
Mettalico	Metallo
Orveglianza	Sorveglianza
Orificera	Oreficeria
Ontaggio	Montaggio
Nin	Non
Matteriale	Materia
Dittte	Impresa
Fabbricaz, abbricazione, fabbricazone, fabbriz, abbric, abbr,	Produzione
Igrosso, inngrosso, ingorosso, ngrosso	Ingrosso
Produ, produz, produzio, roduz, produzone, roduzione, oroduzione, produzionew, prodruzione	Produzione
Pwer	
Tomata, topaia	Tomaia
Tampaggio	Stampa
Strazione	Estrazione
Riparazuone	Riparazione
Realozzazione	Realizzo
Vetreo	Vetro
Tessit, essitura, tessirura	Tessitura

Tabella 29 – Errori ortografici contemplati nell'ambiente di codifica automatica a *livello di DWRD*

Coppie di parole affette da errore ortografico o abbreviazioni	***Parole corrette cui sono ricondotte***
Abbigliamento Conf	Abbigliamento Confezioni
Accessorio Ed	Accessorio Edilizia
Acq Dist	Acqua Distillata
Amb.te	Ambulante
Amm.tiva	Amministrazione
Animale Mang	Animale Mangime
Centro Com	Centro Commercio
Conf Biancheria	Confezionamento Biancheria
Conf.to	Confezionamento

In vista del prossimo censimento, sarebbe quindi opportuno investire maggiormente su questo aspetto.

4.3 Sistema di monitoraggio della qualità

Ogni qual volta l'ambiente di codifica automatica viene utilizzato per codificare i dati di un'indagine, sarebbe opportuno non soltanto verificare la sua tenuta in termini di *efficacia*, ma controllarne il livello di *accuratezza*, al fine di:

- misurare la qualità del processo e garantire quindi che il livello di errore non superi la soglia prefissata

opportuno non soltanto verificare la sua tenuta in termini di *efficacia*, ma controllarne il livello di *accuratezza*, al fine di:

- misurare la qualità del processo e garantire quindi che il livello di errore non superi la soglia prefissata,
- possibilmente intervenire sull'ambiente di codifica per evitare il ripetersi dell'errore.

Operativamente, l'analisi dei risultati della codifica può essere circoscritta ai casi codificati a seguito di un *indirect match* , ossia agli 'Unici' con punteggio minore di 10, in quanto, dando per scontato la correttezza del dizionario, si presume che gli 'Unici' con punteggio uguale a 10 siano corretti.
Quantitativamente, questo significa sottoporre a verifica in media il 30% dei casi codificati.
Questa attività deve necessariamente essere affidata a codificatori esperti e, nel caso di indagini caratterizzate da un'esigua mole di dati da codificare, può essere realizzata a tappeto.
Nel caso di un censimento, invece, sarebbe opportuno strutturare questa fase in modo da sottoporre a controllo soltanto un campione di dati.

Nell'ambito di uno studio già effettuato presso il servizio Metodologie per la Produzione Statistica, si è provveduto a misurare l'*accuratezza* della codifica di un file di circa 350.000 testi da codificare (relativamente alla variabile *Professione*), disegnando un campione costituito esclusivamente da testi 'differenti' (dal file originario sono state individuate le risposte differenti l'una dall'altra ed a ciascuna di esse è stata associata la frequenza con cui sono state rilevate). In pratica, è stata dapprima effettuata una stratificazione in funzione della citata frequenza, quindi, nell'ambito di ciascuno strato, è stato estratto un campione casuale semplice di testi.
Dato un livello di *accuratezza* prefissato (97%), è stato dimostrato che è sufficiente estrarre un campione di 937 testi (corrispondente al 6.79% di testi differenti), per garantire la qualità del campione originario (Macchia S. and D'Orazio, 2000).

Nell'ambito del censimento, quindi, sarebbe opportuno disegnare una procedura, opportunamente parametrizzata, che consenta l'individuazione dei 'testi diversi', la definizione dell'ampiezza campionaria e l'estrazione del campione, per lotti di dati.
Tale procedura dovrebbe essere dotata di un'interfaccia user friendly, in modo da poter essere gestita direttamente dagli utenti finali.
Qualora si reputi strategico realizzare tale sistema, è indispensabile definire le risorse da dedicare a questa attività, in funzione dei tempi ormai ristretti.

4.4 Soluzione dei casi non risolti automaticamente

Resta ovviamente aperto il problema della soluzione dei casi non 'Unici' ai quali il sistema non ha attribuito un singolo codice, oppure di quelli 'Unici' cui il sistema ha attribuito un codice non completo.
Ciò può avvenire per diverse ragioni:

- la risposta non aveva un contenuto informativo sufficiente per l'individuazione di un singolo codice,
- la risposta aveva un contenuto informativo sufficiente, ma il sistema ha fallito a causa di una carenza del dizionario, oppure di un errore ortografico.

In entrambi i casi il sistema può:

1. aver assegnato un unico codice, ma incompleto (con meno di cinque digit),
2. aver proposto un certo numero di codici tra i quali selezionare quello corretto,
3. aver completamente fallito (non aver individuato alcun codice possibile).

Le soluzioni possono essere diverse e vanno dall'affidamento a codificatori esperti all'adozione di procedure automatiche che trattino questa casistica come quella delle mancate risposte e le risolvano, ad esempio, con la metodologia del donatore, con la peculiarità che nei casi rientranti nelle tipologie 1) e 2) l'informazione derivante dalla risposta testuale non sarebbe persa, ma potrebbe essere utilizzata per selezionare il miglior donatore.

Si ritiene che, nel caso del censimento, data la mole di dati da trattare, potrebbe essere adottata una strategia intermedia, consistente nel sottomettere a codificatori esperti i casi 'Falliti', in modo da risolvere quelli che hanno un contenuto informativo sufficiente per essere codificati, ed adottare procedure automatiche per i restanti casi.

Ovviamente, anche l'implementazione di questa procedura richiede risorse da dedicare.

procedure automatiche per i restanti casi.

Ovviamente, anche l'implementazione di questa procedura richiede risorse da dedicare.

4.5 Assetto organizzativo

L'applicazione della codifica automatica nel corso del prossimo censimento richiede un assetto organizzativo diverso rispetto a quello adottato nel corso delle attività del gruppo di lavoro.

Si ritiene infatti indispensabile cedere completamente la responsabilità e la gestione dell'ambiente applicativo di codifica agli esperti del dipartimento delle statistiche economiche perché:
- lì risiedono le competenze specifiche sul tema;
- se tale soluzione non fosse adottata, la struttura MPS/C verrebbe presto a costituire un collo di bottiglia rispetto alle esigenze censuarie, in quanto è attualmente impegnata in una serie di altre attività rispetto alle quali non è possibile, oltre a non essere funzionalmente accettabile, distogliere le risorse.

Il passaggio di responsabilità sarà inoltre facilitato da:
- una dettagliata documentazione resa disponibile sui criteri di codifica adottati e sulle relative specifiche tecniche inerenti il *parsing* (cfr.. il capitolo 5 e l'Allegato 6);
- il manuale di utilizzo di ACTR, di cui è prevista la diffusione entro la fine dell'anno;
- il proseguimento della collaborazione con la struttura MPS/C, che si impegnerà in tutta l'attività di formazione ed affiancamento delle risorse del CUE che sarà necessaria, soprattutto nella prima fase di passaggio di consegne.

5 Criteri classificatori adottati nella costruzione del dizionario

5.1 Criteri generali

Nei file ausiliari contenenti i parametri che il software applica per rimuovere qualsiasi elemento grammaticale o sintattico che possa rendere differenti due frasi con lo stesso contenuto semantico, in particolare in quelli relativi alla definizione dei sinonimi, è stato possibile inserire diverse dizioni di uso comune equiparandole alle corrispondenti presenti nella 'Classificazione delle attività economiche', per esempio:
Vendita = Commercio
Fabbricazione = Produzione
Commercializzazione =Commercio ingrosso
Rappresentante = Mediazione
Si è cercato poi di prevedere tutte le 'Possibili' abbreviazioni univoche come per esempio
Lavoraz. = Lavorazione
Abbigl. = Abbigliamento
Bicicl = Bicicletta.

5.2 Doppie attività economiche

Dall'analisi delle risposte fornite nell'ambito di diverse indagini è emersa pesantemente la casistica delle **doppie attività esercitate** dai rispondenti che, per poter essere codificate, richiedono la definizione di appositi criteri guida, ispirati, nella gran parte dei casi, a far prevalere, tra quelle menzionate, l'attività che fornisce maggior reddito.
I criteri individuati sono a volte validi indipendentemente dalla categoria Ateco cui si riferiscono o dai prodotti oggetto delle attività (*criteri generalizzati*), a volte sono invece peculiari di certe situazioni (*criteri particolari*).

I '*criteri generalizzati*' sono stati risolti nell'ambito dei file di *parsing*, mentre quelli '*particolari*' tramite l'inserimento nel *reference* di apposite *empiriche*.

La complessità della problematica ha tuttavia fatto sì che anche i cosiddetti '*criteri generalizzati*' presentino delle eccezioni che sono state di volta in volta risolte.

5.2.1 Doppie attività: 'criteri generalizzati'

Si esaminano di seguito coppie di attività nell'ambito delle quali è stata stabilita, su indicazione del Servizio CUE, la prevalenza (>)

Si esaminano di seguito coppie di attività nell'ambito delle quali è stata stabilita, su indicazione del Servizio CUE, la prevalenza (>)

- **Produzione/lavorazione** rispetto a **Commercio**

a) Produzione >Commercio
b) Fabbricazione >Commercio
c) Trasformazione >Commercio
d) Lavorazione >Commercio

Per il punto c) vale la seguente **eccezione**:
l'attività della ***Stagionatura*** fa parte della fase di trasformazione di un certo prodotto. Tuttavia, quando viene considerata assieme all'attività del *Commercio,* predomina quest'ultimo nonostante la prevalenza contraria espressa sopra. Come viene mostrato nell'esempio che segue dove l'attività prevalente risulta essere il commercio all'ingrosso:
51.33.1 Stagionatura e commercio di formaggi

- **Produzione** rispetto a **Riparazione**

Produzione >riparazione

- **Installazione** rispetto a **Riparazione**

Installazione >riparazione
Quando l'installazione è da parte delle ditte costruttrici la fabbricazione è sempre prevalente sia sull'installazione che sulla riparazione

- **Riparazione** rispetto a **Commercio**

Commercio >riparazione
Valgono le seguenti **eccezioni,** per le quali l'attività prevalente rimane la riparazione

20.40.0 Riparazione vendita pallets
50.20.4 Commercio riparazione sostituzione di pneumatici

5.2.2 Doppie attività: 'criteri particolari'

Come già accennato, le regole che seguono, pur essendo considerate come ***criteri di codifica***, non sono state risolte nei file di sistema, in quanto sono state ritenute più elastiche perché potrebbero incorrere in più di una eccezione. Tecnicamente, sono state infatti inserite una serie di *empiriche* che rispecchiano questi criteri:
- Lavorazione >stagionatura
- Produzione >ricerca
- Progettazione>consulenza, analisi
- Maglierista >ricamatrice

5.2.3 Criteri adottati nelle singole divisioni

Nella divisione 01 (Agricoltura caccia e relativi servizi)
Si è riscontrato che spesso i rispondenti utilizzano la parola *produzione* come sinonimo di *coltivazione*. Non essendo questo sinonimo estendibile a tutte le classi, sono state duplicate alcune voci con la sostituzione della parola *produzione* al posto di *coltivazione* prevista dalla classificazione.
Si sono stabiliti poi i seguenti codici non completi (a due digit) come:

01 - (ei) Agricoltore
01 - (ei) Agricoltore imprenditore
01 - (ei) Attività agricola
01 - (ei) Coltivazione campi
01 - (ei) Conduzione aziende agricole
01 - (ei) Contadino

Nell'ambito della divisione si sono utilizzati i seguenti **criteri particolari:**

) Si è stabilito che *la Prima lavorazione di prodotti agricoli* -à divisione 01
) *La lavorazione e trasformazione di prodotti agricoli* è stata invece intesa come *Trasformazione* -à divisione 15
) *L'allevamento di bovini è prevalente* su quello degli equini

Nella divisione 02 (Silvicoltura e utilizzazione di aree forestali e servizi connessi)
Non sono stati segnalati dal Servizio CUE criteri particolari da considerare. Nelle varie divisioni, gruppi, classi e categorie si è provveduto, pertanto, ad inserire unicamente delle *empiriche*.

Nella divisione 05 (Pesca piscicoltura e servizi connessi)
Pesca con turbo soffiante è la pesca di molluschi in acque lagunari.

categorie si è provveduto, pertanto, ad inserire unicamente delle *empiriche*.

Nella divisione 05 (Pesca piscicoltura e servizi connessi)

Pesca con turbo soffiante è la pesca di molluschi in acque lagunari.

Poiché non sono stati segnalati dal Servizio CUE criteri particolari o *empiriche* da inserire nel dizionario informatizzato, la divisione contiene unicamente le voci della classificazione ufficiale opportunamente rielaborate come detto sopra.

Nella divisione10 (Estrazione di carbon fossile e lignite; estrazione di torba)

Poiché non sono stati segnalati dal Servizio CUE criteri particolari o *empiriche* da inserire nel dizionario informatizzato, la divisione contiene unicamente le voci della classificazione ufficiale opportunamente rielaborate come detto sopra.

Nella divisione 11 (Estrazione di petrolio greggio e di gas naturale; servizi connessi all'estrazione di petrolio e di gas naturale, esclusa la prospezione)

Poiché non sono stati segnalati dal Servizio CUE criteri particolari o *empiriche* da inserire nel dizionario informatizzato, la divisione contiene unicamente le voci della classificazione ufficiale opportunamente rielaborate come detto sopra.

Nella divisione 12 (Estrazione di minerali e torio)

Poiché non sono stati segnalati dal Servizio CUE criteri particolari o *empiriche* da inserire nel dizionario informatizzato, la divisione contiene unicamente le voci della classificazione ufficiale opportunamente rielaborate come detto sopra.

Nella divisione 13 (Estrazione di minerali metalliferi)

Poiché non sono stati segnalati dal Servizio CUE criteri particolari o *empiriche* da inserire nel dizionario informatizzato, la divisione contiene unicamente le voci della classificazione ufficiale opportunamente rielaborate come detto sopra.

E' stata inserita una sola empirica

13.20.0 - (ei) Estrazione coltivazione e trattamento minerali metallici non ferrosi

Nella divisione 14 (Altre industrie estrattive)

Se non viene effettuata l'*estrazione,* è sottinteso che tutte le altre attività, come per esempio la *frantumazione,* vengano effettuate *fuori cava*

Estrazione di materiali di cava va nella categoria 14.21.0

Cava e lavorazione trachite (coerentemente con la Prodcom) nella categoria 14.11.2

Il calcare può avere un doppio significato 'pietra per calce' e 'pietra da costruzione'

La pozzolana deve andare nella categoria 14.12.2

Nella divisione 15 (Industria alimentare e delle bevande)

Quando si parla contemporaneamente di *trattamento e/o trasformazione e di distribuzione (trasporto),* per esempio *del latte,* è prevalente l'attività di trattamento e trasformazione, altrimenti viene considerato un semplice trasporto (categoria 15.51.1)

Produzione pane e dettaglio va nella categoria 15.81.1

Il commercio di pane + forno va nella categoria 15.81.1

Per le attività che riguardano la produzione di più prodotti si è stabilito per es. che

Produzione pane e pasticceria >produzione pane

Oleificio implica la raffinazione

Palmento (macina del mulino, insieme delle macine dell'attrezzatura di un mulino) prevale sulla raffinazione pertanto *Palmento> Oleificio*

Nella divisione 16 (Industria del tabacco)

Anche l'*Amministrazione dei monopoli* va nella categoria 16.00.0

Nella divisione 17 (Industrie tessili e dell'abbigliamento)

Sarebbe necessario consultare un esperto tecnico del settore in quanto, per esempio, diverse attività intermedie tra la filatura e la tessitura sono difficilmente classificabili, venendo anche svolte senza differenziarle tra cotone, lana, etc.. La Prodcom risolve in parte la questione con le categorie 17.10.0 e 17.20.0, ma per Ateco 91 il problema non è di facile risoluzione.

Poiché sono comunque attività molto diffuse sarebbe opportuno risolvere tale problematica alla radice con indicazioni chiare da parte dei responsabili della classificazione.

Le indicazioni del CUE sono comunque state le seguenti:

- *Lana pettinata >lana cardata*
- Se non viene specificato il tipo di lana *si considera sempre la lana pettinata*
- *Attività di roccatura è* prevista per tutti il filati
- *L'attività* di *roccatura della lana* va nella categoria 17.13.2
- *L'attività di filatura* generica va nel gruppo 17.1
- *Nello stesso gruppo della filatura (*17.1) c'è sempre la preparazione di qualsiasi fibra
- *La produzione di tessuti* prevale sulla produzione di filati
- *La produzione di prodotti tessili in genere* va nella divisione 17
- *La fase di tessitura* prevale sulla fase di smacchiatura e candeggio

- *Nello stesso gruppo della filatura* (17.1) c'è sempre la preparazione di qualsiasi fibra
- *La produzione di tessuti* prevale sulla produzione di filati
- *La produzione di prodotti tessili in genere* va nella divisione 17
- *La fase di tessitura* prevale sulla fase di smacchiatura e candeggio
- *L'attività* di *roccatura generica* va nella categoria 17.17.0
- *L'attività* di *torcitura* senza alcuna specifica va nella categoria 17.17.0
- *Le attività di testurizzazione e torcitura di filati sintetici o artificiali* vanno nella categoria 17.15.0
- *Imbozzimatura, annodatura, orditura, stribbiatura* sono tutte fasi e operazioni di preparazione
- *La fase di tessitura* inizia dal gruppo 17.2
- *La fase di garzatura* riguarda la finitura dei tessuti
- *Gli scardassi* sono degli strumenti per l'industria tessile
- *La fabbricazione di assorbenti in ovatta di cotone* va nella categoria 17.54.1
- *La confezione di maglieria* va nella categoria 17.72.0
- *Attività di finissaggio conto terzi* anche se non viene specificato di tessili va nella categoria 17.30.0

Sono state messe nella ***divisione generica 17*** le seguenti attività:

17 - (ei) Annodatore meccanico presso terzi
17 - (ei) Annodatore tele
17 - (ei) Annodatura
17 - (ei) Annodatura conto terzi non filatura
17 - (ei) Annodatura di subbi per tessitura
17 - (ei) Annodatura meccanica filati
17 - (ei) Annodatura tele conto terzi
17 - (ei) Annodino settore tessile
17 - (ei) Confezione tessili in proprio e c/terzi
17 - (ei) Lavorazione fibre tessili artificiali c/t
17 - (ei) Manifattura lane
17 - (ei) Produzione e commercio di filati e materie tessili
17 - (ei) Produzione prodotti tessili in genere

Sono rimaste invece nella ***categoria*** 17.17.0 le seguenti attività intermedie tra la filatura e la tessitura che sono di non facile interpretazione:

17.17.0 - (ei) Aspatura roccatura
17.17.0 - (ei) Attività preparazione e filatura altre fibre
17.17.0 - (ei) Cernita roccatura fili sintetici
17.17.0 - (ei) Dipanatura c/t
17.17.0 - (ei) Dipanatura fibre
17.17.0 - (ei) Dipanatura filati c/p c/t
17.17.0 - (ei) Dipanatura filati c/t
17.17.0 - (ei) Dipano roccatura
17.17.0 - (ei) Eliminazione dei difetti sul filato prodotti dalla filatura
17.17.0 - (ei) Floccatura
17.17.0 - (ei) Garmettatura e sfilacciatura
17.17.0 - (ei) Garmettatura e sfilacciatura stracci
17.17.0 - (ei) Lavorazione filati (roccatura) c/t
17.17.0 - (ei) Mescolatura di lana e fibre tessili
17.17.0 - (ei) Miscelatura ripettinatura c/t
17.17.0 - (ei) Orditura
17.17.0 - (ei) Orditura c/t
17.17.0 - (ei) Orditura e roccatura filato
17.17.0 - (ei) Orditura filati sintetici
17.17.0 - (ei) Orditura filato c/t
17.17.0 - (ei) Orditura prodotti tessili c/t
17.17.0 - (ei) Orditura roccatura
17.17.0 - (ei) Orditura roccatura conto terzi
17.17.0 - (ei) Preparazione aspatura
17.17.0 - (ei) Preparazione e filatura altre fibre tessili juta rafia filati di carta
17.17.0 - (ei) Preparazione e filatura juta rafia filati di carta
17.17.0 - (ei) Produzione cardati fibre acrilico
17.17.0 - (ei) Ritorcitura
17.17.0 - (ei) Ritorcitura conto terzi filati
17.17.0 - (ei) Ritorcitura retrazione roccatura filati per maglierie
17.17.0 - (ei) Ritorcitura vaporizzo
17.17.0 - (ei) Roccatura
17.17.0 - (ei) Roccatura artigiana
17.17.0 - (ei) Roccatura c/p c/t
17.17.0 - (ei) Roccatura c/t
17.17.0 - (ei) Roccatura dipanatura
17.17.0 - (ei) Roccatura e aspatura c/t
17.17.0 - (ei) Roccatura e stribbiatura di filati c/t

17.17.0 - (ei) Roccatura c/t
17.17.0 - (ei) Roccatura dipanatura
17.17.0 - (ei) Roccatura e aspatura c/t
17.17.0 - (ei) Roccatura e stribbiatura di filati c/t
17.17.0 - (ei) Roccatura e stribbiatura elettronica
17.17.0 - (ei) Roccatura filati
17.17.0 - (ei) Roccatura filati c/t
17.17.0 - (ei) Roccatura filati vari tipi
17.17.0 - (ei) Roccatura filo
17.17.0 - (ei) Roccatura ritorcitura
17.17.0 - (ei) Roccatura stribbiatura
17.17.0 - (ei) Sfilacciatura c/t
17.17.0 - (ei) Stribbiatura conto terzi
17.17.0 - (ei) Tagliatura e cardatura fibre sintetiche
17.17.0 - (ei) Torcitura ritorcitura fibre tessili lana cotone acrilico lane viscosa

Un altro grosso problema legato sia alla divisione 17 che alla divisione 22 è quello inerente la *produzione di pannolini ed assorbenti*. Questi infatti possono essere sia nelle materie tessili (17.54.1) che nella cellulosa (21.22.0). Le imprese li producono con entrambi i materiali e quindi non specificano mai. Resta pertanto la necessità di rimarcare la incongruenza nell'Ateco 91.

Nella divisione 18 (Confezione di articoli di vestiario, preparazione e tintura di pellicce)
Sono state effettuate delle duplicazioni delle descrizioni che contenevano la parola *confezione,* sostituendola con *produzione*. Dal campione di dati esaminato, infatti, si è visto che la parola produzione veniva spesso usata come sinonimo di confezione. Ad esempio nella categoria:
18.23.0 Confezione di biancheria personale
si è aggiunta anche la voce
18.23.0 Produzione di biancheria personale
La confezione di maglieria e abbigliamento va nella categoria 18.22.1, è cioè prevalente l'abbigliamento.
Fabbricazione berretti sciarpe guanti a maglia va nella categoria 18.24.2 prevale cioè la produzione di sciarpe e guanti rispetto a quella dei copricapi.
Tutta *la produzione di cinture in pelle e non* va nella categoria 18.24.2 come dice anche la Prodcom.

Nella divisione 19 (Preparazione e concia del cuoio, fabbricazione di articoli da viaggio, borse, articoli da correggiaio, selleria e calzature)
La fabbricazione di articoli da viaggio in qualsiasi materiale (anche abs alluminio) va nella categoria *19.20.0* anche per la Prodcom
La produzione di marocchinerie va nella categoria 19.20.0

Nella divisione 20 (Industria del legno e dei prodotti in legno e sughero, esclusi i mobili)
Si fa presente che solo la fabbricazione è stata equiparata alla produzione; per le operazioni di trasformazione e di lavorazione non si potuto procedere con il sinonimo perché questo avrebbe potuto generare confusione in altre classi . Per esempio:
02.01.1 Taglio e produzione di legno grezzo
20.10.0 Taglio, piallatura e trattamento del legno

Nella divisione 21 (Fabbricazione della pasta carta, della carta e dei prodotti di carta)
La fabbricazione di assorbenti e tamponi igienici in ovatta di cellulosa o carta va nella categoria 21.22.0
Cartotecnica va nella categoria 21.23.0
Cartotecnica e litografia va nella categoria 21.23.0

Nella divisione 22 (Editoria, stampa e riproduzione di supporti registrati)
Cartotecnica e legatoria va nella categoria 22.23.0
Cartotecnica editoriale va nella categoria 22.23.0

Nella divisione 23 (Fabbricazione di coke, raffinerie di petrolio, trattamento dei combustibili nucleari)
Produzione emulsioni conglomerati bituminosi va nella categoria 23.20.4

Nella divisione 24 (Fabbricazione di prodotti chimici e di fibre sintetiche e artificiali)
La produzione di gomme e resine va nella categoria 24.17.0
Prodotti chimici di sintesi sono i prodotti chimici di base inorganica
Per addolcitori d'acqua s'intendono gli anti calcare e vanno nella categoria 24.13.0

Nella divisione 25 (Fabbricazione di articoli in gomma e materie plastiche)
Abs sono materie plastiche
Quando si parla di produzione di *caschi protettivi e motociclistici* si sottintende che siano di plastica e vanno nella categoria 25.24.0

Nella divisione 25 (Fabbricazione di articoli in gomma e materie plastiche)
Abs sono materie plastiche
Quando si parla di produzione di *caschi protettivi e motociclistici* si sottintende che siano di plastica e vanno nella categoria 25.24.0

Nella divisione 26 (Fabbricazione di prodotti della lavorazione di minerali non metalliferi)
I conglomerati bituminosi sono nella categoria 26.82.0
I conglomerati bituminosi e cementizi *sono nella categoria 26.63.0*
Nella divisione 27 (Produzione di metalli e loro leghe)
Allumina sinterizzata è uguale *ossido di alluminio* va pertanto nella categoria 27.42.0

Nella divisione 28 (Fabbricazione e lavorazione dei prodotti in metallo, escluse macchine e impianti)
Riparazione assistenza caldaie a gas riscaldamento centrale va nella categoria 28.22.0
Produzione di targhe ed insegne stradali va nella categoria 28.75.3
Fabbricazione di griglie e reti di fili di ferro o di acciaio saldati va nella categoria 28.73.0

Nella divisione 29 (Fabbricazione di macchine ed apparecchi meccanici compresi l'installazione il montaggio la riparazione e la manutenzione)
Le conchiglie sono dei particolari stampi per la pressofusione e vanno nella categoria 29.56.3

Nella divisione 30 (Fabbricazione di macchine per ufficio di elaboratori e sistemi informatici)
Il riciclaggio toner, cartucce, nastri stampa va nella categoria 30.02.0.

Nella divisione 31 (Fabbricazione di macchine e apparecchi elettrici n,c,a,)
La fabbricazione di tutte le parti elettriche degli autoveicoli va nella categoria 31.61.0
Il cablaggio di impianti elettrici va nella categoria 31.20.1
Il cablaggio di cavi elettrici va nella categoria 31.30.0

Nella divisione 32 (Fabbricazione di apparecchi radiotelevisivi e di apparecchiature per le comunicazioni)
La fase dei sistemi di cablaggio resta di difficile attribuzione
Il *cablaggio impianti elettronici* va nella categoria 32.10.0
Nella divisione 33 (Fabbricazione di apparecchi medicali, di apparecchi di precisione di strumenti ottici e di orologi)
Laboratorio di odontotecnico va nella categoria 33.10.3
Laboratorio oftalmico va nella categoria 33.40.2

Nella divisione 34 (Fabbricazione di autoveicoli, rimorchi e semirimorchi)
La fabbricazione di autoveicoli va nella categoria 34.10.0
La fabbricazione di autocisterne, sebbene per Ateco 91 sia nella divisione 34.20.0, sulla base della nuova interpretazione Nace è stata spostata nella categoria 34.10.0 dove si trovano tutti gli autoveicoli.
La fabbricazione di autopompe va nella categoria 34.10.0.

Nella divisione 35 (Fabbricazione di altri mezzi di trasporto)
Cantiere navale va nella categoria 35.11.1
Quando si parla di *Barche* senza nessuna specifica si sottintende che siano da *Diporto*

Nella divisione 36 (Fabbricazione di mobili altre industrie manifatturiere)
La rigenerazione nastri stampanti, nastri toner, supporti di stampa va nella categoria 36.63.6

La produzione di reti da letto è contemplata temporaneamente nella fabbricazione di materassi (categoria 36.15.0), non essendo le reti menzionate nella classificazione Istat ma nel Repertorio Merceologico della Rilevazione Annuale della Produzione Industriale.
Per ciò che riguarda invece la doppia attività di *ebanista restauratore* va nella categoria 36.14.1.
Per l'attività di ***restauro*** si è stabilito che va nella categoria del prodotto restaurato.
Il restauro di mobili di legno pertanto va nella categoria 36.14.1.
Fabbricazione di mobili per l'arredamento si sottintende per la casa, pertanto va nella categoria 36.14.1.

Nella divisione 37 (Recupero e preparazione per il riciclaggio)
Autodemolizioni e commercio ricambi usati va nella categoria 37.10.0

Cernita stracci va nella categoria 37.20.2

Riciclaggio rifiuti va nella categoria 37.20.2

Nella divisione 40 (Produzione di energia elettrica, di gas di vapore e acqua calda)
Ente nazionale energia elettrica va nella categoria 40.10.0
Azienda gas municipale va nella categoria 40.20.1
Gestione centrali di teleriscaldamento va nella categoria 40.30.0

Nella divisione 41 (Raccolta depurazione e distribuzione d'acqua)
Gestione acquedotti va nella categoria 41.00.1
Gestione acqua irrigua va nella categoria 41.00.2

Gestione centrali di teleriscaldamento va nella categoria 40.30.0

Nella divisione 41 (Raccolta depurazione e distribuzione d'acqua)
Gestione acquedotti va nella categoria 41.00.1
Gestione acque irrigue va nella categoria 41.00.2

Nella divisione 45 (Costruzioni)
In questa divisione della classificazione si è usato il criterio di assegnare il codice a cinque digit, quando possibile, assegnando i codici relativi alla categoria altro…..
Per esempio :
45.4 *Lavori di completamento di edifici*
è diventato
45.45.2 *Altri lavori di completamento di edifici*
ed in generale
45.21.0 *Edilizia abitativa*
In questa divisione si trovano quasi tutti i lavori di *installazione* come per esempio del telefono, del gas…
Demolizione e ricostruzione di edifici va nella categoria *45.21.0*
Lavori di isolamento edile ed impermeabilizzazione va nella categoria *45.32.0*
Riparazione assistenza caldaie a gas da appartamento va nella categoria *45.33.0* in base ad una specifica Nace

Nella divisione 50 (Commercio, manutenzione e riparazione di autoveicoli e motocicli; vendita al dettaglio di carburante per autotrazione)
Autofficina prevale rispetto al soccorso stradale va nella categoria 50.20.1
Autofficina (autoriparazioni) e vendita ricambi va nella categoria 50.20.1
Autofficina e autorimessa va nella categoria 50.20.1
Autocarrozzeria e soccorso stradale va nella categoria 50.20.2
Concessionaria auto ricambi officina va nella categoria 50.10.0
La riparazione delle auto >riparazione mezzi agricoli
Riparazione e commercio pneumatici va nella categoria 50.20.4
Riparazione e commercio dettaglio pneumatici va nella categoria 50.20.4
Riparazione e commercio ingrosso pneumatici va nella categoria 50.30.0
Vendita di carburanti e vendita gomme e accessori auto va nella categoria 50.50.0
La vendita di motocicli e di biciclette va nella categoria 50.40.1
La vendita dei motocicli e di accessori va nella categoria 50.40.1

Nella divisione 51 (Commercio ingrosso e intermediari del commercio, autoveicoli e motori esclusi)
In generale il *Commercio all'Ingrosso* è prevalente sul *Commercio al Dettaglio*
Intermediari commercio con deposito è inteso come *commercio all'ingrosso*
Commercializzazione è intesa come commercio all'ingrosso
Etichettamento e vendita è inteso come commercio all'ingrosso
Commercio a stock è inteso come commercio all'ingrosso
Tutto il commercio, sia all'ingrosso che al dettaglio, esclude quello di autoveicoli e di motocicli
Non esiste il commercio all'ingrosso fatto con distributori automatici
Non esiste il commercio di libri usati all'ingrosso
Rappresentanze con deposito vanno intese come *commercio despecializzato* e vanno nella categoria 51.70.0
Commercio all'ingrosso senza specifica di prodotto va nella categoria *51.70.0*
Intermediari commercio all'ingrosso senza specifica di prodotto va nella categoria 51.19.0
La gastronomia all'ingrosso va nella categoria 51.39.4
Commercio ingrosso lattiero caseari salumi va nella categoria 51.33.1
Agenzia di commercio despecializzata va nella categoria 51.19.0
Agenzia di commercio specializzata va nella categoria 51.18.0
Agente di commercio va nella categoria 51.19.0
Intermediario del commercio va nella categoria 51.19.0
Esiste il commercio all'ingrosso di indumenti usati (sono gli Stockisti che importano dagli Stati Uniti, vedi mercato di Latina) e va nella categoria 51.42.1
Il consorzio agrario per convenzione consolidata e riscontrata va nella categoria 51.55.0 data la prevalenza nella loro attività di vendita all'ingrosso di fertilizzanti, diserbanti, fitofarmaci
Il commercio all'ingrosso di mobili per ufficio va nella categoria 51.47.1
Gli apparecchi sanitari sono articoli medicali venduti per lo più all'ingrosso
Non esiste il commercio all'ingrosso fatto con distributori automatici

Per taluni prodotti non è stato necessario specificare il tipo di commercio, ingrosso o dettaglio, perché difficilmente commerciabili al dettaglio essi sono:

- Nella categoria 51.21.1: commercio settore risicolo
- Nella categoria 51.21.2: alimenti per animali da allevamento, mangimi, mangimi cereali, paglia fieno, prodotti per l'agricoltura sementi, prodotti zootecnici,
- Nella categoria 51.22.0: export fiori, importazione esportazione bulbi e talee

- Nella categoria 51.21.1: commercio settore risicolo
- Nella categoria 51.21.2: alimenti per animali da allevamento, mangimi, mangimi cereali, paglia fieno, prodotti per l'agricoltura sementi, prodotti zootecnici,
- Nella categoria 51.22.0: export fiori, importazione esportazione bulbi e talee
- Nella categoria 51.23.2: bovini suini vivi, bestiame, suini vivi
- Nella categoria 51.24.1: Import export pelli finite e pellame, pelli grezze
- Nella categoria 51.31.0: prodotti agricoli, distribuzione ortofrutta
- Nella categoria 51.32.3: prosciutti
- Nella categoria 51.33.1: derivati del latte burro, stagionatura confezionamento formaggio grana e prodotti alimentari, distribuzione uova e ovoprodotti, stagionatura e commercio gorgonzola
- Nella categoria 51.33.2: olio, olio conferito dai soci, olio vegetale
- Nella categoria 51.34.1: commercio e imbottigliamento vini, commercio vini non prodotti, esportazione vini
- Nella categoria 51.34.2: acqua oligominerale, acque gasate, concessionario di bevande, distribuzione bevande
- Nella categoria 51.35.0: magazzino vendita generi monopolio di stato
- Nella categoria 51.38.2: distribuzione prodotti alimentari
- Nella categoria 51.39.2: importazione di olio di fegato di merluzzo
- Nella categoria 51.39.3: etichettamento e vendita derivati pomodoro
- Nella categoria 51.39.4: sale (per uso strade, alimentazione animali ecc.), materie prime alimentari per bar pasticceria, distribuzione di pasta fresca
- Nella categoria 51.41.1: tessuto grezzo, tessuti stock
- Nella categoria 51.41.4: pellame vegetale
- Nella categoria 51.42.1: esportazione articoli abbigliamento
- Nella categoria 51.42.2: compravendita pellicce
- Nella categoria 51.42.4: import export calzature
- Nella categoria 51.42.5: import export di abbigliamento e maglieria
- Nella categoria 51.43.4: apparati telecomunicazioni, prodotti per telecomunicazioni
- Nella categoria 51.43.5: quadri elettrici, semiconduttori e microsistemi, elettroforniture
- Nella categoria 51.44.2: forniture alberghiere di porcellane e vetrerie
- Nella categoria 51.46.1: deposito con rappresentanze farmaceutiche, deposito rappresentanza medicinali, distribuzione rappresentante prodotti farmaceutici, gestione distribuzione intermedia farmaci, importazione prodotti medicinali sieri vaccini emoderivati
- Nella categoria 51.46.2: gas tecnici puri ossigeno e medicinali, lastre radiografiche, leghe dentarie e materiale odontoiatrico, prodotti ospedalieri, distribuzione dispositivi apparecchi medicali, fornitura strumenti medicali
- Nella categoria 51.47.1: arredi per pubblici esercizi, importazione e esposizione mobili, arredamenti negozi
- Nella categoria 51.47.2: carta
- Nella categoria 51.47.5: forniture orologeria, pietre per oreficeria
- Nella categoria 51.51.1: prodotti petroliferi
- Nella categoria 51.51.3: gas kerosene
- Nella categoria 51.52.1: acciai, materiali ferrosi, metalli ferrosi, metalli ferrosi semilavorati, prodotti siderurgici
- Nella categoria 51.52.3: barre bronzo, metalli non ferrosi, metalli preziosi e semilavorati, semilavorati e rottami di alluminio
- Nella categoria 51.53.1: legname, legnami esteri, legnami travi ferro, prefabbricati in legno, trucioli legno, legnami
- Nella categoria 51.53.2: blocchi marmo bianco venato, blocchi marmo Carrara, calce idrata, canale tubi rame, laterizi, marmi e pietre, marmi graniti pietre, marmi graniti pietre lapidei onici travertini, pietrame, sabbia ghiaia inerti, conglomerati bituminosi, manufatti di gesso, inerti conglomerati calcestruzzo
- Nella categoria 51.53.5: accessori e materie prime per la produzione di isolante, grassello di calce, scale finestre caminetti
- Nella categoria 51.54.2: impianti riscaldamento e condizionamento, impianti termo tecnici, materiale termoidraulico, fornitura per impianti termici
- Nella categoria 51.55.0: concimi, prodotti chimici petrolchimici e materie plastiche, fertilizzanti fitofarmaci, granuli termoplastici, materie prime per uso farmaceutico e cosmetico, prodotti chimici industria, prodotti per l'agricoltura fertilizzanti, riattivazione carboni attivi
- Nella categoria 51.56.1: materie prime tessili
- Nella categoria 51.57.1: rottami ferrosi e non ferrosi, recupero ferrosi, rottami ferrosi e non, rottami metallici, rottami parti meccaniche, trasformazione rottami, rottami autotrasporti
- Nella categoria 51.57.2: scarti lavorazione pelle
- Nella categoria 51.57.3: biomasse, recupero carta da macero imballaggi in cartone metallo

- Nella categoria 51.57.1: rottami ferrosi e non ferrosi, recupero ferrosi, rottami ferrosi e non, rottami metallici, rottami parti meccaniche, trasformazione rottami, rottami autotrasporti
- Nella categoria 51.57.2: scarti lavorazione pelle
- Nella categoria 51.57.3: biomasse, recupero carta da macero imballaggi in cartone metallo legno e plastica, materiali vari recupero, recupero smaltimento rifiuti speciali ferro carta legno, stracci, rottami non ferrosi
- Nella categoria 51.61.0: macchine lavorazione legno, macchine utensili, macchine fresatrici, macchine lavorazione legno e alluminio
- Nella categoria 51.62.0: gru pedane caricatrici, macchine edili, mezzi sollevamento, noleggio macchine movimento terra
- Nella categoria 51.64.1: hardware, sistemi telematici, assistenza registratori-di cassa
- Nella categoria 51.64.2: scaffalature terminali sistemi presenze
- Nella categoria 51.65.0: apparati elettronici, apparecchiature pneumatici, attrezzature accessori per industrie materie plastiche, attrezzature industriali, attrezzature per garage, automatismi elettrici, bilance affettatrici, carrelli elevatori manutenzione attrezzature per magazzino e interni, componenti elettronici, cuscinetti a sfere e affini, cuscinetti articoli tecnici, articoli tecnici per forniture industriali, attrezzature macchinari, installazione macchine per la gastronomia, isolanti termoelettrici, macchinari per l'enologia, macchine attrezzature elettroniche e ottiche, macchine e accessori per calzaturifici, macchine industriali, macchine per conceria, macchine per il trattamento della carta, macchine per industria grafica, macchine per marmi, macchine pulizia, materiale ferroviario, materiale rotabile ferroviario, ricambi industriali, ricambi rotabili ferroviari, riparazione bilance affettatrici attrezzature per la trasformazione alimentare, sistemi fibre ottiche, strumentazioni e componenti elettrici e elettronici in genere, utensili per fornaci, veicoli carrelli elevatori, lettori banda magnetica, assistenza sistemi per l'automazione industriale, attrezzature panifici panetterie, bilance misuratori fiscali, attrezzature per oleifici, assistenza attrezzature professionali per la ristorazione e il commercio, assistenza carrelli elevatori, impianti refrigeranti, ricambi presse ceramica, strumentazione per rilevamento inquinamento atmosferico, strumenti di pesatura
- Nella categoria 51.70.0: articoli per equitazione, gas materiali saldatura merci varie, imballaggi o parte di essi, imballi in legno e non

Nella divisione 52 (Commercio al dettaglio, escluso quello di autoveicoli e di motocicli; riparazione di beni personali per la casa)

Per il *commercio al dettaglio*:

Nella codifica si sono considerati sempre il negozio, il negoziante, dettagliante e la rivendita come commercio al dettaglio.

Inoltre:

Commercio al pubblico = commercio dettaglio

Commercio al minuto = commercio al dettaglio

Punto vendita = commercio dettaglio

Commercio dettaglio fisso viene inteso come *commercio al dettaglio*

Nel **commercio al dettaglio** non esiste la fase di *etichettamento*

Commercio al dettaglio per mezzo di agenti di vendita viene classificato nel commercio al dettaglio in base al prodotto venduto. Se il prodotto non viene specificato va nella categoria 52.48.9

Rivendita materiali edili va nella categoria 52.46.4

Vendita di libri senza ulteriore specifica va nella categoria 52.47.1

Vendita di articoli di cartoleria senza ulteriore specifica va nella categoria 52.47.3

La vendita di generi di monopolio e predominante su, cartoleria, cartolibreria, ricevitoria lotto e totocalcio, mercerie e bazar.

La vendita di generi di monopolio ed edicola va invece nella categoria 52.47.2

Commercio al dettaglio di mobili antichi è predominante rispetto *alla vendita di oggetti di antiquariato* e va nella categoria 52.50.2

Commercio al dettaglio di antiquariato è considerato come *mobili antichi* e va nella categoria 52.50.2

Il dispensario farmaceutico è inteso al minuto

La vendita di porchetta sottintende l'ambulante

Commercio dettaglio bevande sottintende non alcoliche

La vendita di ricambi ciclo va nella categoria 52.48.5

Riparazione di elettrodomestici va nella categoria 52.72.0

Negozio di generi alimentari va nella categoria 52.11.4

Produzione di pane e dettaglio alimentari va nella categoria 52.11.4 è cioè prevalente la vendita degli alimentari

Dettaglio di pane e alimentari va nella categoria 52.11.4 è cioè prevalente la vendita degli alimentari

Dettaglio solo pane va nella categoria 52.24.1

Commercio minuto di generi alimentari e bevande va nella categoria 52.27.4

Dettaglio di pane e alimentari va nella categoria 52.11.4 e cioè prevalente la vendita degli alimentari
Dettaglio solo pane va nella categoria 52.24.1
Commercio minuto di generi alimentari e bevande va nella categoria 52.27.4
Il commercio di pane +alimentare va nella categoria 52.11.4
Vendita di carni fresche si intendono sempre bovine suine
Commercio di calzature, pelletteria valigeria va nella categoria 52.43.1 (la vendita delle scarpe è prevalente)
Commercio minuto di casalinghi, giocattoli, articoli da regalo va nella categoria 52.44.5
Galleria d'arte è intesa come commercio dettaglio oggetti d'arte e va nella categoria 52.48.6
La vendita di antiquariato è prevalente *su cose usate e oggetti usati* e va nella categoria 52.50.2
La vendita minuto di articoli da pesca è prevalente *sulla vendita di attrezzature e animali domestici* e va nella categoria 52.48.5

In base alle indicazione avute dal CUE la parola '*Negozio',* come già detto, sottintende il *'Commercio al dettaglio'* pertanto è stato introdotto il sinonimo di *negozio/negoziante =dettaglio*.
Differentemente dal commercio all'ingrosso, dove sono state inserite *empiriche* nella classe corrispondente dell'ingrosso associate solo alla parola commercio con prodotti specifici, per il commercio al dettaglio sono state inserite nei file di *parsing* le seguenti trasformazioni che hanno permesso di considerare sempre come dettaglio il commercio associato a taluni prodotti o attività:
COMMERCIO MACELLERIA--à *DETTAGLIO MACELLERIA*
COMMERCIO MERCERIA ---à *DETTAGLIO MERCERIA*
COMMERCIO CHINCAGLIERIA--à *DETTAGLIO CHINCAGLIERIA*
COMMERCIO RIPARAZIONE--à *DETTAGLIO RIPARAZIONE*

Per il *commercio ambulante*:
Commercio ambulante senza altre specifiche = commercio ambulante a posteggio fisso pertanto va nella categoria 52.62.7
Questo criterio è stato ricavato dal CUE in base all'analisi dei dati LongForm: su 352 record con 'ambulante', infatti, 42 hanno specificato ' a posteggio fisso' e 8 'a posteggio mobile'. Dato che sia la Classificazione Ateco 91 che la NACE Rev.1 concordano che per '*posteggio fisso*' s'intende '*il commercio al dettaglio di qualsiasi tipo di prodotto in banchi, di solito smontabili, situati su una strada pubblica oppure in un posto fisso al mercato*' e che per 'posteggio mobile' s'intende solo quello su *mezzi mobili*, allora si è stabilito che, se la descrizione dell'attività contiene solo '*commercio ambulante',* senza la specifica di *fisso*' o '*mobile',* di assegnare l'attività di *ambulante a posteggio fisso* a seconda dei prodotti indicati, altrimenti, se manca anche la specifica del prodotto, è codificata nella categoria generica 52.62.7 come detto sopra.
La vendita di porchetta sottintende il commercio ambulante
Eccezione
Commercio ambulante legna da ardere va nella categoria 52.63.5

Per il commercio diretto
Commercio porta a porta= commercio diretto
Nella divisione 55 (Alberghi e ristoranti)
Nelle attività doppie che riguardano ***Alberghi e Ristoranti*** si è stabilito che:
Ristorante> bar
Albergo > ristorante
Hotel > villaggio
Mensa>catering
Attività alberghiera>attività termale
Ristorazione senza nessuna specifica va nella divisione 55
Attività di ristorazione va nella divisione *55*
Albergo senza nessuna specifica va nella categoria 55.12.0
Garni' è l'affitto della stanza con la prima colazione sulle Dolomiti
Nelle attività doppie che riguardano il ***Bar*** si è stabilito che :
Bar>gelateria
Bar>pasticceria
Bar>pizzeria a taglio
Bar>paninoteca
Bar>panineria
Bar>osteria
Bar>birreria
Bar tabacchi>bar
Bar >night club e va nella categoria 55.40.4
Gestione discoteca bar va nella categoria 92.34.1
Bar discoteca spettacoli va nella categoria 55.40.4
Bar trattenimenti danzanti va nella categoria 55.40.4
Le stesse prevalenze valgono per il caffè inteso come bar .

Nella divisione 60 (Trasporti terrestri; trasporti mediante condotte)

Bar discoteca spettacoli va nella categoria 55.40.4
Bar trattenimenti danzanti va nella categoria 55.40.4
Le stesse prevalenze valgono per il caffè inteso come bar .

Nella divisione 60 (Trasporti terrestri; trasporti mediante condotte)
I trasporti regolari non sono mai a noleggio
Poiché non sono stati segnalati dal Servizio CUE altri criteri particolari o *empiriche* da inserire nel dizionario informatizzato, la divisione contiene unicamente le voci della classificazione ufficiale opportunamente rielaborate come detto sopra.

Nella divisione 61 (Trasporti marittimi e per vie d'acqua)
Trasporti marittimi locali=costieri e vanno nella categoria 61.12.0
Trasporti marittimi>trasporti per vie costiere
Pilota di aliscafo va nella categoria 61.12.0 perché viene inteso come trasporto costiero

Nella divisione 62 (Trasporti aerei)
Poiché non sono stati segnalati dal Servizio CUE criteri particolari la classe contiene unicamente le voci della classificazione ufficiale opportunamente rielaborate come detto sopra. Sono state inserite cinque sole *empiriche*

Nella divisione 63 (Attività di supporto ed ausiliarie dei trasporti; attività delle agenzie di viaggio)
Autorimessa >soccorso Aci va nella categoria 63.21.0
Il facchinaggio inteso come manovalanza, carico scarico va nella categoria 63.11.3
I servizi di facchinaggio vanno nella categoria 63.11.3
Le attività di pilotaggio e ancoraggio all'interno del porto vanno nella categoria 63.22.0

Nella divisione 64 (Poste e telecomunicazioni)
Poiché non sono stati segnalati dal Servizio CUE criteri particolari la classe contiene le voci della classificazione ufficiale opportunamente rielaborate come detto sopra. Si è poi implementato il dizionario con l'inserimento di varie *empiriche* (circa 58).

Nella divisione 65 (Intermediazione monetarie e finanziaria)
Poiché non sono stati segnalati dal Servizio CUE criteri particolari la classe contiene le voci della classificazione ufficiale opportunamente rielaborate come detto sopra. Si è poi implementato il dizionario con l'inserimento di varie *empiriche* (circa 85).

Nella divisione 66 (Assicurazione e fondi pensione, escluse le assicurazioni sociali obbligatorie)
Poiché non sono stati segnalati dal Servizio CUE criteri particolari la classe contiene le voci della classificazione ufficiale opportunamente rielaborate come detto sopra. Si è poi implementato il dizionario con l'inserimento di varie *empiriche* (circa 12).

Nella divisione 67 (Attività ausiliarie dell'intermediazione finanziaria)
L'attività *di perito e di assicuratore predomina* rispetto all'assicurazione e va nella categoria 67.20.2

Nella divisione 70 (Attività immobiliari)
L'attività immobiliare generica senza specificare se su beni propri o per conto terzi, si è stabilito di considerala sempre come una attività per conto terzi**.**
Cooperative edilizie e costruzione di alloggi per soci va nella categoria 70.11.0

Nella divisione 71 (Noleggio di macchinari e attrezzature senza operatore e di beni per uso personale e domestico)
Noleggio a freddo = noleggio senza operatore

Attività di noleggio e commercio nel caso di *videocassette e cd* predomina l'attività di noleggio e va nella categoria 71.40.2
Attività di gestione e noleggio nel caso di *giochi d'intrattenimento* predomina il noleggio e va nella categoria 71.40.2

Nella divisione 72 (Informatica ed attività annesse)
Si è convenuto di considerare per *sistemi informatici* anche l'hardware pertanto per esempio

Consulenza tecnica di sistemi informatici va nella categoria 72.10.0

Nella divisione 73 (Ricerca e sviluppo)
Biologo marino va nella classe della ricerca 73.10.0 in quanto l'attività solo tecnica propria della categoria 74.20.6 è più rara

Nella divisione 74 (Altre attività professionali ed imprenditoriali)
Tutte le attività che contemplano *la pubblicità* vann*o* nella classe 74.40, come da accordi con la Prodcom
Produzione e noleggio stand fieristici va nella categoria 71.34.0
Produzione di targhe ed insegne stradali pubblicitarie va nella categoria 74.40.1
La pulizia e il facchinaggio vanno nella categoria 74.70.1

Nella divisione 75 (Pubblica amministrazione e difesa; assicurazione sociale obbligatoria)

Produzione di targhe ed insegne stradali pubblicitarie va nella categoria 74.40.1
La pulizia e il facchinaggio vanno nella categoria 74.70.1

Nella divisione 75 (Pubblica amministrazione e difesa; assicurazione sociale obbligatoria)
La classe ha subito ampliamenti meno consistenti rispetto alle altre in quanto la Pubblica amministrazione non è oggetto di rilevazione nelle indagini utilizzate per l'addestramento (Short Form, Long Form, Censimento dell'industria). Nonostante questo il dizionario informatizzato contiene le *empiriche* inserite ricavate in massima parte da precedenti esperienze sul Censimento della popolazione.

Nella divisione 80 (Istruzione)
La scuola di musica senza nessuna specifica va nella categoria 80.42.2
L'*asilo sia pubblico che privato* è stato considerato come *scuola materna* e va nella categoria 80.10.1
L'*asilo nido* invece va nella categoria 85.32.0

Nella divisione 85 (Sanità ed altri servizi sociali)
Il medico generico va nella categoria 85.12.2
Il medico condotto va nella categoria 85.12.1
Il presidio sanitario privato va nella categoria 85.12.3
Lo psicologo professionista va nella categoria 85.32.0
Il medico psicologo va nella categoria 85.12.3
Per l'attività che riguarda solo l'attività di ***psicologo*** si è stabilito che:
85.32.0 - (ei) Attivita professionale svolta da psicologi
85.32.0 - (ei) Libera professione psicologa
85.32.0 - (ei) Psicanalista
85.32.0 - (ei) Psicoanalisi
85.32.0 - (ei) Psicologia clinica psicoterapia
I servizi di assistenza agli anziani senza ulteriore specifica vanno nel gruppo 85.3
I servizi di assistenza agli anziani ed ammalati vanno nella categoria 85.31.0 in quanto si sottintende l'assistenza residenziale

Nella divisione 90 (Smaltimento di rifiuti solidi, delle acque di scarico e simili)
Bonifica residuati bellici va nella categoria 90.00.1 (la nuova voce riporta anche la bonifica di terreni inquinati)
Le *bonifiche ambientali* vanno nella categoria 90.00.1 (come da nuova Nace)

Nella divisione 91 (Attività di organizzazioni associative n.c.a.)
Centro lega antidroga va nella categoria 91.33.0

Nella divisione 92 (Attività ricreative culturali e sportive)
Attività culturali e ricreative senza altre specifiche vanno nella divisione 92
Il restauratore d'arte generico va nella categoria 92.31.0
Relativamente a questa divisione il dizionario si è implementato con l'aggiunta di numerose empiriche (circa 498)

Nella divisione 93 (Altre attività dei servizi)
L'attività di parrucchiere è predominante sul solarium e va nella categoria 93.02.2

Nella divisione 95 (Servizi domestici presso famiglie e convivenze)
Poiché non sono stati segnalati dal Servizio CUE criteri particolari la classe contiene le voci della classificazione ufficiale opportunamente rielaborate come detto sopra. Si è poi implementato il dizionario con l'inserimento di varie *empiriche* (circa 16)

Nella divisione 99 (Organizzazioni ed organismi extraterritoriali)
Si è inserita una sola *empirica*

99.00.0 - (ei) Ambasciata

5.4 Informazioni di carattere generale

Nell'ambito poi delle varie divisioni permangono queste informazioni di carattere generale:

- *Se non è specificato il metallo si presume il ferro*
- *Estrusione è un processo di lavorazione per preparare tubi, lastre, barre, profilati*
- *Per le barche* si sottintende che siano da *diporto*
- *Oli non commestibili* si intendono gli oli vegetali come quelli di sansa finali
- *Oli essenziali* si intendono come prodotti chimici
- *Lavorazione stocco* deve essere inteso come lavorazione dello stoccafisso, baccalà
- *Addolcitori acqua* sono gli anticalcare
- *Ovoprodotti* come dice la Prodcom sono albumina di uova, tuorli congelati, essiccati, freschi

- *Lavorazione stocco* deve essere inteso come lavorazione dello stoccafisso, baccalà
- *Addolcitori acqua* sono gli anticalcare
- *Ovoprodotti* come dice la Prodcom sono albumina di uova, tuorli congelati, essiccati, freschi liquidi (particolarmente nell'industria alimentare); uova di volatili essiccate conservate ecc.
- *Le motoseghe* vengono considerate accessori e utensili agricoli
- *Gli articoli igienico-sanitari* sono bagni e docce intesi come materiali da costruzione
- *Rottami* sono più che altro metallici
- *Sistema operativo=software*
- *Quando si parla di carburante per combustione* è sottinteso per uso domestico
- *Quando si parla di progettazione e realizzazione (produzione)* predomina la realizzazione
- *Engineering = ingegneria integrata*
- *Guardia giurata=sorveglianza privata*

5.5 Considerazioni finali

Da quanto esposto sopra emerge pertanto che, nelle **attività doppie in cui è presente la produzione**, intesa anche come **fabbricazione,** essa risulta essere quasi sempre l'**attività prevalente**.

Per le attività doppie in cui sono presenti la **fabbricazione e manutenzione** non si è potuto dare una priorità univoca perché per alcuni prodotti la categoria legata alla fabbricazione è diversa dalla categoria di fabbricazione e manutenzione. Per esempio:

30.01.0 Fabbricazione di macchine per ufficio (esclusa riparazione)

72.50.0 Manutenzione e riparazione macchine per ufficio e di elaboratori elettronici

Per ciò che riguarda invece le **attività doppie legate al commercio** si è stabilito che:

- *Macellazione > commercio*
- *Tranciatura commercio>tranciatur*a
- *La vendita prevale rispetto al rimessaggio*

5.6 Trattamento delle risposte assolutamente non significative

L'analisi dei casi rilevati ha dimostrato una casistica di risposte prive di un contenuto sufficiente per l'attribuzione di un codice seppur generico, che tuttavia possono presentare il rischio di essere associate a codici particolari (l'errore di codifica può derivare dal fatto che sono spesso risposte molto brevi rappresentate da parole piuttosto rare, alle quali l'algoritmo per l'attribuzione dei pesi alle parole assegna un valore elevato).

Per ovviare a tale problematica, queste probabili risposte sono state inserite nel dizionario associate alla sigla *n.c.*, che sta per 'non codificabile'.

L'elenco di tali attività è stato fornito dal CUE e sono:

n.c - (ep) Famiglia
n.c - (ep) Pavimenti industriali
n.c - (ep) Terzo settore
n.c - (ep) Alleanza nazionale
n.c. - (ei) Conto lavoro per terzo
n.c. - (ei) Servizi svolti per conto soci
n.c. - (ep) Arredi sanitari
n.c. - (ep) Azienda addetti alle vendite
n.c. - (ep) Azienda commerciale
n.c. - (ep) Azienda impianti
n.c. - (ep) Azienda speciale multiservizi
n.c. - (ep) Società di distribuzione
n.c. - (ep) Ufficio tecnico
n.c. - (ep) Vendita ed assistenza tecnica
n.c. - (nc) Abbigliamento
n.c. - (nc) Altro fibra
n.c. - (nc) Attivita' autonoma
n.c. - (nc) Attivita' non specializzata
n.c. - (nc) C/p c/t
n.c. - (nc) Caccia pesca
n.c. - (nc) Cessata in liquidazione
n.c. - (nc) Civile
n.c. - (nc) Commercio e confezione di propria produzione
n.c. - (nc) Commercio parti ed accessori
n.c. - (nc) Confezione
n.c. - (nc) Confezione industriale
n.c. - (nc) Confezione riparazione

n.c. - (nc) Commercio parti ed accessori
n.c. - (nc) Confezione
n.c. - (nc) Confezione industriale
n.c. - (nc) Confezione riparazione
n.c. - (nc) Confezioni
n.c. - (nc) Confezioni c/t
n.c. - (nc) Consorzio
n.c. - (nc) Conto
n.c. - (nc) Conto socio
n.c. - (nc) Cooperativa sociale
n.c. - (nc) Costruzioni
n.c. - (nc) Fiduciarie
n.c. - (nc) Fiore
n.c. - (nc) Impiantistica
n.c. - (nc) Impiegato statale
n.c. - (nc) In liquidazione
n.c. - (nc) Intimo
n.c. - (nc) Intrecciati
n.c. - (nc) Lavorazione autonoma
n.c. - (nc) Lavorazione in proprio c/terzi
n.c. - (nc) Lucidatura
n.c. - (nc) Lucidatura c/t
n.c. - (nc) Manutenzione e riparazione apparecchiature
n.c. - (nc) Miscelazione
n.c. - (nc) Monopolio
n.c. - (nc) Montaggi meccanici
n.c. - (nc) Montaggio
n.c. - (nc) Ora in liquidazione
n.c. - (nc) Pensionato
n.c. - (nc) Prestazioni conto terzi
n.c. - (nc) Produzione
n.c. - (nc) Produzione e lavorazione
n.c. - (nc) Produzione e lavorazione c/t
n.c. - (nc) Produzione e lavorazione c/t parti
n.c. - (nc) Produzione e servizi
n.c. - (nc) Riparazione
n.c. - (nc) Riparazione c/t
n.c. - (nc) Servizio automobilistico
n.c. - (nc) Società in liquidazione
n.c. - (nc) Società in liquidazione volontaria
n.c. - (nc) Ufficio di rappresentanza
n.c. - (nc) Vendita e consulenza

5.7 Problemi e incongruenze Ateco 91

Si segnalano infine alcuni problemi ed incongruenze riscontrate nella classificazione in esame.

- *La produzione di pannolini ed assorbenti* sta sia nelle materie tessili (17.54.1) che nella cellulosa (21.22.0); le imprese li producono contemporaneamente con entrambi i materiali e quindi non specificano mai il materiale
- La *riparazione apparecchi elettrici ed elettronici* rientra nella categoria 32.20.3, se trattasi di *apparati, impianti professionali, attrezzature,* ecc., mentre rientra nella 52.45.1, se trattasi di *elettrodomestici* e nella 52.72.0 se trattasi di apparecchi elettronici, quali i *telefoni cellulari*
- *Finitura di mobili* per Ateco 91 sta nella categoria 36.11.2 (Fabbricazione di poltrone e divani). Poiché tale fase riguarda la finitura di qualsiasi tipo di mobile si è ritenuta più consona la categoria 36.14.1 (Fabbricazione di altri mobili in legno).

Inoltre sono stati riscontrate alcune incongruenze rispetto alla classificazione PRODCOM, relativamente alle quali si rimanda all'allegato 4.

6 L'impatto sull'applicazione dell'Ateco 2001

6.1 Modifiche ATECO '91 - ATECO 2001

La classificazione dell'attività economica 2001 non determina grossi stravolgimenti rispetto a quella precedente; si tratta della trasposizione italiana della NACE Rev. 1.1 che, a sua volta, costituisce un semplice aggiornamento della NACE Rev.1 e non una sua significativa riorganizzazione. L'ATECO '91 è caratterizzata da un altro problema: non è

La classificazione dell´attività economica 2001 non determina grossi stravolgimenti rispetto a quella precedente; si tratta della trasposizione italiana della NACE Rev. 1.1 che, a sua volta, costituisce un semplice aggiornamento della NACE Rev.1 e non una sua significativa riorganizzazione. L'ATECO '91 è caratterizzata da un altro problema: non è perfettamente allineata alla NACE Rev.1; infatti essa è basata sulla versione non emendata della NACE Rev.1 pubblicata sulla G.U. CEE L.293 del 24 ottobre 1990. L'Eurostat e gli altri paesi UE adottano invece la versione NACE Rev.1 emendata con il Regolamento CEE n.761/93 G.U. CEE L.83 del 3 marzo 1993. Ne consegue che la NACE Rev.1 emendata non ha corrispondenza diretta con le prime 4 cifre di 25 codici ATECO '91 (cfr.. Tabella 30).

L'ATECO 2001 costituisce dunque solo una parziale revisione e aggiornamento della classificazione precedente; la revisione vera e propria avverrà nel 2007 in previsione dei Censimenti del 2011. L'ATECO 2001 deve però soddisfare quattro punti fondamentali:

1. Allineamento alla NACE Rev.1 emendata;
2. Comprensione delle nuove attività non ancora esistenti al momento dell'elaborazione della NACE Rev.1;
3. Considerazione delle attività cresciute in importanza dopo l'elaborazione della NACE Rev.1 in seguito a cambiamenti tecnologici o a modifiche della realtà economica;
4. Correzione degli errori presenti nella NACE Rev.1 originaria, errori già evidenti all'epoca e non dovuti a cambiamenti della filosofia dell'operazione.

La NACE Rev.1.1 contiene pochi elementi aggiuntivi. Oltre ad alcuni cambiamenti nei titoli e a modifiche dovute allo scadere del trattato CECA del luglio 2002, i principali cambiamenti sono:

- La ripartizione della NACE 29.40 (fabbricazione di macchine utensili) in tre classi, portatili, per la metallurgia e altre.
- La ripartizione della NACE 40.10 (produzione e distribuzione di energia elettrica) in tre nuove classi, una per la produzione, una per il trasporto e una per la distribuzione e il commercio.
- La ripartizione della NACE 40.20 (produzione di gas, distribuzione di combustibili gassosi mediante condotte) in una classe per la produzione e una classe per la distribuzione e il commercio.
- La ripartizione in due nuove classi: 51.84 e 51.85 della ex classe di commercio all'ingrosso NACE 51.64 (commercio all'ingrosso di macchine e di attrezzature per ufficio) e la ripartizione in due nuove classi:51.86 e 51.87 della ex classe NACE 51.65 (commercio all'ingrosso di altre macchine per l'industria, il commercio e la navigazione).
- Una nuova classe: 74.86 per le attività dei call center.
- Una nuova classe: 72.21 per l'edizione di software.
- La ripartizione della NACE 90.00 (smaltimento dei rifiuti solidi, delle acque di scarico e simili) in tre classi per la raccolta e depurazione delle acque di scarico, la raccolta e lo smaltimento di rifiuti solidi e la pulizia delle aree pubbliche, decontaminazione e disinquinamento dell'ambiente.

Il 14 settembre è stata ufficialmente chiusa la NACE 2001 che sarà in vigore dal gennaio 2002; è stata quindi confermata la bozza sulla quale stavamo già lavorando per definire la quinta cifra dell'ATECO 2001. Si ricorda che la quinta cifra costituisce il dettaglio nazionale della NACE ed è quella sulla quale lavorare per cogliere le sfumature del proprio sistema produttivo e di servizi

L'ATECO 2001 è soggetta alle seguenti modificazioni:

1. Recepire le correzioni della Nace Rev.1.1, come già visto sopra, non solo in termini di modificazione di codici ma soprattutto in termini di descrizione diverse, più complete e più aggiornate in linea con le esigenze delle attività economiche che si sono venute modificando in questo arco temporale;
2. Lavoro di eliminazione dei settori che sono cambiati o che sono diventati inutili;
3. Rivedere la V cifra nel complesso e inserire eventuali nuove attività.

Nonostante si pensi di limitare al minimo il terzo punto (ad es., l'agro-alimentare deve essere completamente rivisto ma si pensa di farlo direttamente per il 2007), occorre comunque provvedere a considerare tutte le nuove attività; in particolare il settore che è maggiormente soggetto a revisione è quello del Credito e delle Assicurazioni.

In conclusione, nonostante l'ATECO 2001 si presenti come un'operazione di "ripulitura" e non di vera e propria revisione, le modifiche da tenere sotto controllo al fine del lavoro di codifica automatica saranno abbastanza gravose; riguarderanno infatti, in maniera relativamente contenuta, settori nuovi o rivisti; saranno invece più gravose le modifiche relative alle descrizioni.

Tabella 30 - Codici ATECO '91 senza corrispondenza con la NACE Rev.1 emendata

relative alle descrizioni.

Tabella 30 - Codici ATECO '91 senza corrispondenza con la NACE Rev.1 emendata

05.03.0	Attività dei servizi connessi alla pesca e alla piscicoltura
11.11.0	Estrazione di petrolio greggio
11.12.0	Estrazione di gas naturale
11.13.0	Estrazione di sabbie e scisti bituminosi
15.99.0	Fabbricazione di altre bevande analcoliche
17.73.0	Fabbricazione di altra maglieria esterna
17.74.0	Fabbricazione di maglieria intima
17.75.0	Fabbricazione di altri articoli ed accessori a maglia
28.75.5	Fabbricazione di elementi assemblati per ferrovie o tramvie
29.56.1	Fabbricazione e installazione di macchine e apparecchi per le industrie chimiche petrolchimiche e petrolifere (compresi parti e accessori, manutenzione e riparazione)
29.56.2	Fabbricazione e installazione di macchine automatiche per la dosatura, la confezione e per l'imballaggio (compresi parti e accessori, manutenzione e riparazione)
29.56.4	Fabbricazione e installazione di macchine per la lavorazione del legno e materie similari (compresi parti e accessori, manutenzione e riparazione)
51.38.1	Commercio all'ingrosso non specializzato di prodotti surgelati
51.38.2	Commercio all'ingrosso non specializzato di prodotti alimentari, bevande e tabacco
51.39.1	Commercio all'ingrosso di prodotti della pesca freschi
51.39.2	Commercio all'ingrosso di prodotti della pesca surgelati
51.39.3	Commercio all'ingrosso di conserve alimentari e prodotti affini
51.39.4	Commercio all'ingrosso di altri prodotti alimentari
51.54.3	Commercio all'ingrosso di coltelleria e posateria
51.42.4	Commercio al dettaglio di merceria, cucirini, filati, ricami
55.40.2	Gelaterie
60.24.0	Altri trasporti terrestri di passeggeri
60.25.0	Trasporto di merci su strada
61.11.0	Trasporti marittimi
61.12.0	Trasporti costieri

6.2 Impatto sull'applicazione di codifica automatica

Nelle tabelle di seguito riportate si vuole soltanto dare un'idea quantitativa sul numero di testi presenti nel dizionario informatizzato corrispondenti ai codici soggetti alle variazioni citate nel paragrafo precedente, in quanto tali testi dovranno essere esaminati ad uno ad uno e ricollocati nelle classi di competenza.

Tabella 31 – Principali cambiamenti NACE Rev.1.1

Da elementi attualmente previsti da ATECO 91	**Dizioni ufficiali**	**Altri testi del dizionario.**	**Classi previste dalla nuova classificazione**
29.40 – Fabbricazione di macchine utensili	32	86	3 classi – portatili, per la metallurgia e altre.
40.10 – Produzione e distribuzione di energia elettrica	19	31	3 classi – produzione, per il trasporto e la distribuzione ed il commercio
40.20 – Produzione di gas, distribuzione di combustibili gassosi mediante condotte	15	27	2 classi – produzione, e per la distribuzione ed il commercio.
51.64 – Commercio all'ingrosso di macchine e di attrezzature per ufficio	9	71	2 nuove classi: 51.84 e 51.85
51.65 – Commercio all'ingrosso di altre macchine per l'industria, il commercio e la navigazione	15	173	2 nuove classi:51.86 e 51.87
			Nuova classe: 74.86 per le attività dei call center.

la navigazione			
			Nuova classe: 74.86 per le attività dei call center.
			Nuova classe: 72.21 per l'edizione di software.
90.00 – Smaltimento dei rifiuti solidi, delle acque di scarico e simili	34	195	3 classi – raccolta e depurazione delle acque di scarico, raccolta e smaltimento di rifiuti solidi e pulizia delle aree pubbliche, decontaminazione e disinquinamento dell'ambiente
Totale testi da esaminare	124	583	

Tabella 32 -- Codici ATECO '91 senza corrispondenza con la NACE Rev.1 emendata (dimensioni ultima release 09/2001)

Codice	**Descrizione**	**Dizioni ufficiali**	**Altri testi del dizionario.**
05.03.0	Attività dei servizi connessi alla pesca e alla piscicoltura	4	6
11.11.0	Estrazione di petrolio greggio	3	6
11.12.0	Estrazione di gas naturale	11	13
11.13.0	Estrazione di sabbie e scisti bituminosi	5	6
15.99.0	Fabbricazione di altre bevande analcoliche	5	6
17.73.0	Fabbricazione di altra maglieria esterna	3	26
17.74.0	Fabbricazione di maglieria intima	3	12
17.75.0	Fabbricazione di altri articoli ed accessori a maglia	2	7
28.75.5	Fabbricazione di elementi assemblati per ferrovie o tramvie	5	6
29.56.1	Fabbricazione e installazione di macchine e apparecchi per le industrie chimiche petrolchimiche e petrolifere (compresi parti e accessori, manutenzione e riparazione)	19	32
29.56.2	Fabbricazione e installazione di macchine automatiche per la dosatura, la confezione e per l'imballaggio (compresi parti e accessori, manutenzione e riparazione)	25	37
29.56.4	Fabbricazione e installazione di macchine per la lavorazione del legno e materie similari (compresi parti e accessori, manutenzione e riparazione)	14	25
51.38.1	Commercio all'ingrosso non specializzato di prodotti surgelati	1	8
51.38.2	Commercio all'ingrosso non specializzato di prodotti alimentari, bevande e tabacco	4	17
51.39.1	Commercio all'ingrosso di prodotti della pesca freschi	1	8
51.39.2	Commercio all'ingrosso di prodotti della pesca surgelati	4	10
51.39.3	Commercio all'ingrosso di conserve alimentari e prodotti affini	1	5
51.39.4	Commercio all'ingrosso di altri prodotti alimentari	1	29
51.54.3	Commercio all'ingrosso di coltelleria e posateria	3	4
51.42.4	Commercio al dettaglio di merceria, cucirini, filati, ricami	22	11
55.40.2	Gelaterie	1	5
60.24.0	Altri trasporti terrestri di passeggeri	2	10
60.25.0	Trasporto di merci su strada	9	83
61.11.0	Trasporti marittimi	3	5
61.12.0	Trasporti costieri	5	10
	Totale testi da esaminare	***156***	***387***
TOTALE TESTI DEL DIZIONARIO		**7711**	**24934**

Tuttavia il lavoro non potrà ridursi all'esame di questi testi, in quanto dovrà essere verificata la coerenza di tutto l'ambiente di codifica, sia rispetto alla gestione del parsing che alle logiche di codifica adottate.

Tuttavia il lavoro non potrà ridursi all'esame di questi testi, in quanto dovrà essere verificata la coerenza di tutto l'ambiente di codifica, sia rispetto alla gestione del parsing che alle logiche di codifica adottate.
Inoltre, il riesame della quinta cifra ATECO, che costituisce il dettaglio nazionale, comporterà l'impatto più pesante sull'applicazione di codifica. Si teme infatti che tali variazioni potranno soltanto parzialmente essere automatizzate (nei casi in cui il contenuto informativo relativo ad un codice a cinque cifre transiti completamente su un altro codice), mentre ogni qual volta siano state modificate le dizioni associate al codice a cinque digit, dovranno essere singolarmente analizzati tutti i testi corrispondenti a quel codice e presenti nel dizionario, per verificare quali di questi debbano permanere nella stessa categoria e quali transitare in altre.

Relativamente ai prossimi censimenti, come è noto, i dati saranno codificati sulla base della vecchia classificazione, ma dovranno successivamente essere riportati a quella nuova.
Questa operazione viene tradizionalmente realizzata tramite una transcodifica, quasi sempre effettuata a livello di aggregati.

L'aggiornamento dell'ambiente di codifica rispetto alla classificazione revisionata, pur essendo un lavoro indubbiamente gravoso, comporterebbe senz'altro una innovazione nel processo di transcodifica che potrebbe essere effettuato direttamente sui dati elementari.

Si dovrebbe infatti ritornare alle dizioni fornite dai rispondenti e sottoporle nuovamente al sistema di codifica automatica, aggiornato secondo la nuova classificazione; soltanto la percentuale di casi non risolti automaticamente dovrebbe essere analizzata manualmente, oppure transcodificata tramite le tradizionali procedure che lavorano sugli aggregati.

E' superfluo infine osservare che l'aggiornamento dell'applicazione di codifica la renderebbe disponibile da ora in poi per qualsiasi altra indagine dell'Istituto che rilevi l'attività economica tramite quesito a testo libero.

7 Variabile 'Natura giuridica delle imprese'

Relativamente a questa variabile, non erano state effettuate sperimentazioni antecedenti alle attività del gruppo di lavoro, tuttavia la sua minore complessità ha consentito il raggiungimento di buoni risultati in un periodo abbastanza breve.

L'ambiente applicativo è stato predisposto tramite la rielalaborazione della classificazione ufficiale disponibile all'epoca della Long-Form ed arricchito a seguito dell'analisi dei risultati delle applicazioni di codifica automatica sulle prime tranche di dati dell'indagine stessa.

In dettaglio, è stato raggiunto un tasso di codifica del 94% ed un livello di correttezza assoluto, con un dizionario di 105 testi, a fronte di una classificazione ufficiale che prevede 28 modalità.

Tabella 33 -- Risultati dell'applicazione di codifica automatica

Testi del dizionario	Sinonimi	Efficacia del sistema
N.	**N.**	**%** di testi codificati automaticamente
105	255	94

Bibliografia

Appel M. and Hellerman E. (1983). "Census Bureau experience with Automated Industry and Occupation Coding". In American Statistical Association, *Proceedings of Section on Survey Research Methods*, pages 32-40.

Chen B., Creecy R. and Appel M. (1993). "Error control of automated industry and occupation coding", *Journal of Official Statistics*, vol. 9: 729-745.

Cochran W. G. (1977). *Sampling Techniques, 3rd ed.*. Wiley, New York.

De Angelis R., Macchia S. and Mazza L. (2000), "Applicazioni sperimentali della codifica automatica: analisi di qualità e confronto con la codifica manuale", sta in (a cura di Istat) Quaderni di ricerca – Rivista di statistica Ufficiale, n. 1, 29-54

Istat, (1991), "Classificazione delle attività economiche", Metodi e norme. Serie C – n.11

Lyberg L. and Dean P. (1992). "Automated Coding of Survey Responses: an international review." In *Conference of European Statisticians, Work session on Statistical Data Editing*, Washington DC

Istat, (1991), "Classificazione delle attività economiche", Metodi e norme. Serie C – n.11

Lyberg L. and Dean P. (1992). "Automated Coding of Survey Responses: an international review." In *Conference of European Statisticians, Work session on Statistical Data Editing*, Washington DC.

Kalpic D. (1994). "Automated coding of census data", *Journal of Official Statistics*, vol. 10: 449-463.

Knaus R. (1987). "Methods and problems in coding natural language survey data", *Journal of Official Statistics*, vol. 1, 45-67.

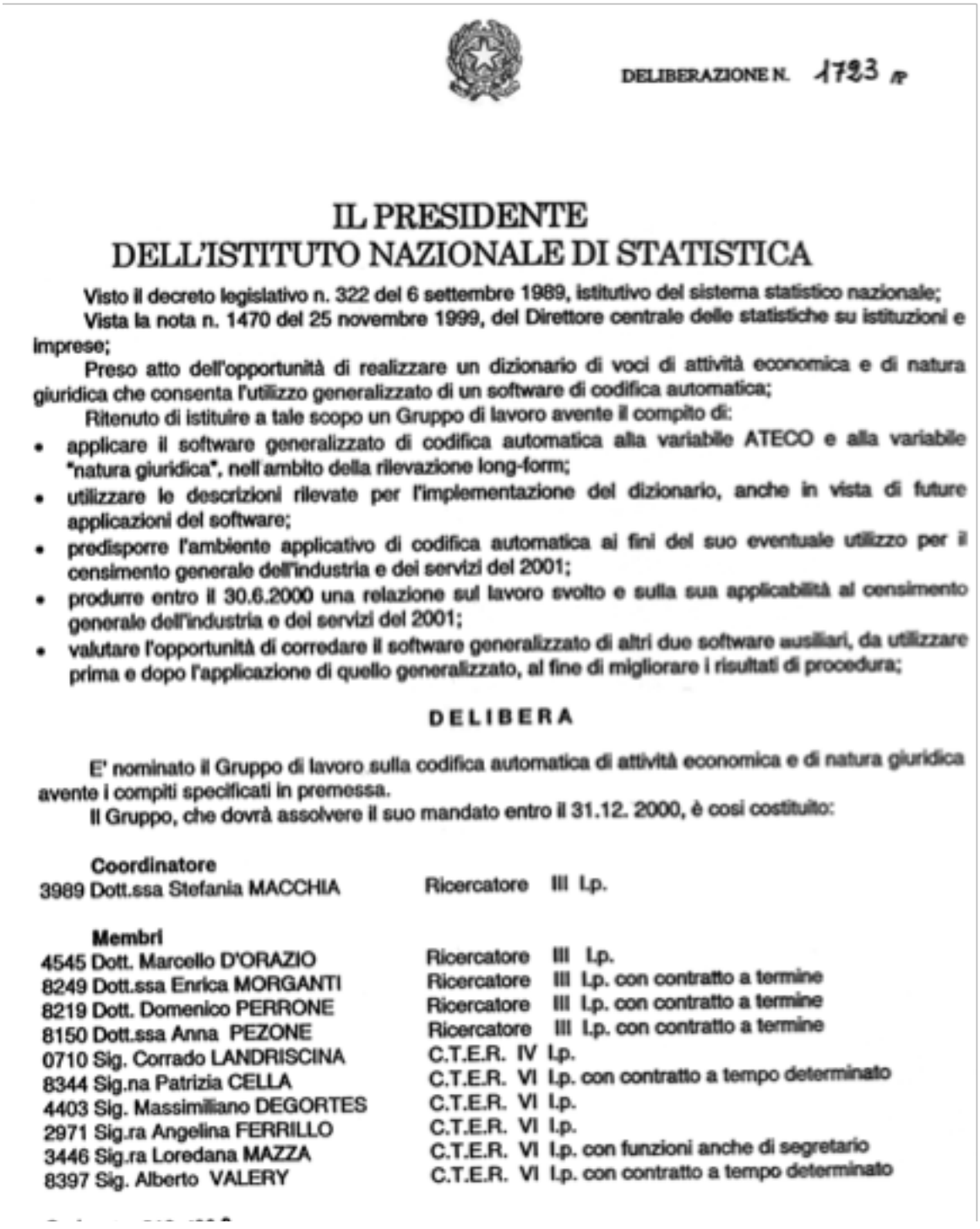
Macchia S. (2001). " Integration of sources to build a dictionary for Automated Coding of Industry." In CLADAG Conference, Palermo, 5-6 luglio 2001

Macchia S. and D'Orazio (2000), "Analysis of Textual data for integrating an automated coding environment system and building a system to monitor the quality of its results", *5th Journées Internationales d'Analyse Statistique des Données Textuelles*, Lausanne, Switzerland, 9-11 Mars 2000, 407-414

Massingham R. (1997). "Data capture and Coding for the 2001 Great Britain Census". In *XIV Annual International Symposium on Methodology Issues*, 5-7 November, Hull, Canada.

Tourigny J.Y. and Moloney J. (1995). The 1991 Canadian Census of Population experience with automated coding. In United Nations Statistical Commission, Statistical Data Editing, 2.

Wenzowski M.J. (1988). ACTR – A Generalised Automated Coding System. Survey Methodology, vol. 14: 299-308.



DELIBERAZIONE N. 1723

IL PRESIDENTE DELL'ISTITUTO NAZIONALE DI STATISTICA

Visto il decreto legislativo n. 322 del 6 settembre 1989, istitutivo del sistema statistico nazionale;

Vista la nota n. 1470 del 25 novembre 1999, del Direttore centrale delle statistiche su istituzioni e imprese;

Preso atto dell'opportunità di realizzare un dizionario di voci di attività economica e di natura giuridica che consenta l'utilizzo generalizzato di un software di codifica automatica;

Ritenuto di istituire a tale scopo un Gruppo di lavoro avente il compito di:

- applicare il software generalizzato di codifica automatica alla variabile ATECO e alla variabile "natura giuridica", nell'ambito della rilevazione long-form;
- utilizzare le descrizioni rilevate per l'implementazione del dizionario, anche in vista di future applicazioni del software;
- predisporre l'ambiente applicativo di codifica automatica ai fini del suo eventuale utilizzo per il censimento generale dell'industria e dei servizi del 2001;
- produrre entro il 30.6.2000 una relazione sul lavoro svolto e sulla sua applicabilità al censimento generale dell'industria e dei servizi del 2001;
- valutare l'opportunità di corredare il software generalizzato di altri due software ausiliari, da utilizzare prima e dopo l'applicazione di quello generalizzato, al fine di migliorare i risultati di procedura;

DELIBERA

E' nominato il Gruppo di lavoro sulla codifica automatica di attività economica e di natura giuridica avente i compiti specificati in premessa.

Il Gruppo, che dovrà assolvere il suo mandato entro il 31.12. 2000, è così costituito:

Coordinatore

3989 Dott.ssa Stefania MACCHIA	Ricercatore III l.p.

Membri

4545 Dott. Marcello D'ORAZIO	Ricercatore III l.p.
8249 Dott.ssa Enrica MORGANTI	Ricercatore III l.p. con contratto a termine
8219 Dott. Domenico PERRONE	Ricercatore III l.p. con contratto a termine
8150 Dott.ssa Anna PEZONE	Ricercatore III l.p. con contratto a termine
0710 Sig. Corrado LANDRISCINA	C.T.E.R. IV l.p.
8344 Sig.na Patrizia CELLA	C.T.E.R. VI l.p. con contratto a tempo determinato
4403 Sig. Massimiliano DEGORTES	C.T.E.R. VI l.p.
2971 Sig.ra Angelina FERRILLO	C.T.E.R. VI l.p.
3446 Sig.ra Loredana MAZZA	C.T.E.R. VI l.p. con funzioni anche di segretario
8397 Sig. Alberto VALERY	C.T.E.R. VI l.p. con contratto a tempo determinato

2971 Sig.ra Angelina FERRILLO — C.T.E.R. VI l.p.
3446 Sig.ra Loredana MAZZA — C.T.E.R. VI l.p. con funzioni anche di segretario
8397 Sig. Alberto VALERY — C.T.E.R. VI l.p. con contratto a tempo determinato

Sede, 06 DIC. 1999

IL PRESIDENTE

codifica.doc

DELIBERAZIONE N. 205 /P

IL PRESIDENTE DELL'ISTITUTO NAZIONALE DI STATISTICA

Visto il decreto legislativo n. 322 del 6 settembre 1989, istitutivo del sistema statistico nazionale;

Vista la deliberazione n. 1723/P del 6 dicembre 1999, con la quale è stato costituito il Gruppo di lavoro sulla codifica automatica di attività economica e di natura giuridica;

Vista la nota n. 24 del 15.2.2000 del coordinatore il Gruppo di cui sopra;

Ritenuto di inserire come membri, a supporto delle attività previste, relative all'arricchimento del dizionario elaborabile della classificazione delle attività economiche ed alla sua integrazione con la classificazione dei prodotti, il dott. Alberto SOCINI collaboratore tecnico enti di ricerca IV l.p. – ed il sig. Piero BRETTI – collaboratore tecnico enti di ricerca VI l.p., entrambi della Direzione centrale delle statistiche su istituzioni e imprese;

D E L I B E R A

Il dott. Alberto SOCINI – matr. 2909 collaboratore tecnico enti di ricerca di quarto livello professionale ed il sig. Piero BRETTI – matr. 2672 collaboratore tecnico enti di ricerca di sesto livello professionale – sono chiamati a far parte, in qualità di membri, del Gruppo di lavoro sulla codifica automatica di attività economica e di natura giuridica.

Sede, 06 MAR. 2000

IL PRESIDENTE

Codifica2.doc

ALLEGATO 1

NOME PROGRAMMA	GLA.EXE	DATA:	2001	
LINGUAGGIO	VBASIC+PROMPT DOS(ACTR)			

DESCRIZIONE	
PERMETTE DI INSERIRE IN ALCUNI RECORDS DI UN FILE DI INPUT PER ACTR(ATECO_T1_T8.TXT) UN DETERMINATO FILTRO (VEDI ALLEGATO) ED UN FILTRO FITTIZIO("00") PER LE ECCEZIONI (CHE NELL'ULTIMO PASSAGGIO DELLA PRIMA FASE SARÀ PORTATO A " ") AL FINE DI CONTROLLARE ED AVERE LA CERTEZZA CHE GLI UNICI CODIFICATI, DA ACTR, SIANO CODIFICATI CON IL GIUSTO CODICE	
SPECIFICA	
CREA DB FILTRO RIDOTTO CREARE IL CONTESTO(RIDOTTO) DEI SOLI FILTRI ESEGUIRE IL FILE CREADB_FILTRO.BAT BCODE FILTRO SFRUTTA LA CODIFICA ACTR PER ASSEGNARE IL CODICE, CHE IN REALTÀ SARÀ IL NOSTRO FILTRO. ESEGUE BCODE_FILTRO.BAT ASSEGNA FILTRO TRASFORMA IL FILE DI INPUT ORIGINALE IN UN FILE CON IL FILTRO O CON " " NEI PRIMI DUE BYTES. SE IL FILTRO È UGUALE A "00"" VIENE INSERITO " ".	
SINTASSI	ANNOTAZIONI
CREARE UN'ICONA SUL DESKTOP DI WINDOWS E FARE DOPPIO CLICK CON IL MOUSE	TRACCIATO CHE NON PUÒ ESSERE VARIATO NOME DEL FILE DI INPUT FISSO PER ESSERE ESEGUITO HA BISOGNO DELLE LIBRERIE DI RUNTIME DI VISUAL BASIC 6.

ALLEGATO 2

Filtro	Descrizione
00	- acconciatura
00	- acido borico

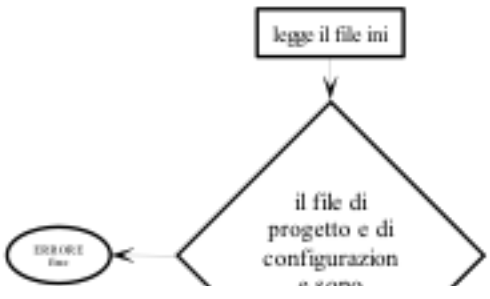
Filtro	Descrizione
00	- commercio biglietteria
00	- commercio camere di commercio

00	- acconciatura	00	- commercio biglietteria
00	- acido borico	00	- commercio camere di commercio
00	- affresco	00	- commercio cartotecnica
00	- pittore	00	- commercio cava
00	- ritrattista	00	- commercio cesellatore
00	- produzione uova	00	- commercio cessione crediti
00	- agente	00	- commercio confezionamento
00	- agenzia	00	- commercio consulenza
00	- trasporto	00	- commercio conto vendite
00	- corso	00	- commercio costruzione
00	- corr.	00	- costruzione produzione
00	- corrispondenza	00	- commercio crediti
00	- agenzia commercio gestione immobili	00	- commercio distribuzione
00	- agenzia commercio immobili	00	- commercio economia
00	- agenzia commercio gestione immobiliare	00	- commercio edizioni musicali
00	- agenzia commercio immobiliare	00	- commercio escavazione
00	- agenzia commercio permuta	00	- commercio estetista
00	- agenzia immobili	00	- commercio formulari
00	- agenzia locazione	00	- commercio immobili commercio
00	- allestimento	00	- commercio incisioni
00	- area	00	- commercio industrie
00	- argilla	00	- commercio informazioni
00	- arredamenti	00	- commercio laboratorio fotografico
00	- arrotino	00	- commercio lavanderia
00	- asfalto	00	- commercio lavori
00	- associazione professionale	00	- commercio macchine automatiche
00	- attrezzature trsf.	00	- commercio manutenzione
00	- audio	00	- commercio ministero commercio
00	- autolavaggio	00	- commercio montaggio
00	- ballo	00	- commercio motori elettrici
00	- bar alimentari	00	- commercio noleggio
00	- alimento	00	- commercio obbligazioni
00	- baritina	00	- commercio opuscoli commerciali
00	- bitume	00	- commercio pellicole
00	- borato	00	- commercio periti commerciali
00	- calcestruzzo	00	- commercio pubblica amministrazione
00	- caolino	00	- commercio pubblicitari
00	- carbonato	00	- commercio raccolta
00	- cereali	00	- commercio radiodiffusione
00	- commercio abbattimento	00	- commercio recupero
00	- commercio acconciatore	00	- commercio restauro
00	- commercio affari generali	00	- commercio ricerche
00	- commercio affitto	00	- commercio rigenerazione
00	- commercio allevamento	00	- commercio rimessaggio
00	- commercio amministrativo	00	- commercio riparazionegomme
00	- commercio assistenza alla vendita	00	- commercio riparazionepneumatici
00	- commercio attività amministrative	00	- commercio servizi
00	- commercio autofficina	00	- commercio sistemi elettronici

00	- commercio assistenza alla vendita	00	- commercio riparazionepneumatici
00	- commercio attività amministrative	00	- commercio servizi
00	- commercio autofficina	00	- commercio sistemi elettronici
00	- commercio azioni	00	- commercio sistemi telematici
00	- commercio banche	00	- commercio software
00	- commercio bar	00	- commercio studio commerciale
00	- commercio biglietti ferroviari	00	- commercio studio legale commerciale
00	- commercio biglietti	00	- commercio tecnico
00	- commercio titoli	00	- olio d'oliva
00	- commercio tombe	00	- pallets
00	- commercio transazioni	00	- pietra calcarea
00	- commercio vendita assistenza tecnica	00	- pirite
00	- commercio viaggi	00	- pomice
00	- compravendita	00	- porchetta
00	- conglomerati	00	- pozzolana
00	- consulente assicurazione	00	- pratiche auto assicurazioni
00	- cuscini campeggio	00	- precucinati
00	- decorazione	00	- produzione agrumi
00	- demolizione edifici	00	- partecipazioni
00	- commercio urbanizzazione aree	00	- holding
00	- urbanizzazione aree	00	- produzione audiovisiva
00	- demolizione edilizia	00	- produzione barbabietola zucchero
00	- demolizione escavazione	00	- produzione cereali
00	- demolizione navi	00	- produzione colture
00	- demolizione noleggio	00	- produzione cotone
00	- demolizione strutture	00	- produzione fiori
00	- dottore commercio	00	- produzione florovivaistica
00	- eliocopie	00	- produzione frutta
00	- estrazione lavorazione	00	- produzione ittica
00	- estrazione produzione	00	- produzione latte
00	- estrazione scavi sterri	00	- produzione legumi
00	- fermenti	00	- produzione miele
00	- fluorite	00	- produzione mitili
00	- fostati	00	- produzione olive
00	- fotoceramiche	00	- produzione ortaggi
00	- fotografo	00	- produzione ortofloreale
00	- gestione commercio	00	- produzione ortovivaistica
00	- ghiaia	00	- produzione patate
00	- humus	00	- produzione piante ornamentali
00	- inerti	00	- produzione pizza
00	- ingrosso giornali	00	- produzione procedure informatiche
00	- insegnante	00	- produzione semi e frutti oleosi
00	- investimenti	00	- produzione sistemi informativi
00	- inscatolamento	00	- produzione software
00	- laboratorio	00	- produzione tabacchi
00	- lavorazione cemento	00	- produzione televisiva
00	- lavorazione macchine	00	- produzione uva
00	- lavorazione pellicceria	00	- produzione video
00	- lavorazione utensili	00	- produzione videogiochi
00	- legnaardere	00	- produzione vite
00	- lucidatura	00	- produzioni cinematografiche
00	- magazzino	00	- produzioni film

00	- lavorazione utensili	00	- produzione videogiochi
00	- legnaardere	00	- produzione vite
00	- lucidatura	00	- produzioni cinematografiche
00	- magazzino	00	- produzioni film
00	- materassaio	00	- produzioni teatrali
00	- mediazione	00	- noleggio auto
00	- mobile	00	- pubblicità
00	- modello	00	- pubblico
00	- olio	00	- pubbl.
00	- riciclaggio produzione	20	- commercio vinificazione
00	- rifiuti	20	- essiccazione
00	- rigenerazione	20	- falegnameria
00	- riparazione commerciopneumatico	20	- finissaggio
00	- sabbia	20	- fonderia
00	- salamoia	20	- frantoio commercio
00	- sale	20	- macellazione
00	- scuola taglio	20	- macinazione
00	- servizio mensa	20	- mobiliere
00	- sociale abbligatoria	20	- produzione derivati
00	- solfati	20	- produzione olio
00	- stagionatura formaggi	20	- produzione olio oliva
00	- stampati commercio	20	- produzione alimento
00	- taglio alberi	20	- raffinazione
00	- taglio boschi	20	- restauratore ambulante
00	- taglio pietra	20	- salagione
00	- ufficio commerciale	20	- sbavatura
00	- vivaio commercio	20	- segheria commercio
00	- vulcanizzazione commercio	20	- stagionatura
00	- zolfo	20	- taglio
00	- commercio segagione	20	- tappezziere
00	- sede	20	- torrefazione
00	- controllo pasti	20	- trafilatura
00	- pulizia mensa	20	- vinificazione
00	- panificio	20	- produzione auto
00	- panetteria	20	- produzione motociclo
00	- pasticceria	40	- agente commercio
00	- ricambio	40	- agenzia commercio
00	- forno	40	- intermediario commercio
00	- artigiano	50	- ambulante
00	- artigiano edile	50	- commencio dettaglio ambulante
00	- istituto	50	- commercio a domicilio
00	- giudiziario	50	- commercio dettaglio a domicilio
00	- pneumatici	50	- commercio ambulante
00	- confezione	50	- commercio aree pubbliche
00	- promozione	50	- commercio dettaglio aree pubbliche
00	- ragioniere	50	- commercio corrispondenza
00	- pane	50	- commercio dettaglio corr.
10	- estrazione	50	- commercio dettaglio corrispondenza
20	- commercio essiccazione	50	- commercio dimostratore
20	- commercio falegnameria	50	- commercio diretto
20	- commercio fonderia	50	- commercio fisso
20	- commercio incisoria	50	- commercio dettaglio diretto
20	- commercio macinazione	50	- commercio dettaglio fisso
20	- commercio mobiliere	50	- commercio itinerante

20 - commercio fonderia
20 - commercio incisoria
20 - commercio macinazione
20 - commercio mobiliere
20 - commercio produzione
20 - commercio raffinazione
20 - commercio riproduzione
20 - commercio salagione
20 - commercio stagionatura
20 - commercio timbrificio
20 - commercio torrefazione
20 - commercio trasformazione
60 - commercio dettaglio
60 - commercio dettaglio alimento
60 - commercio ingrosso
60 - commercio ingrosso alimento
60 - commercio macchine
60 - commerciogomme
60 - commerciogomma
60 - commerciopneumatico
60 - ingrossogomme
60 - ingrossopneumatico
70 - affittacamere
70 - autogrill
70 - bar
70 - bottiglieria
70 - catering
70 - ferie
70 - mensa
70 - ostello
70 - osteria
70 - paninoteca
70 - pasti
70 - pizza a taglio
70 - pizzeria
70 - produzione rosticceria
70 - rifugio
70 - ristorante
70 - rosticceria
70 - somministrazione commercio
70 - tavolacalda
70 - trattoria
70 - vacanze
70 - villaggio
80 - agente assicurazione
80 - agenzia assicurazione
80 - intermediario assicurazione
50 - commercio fisso
50 - commercio dettaglio diretto
50 - commercio dettaglio fisso
50 - commercio itinerante
50 - commercio posteggio
50 - commercio posteggio fisso
50 - commercio dettaglio posteggio
60 - intermediario commercio auto
60 - intermediario auto
60 - commercio auto
60 - intermediario commercio motociclo
60 - COMMERCIO



ALLEGATO 3

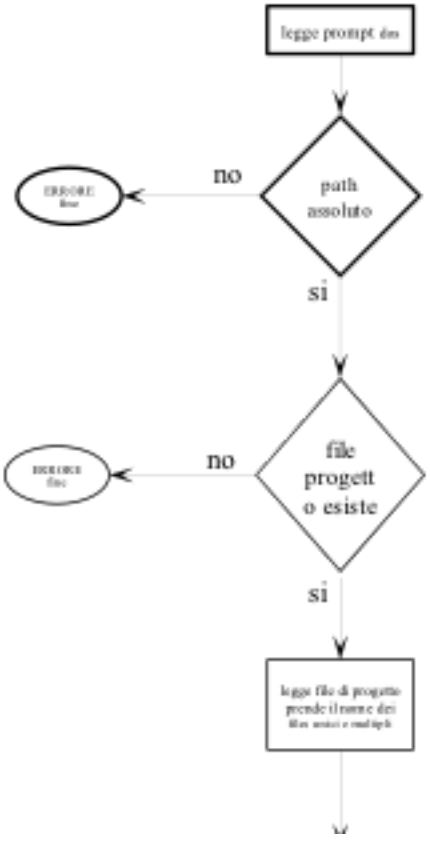
NOME PROGRAMMA	MULTIPLI2UNICI.EXE	DATA:	2001	
LINGUAGGIO	VBASIC			

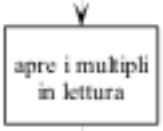
DESCRIZIONE
PROGRAMMA CHE, DOPO UNA ELABORAZIONE DI ACTR, ANALIZZA IL FILE DEI MULTIPLI ED ESTRAE ALCUNI RECORDS E LI ACCODA AL FILE DEGLI UNICI METTENDO IL FLAG SPECIFICO

SPECIFICA	FLUSSO
	CONTROLLA ESISTENZA DEL FILE DI PROGETTO ACTR LEGGE I NOMI DEI FILE UNICI E MULTIPLI

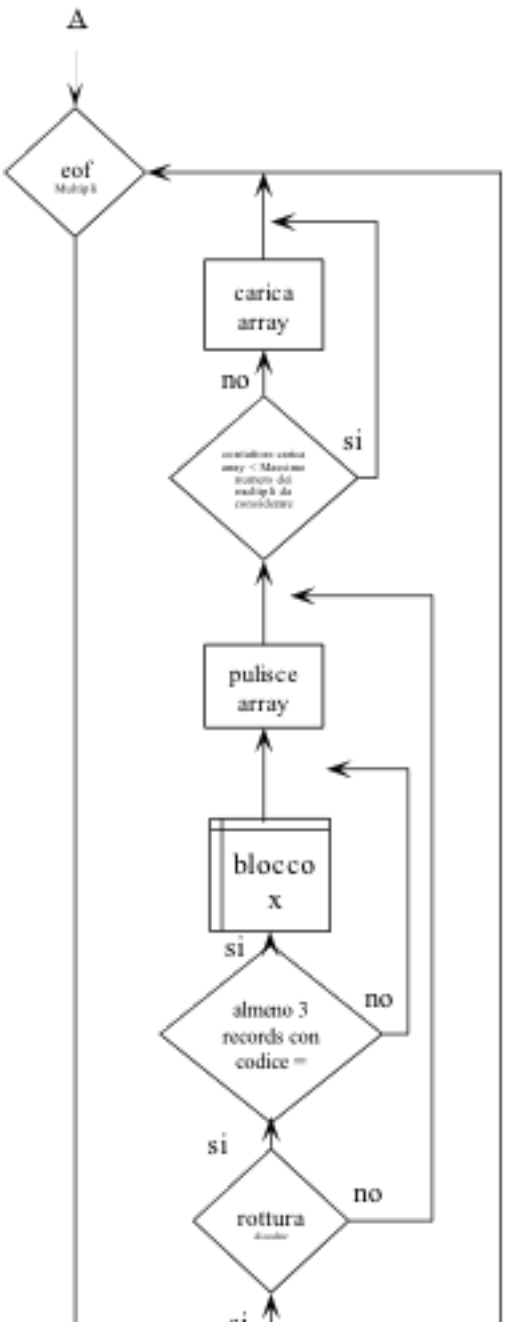
SPECIFICA	PROCESSO
SELEZIONA I RECORDS DA INSERIRE IN BASE A QUESTI CRITERI: 5 RECORDS CON IL CODICE UGUALE A 5 DIGIT calcola il punteggio maggiore e prende il record. 5 RECORDS CON IL CODICE UGUALE A 4 DIGIT. PRENDE IL CODICE, MA LA DESCRIZIONE È BLANK PERCHÉ I CODICI POSSONO ESSERE COMPLETI. 3 RECORDS TOTALI CON IL CODICE UGUALE A 5 DIGIT. riporta il testo della stringa del dizionario a cui corrisponde il punteggio maggiore. 4 RECORDS TOTALI CON IL CODICE UGUALE A 5 DIGIT riporta il testo della stringa del dizionario a cui corrisponde il punteggio maggiore.	CONTROLLA ESISTENZA DEL FILE DI PROGETTO ACTR LEGGE I NOMI DEI FILE UNICI E MULTIPLI LEGGE I MULTIPLI ORDINA IN BASE AL CODICE ACTR FLAG CHE VENGONO UTILIZZATI: '55' PROVENIENZA DA 5 RECORDS UGUALI A 5 DIGIT '54' PROVENIENZA DA 5 RECORDS UGUALI A 4 DIGIT '35' PROVENIENZA DA 3 RECORDS TOTALI UGUALI A 5 DIGIT '45' PROVENIENZA DA 4 RECORDS TOTALI UGUALI A 5 DIGIT VISUALIZZA IL RISULTATO DELL'ELABORAZIONE CON IL NUMERO DEI RECORDS INSERITI E LA PERCENTUALE RISPETTO AI MULTIPLI COME SINGOLI RECORDS

SINTASSI	ANNOTAZIONI
DAL PROMPT DEL DOS: C:> MULTIPLI2UNICI NOMEPROGETTOACTR	PER POTER ESSERE ESEGUITO HA BISOGNO DELLE LIBRERIE DI RUNTIME DEL VISUL BASIC 6.

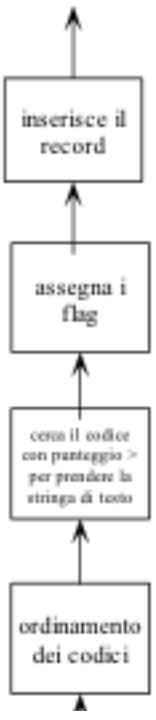


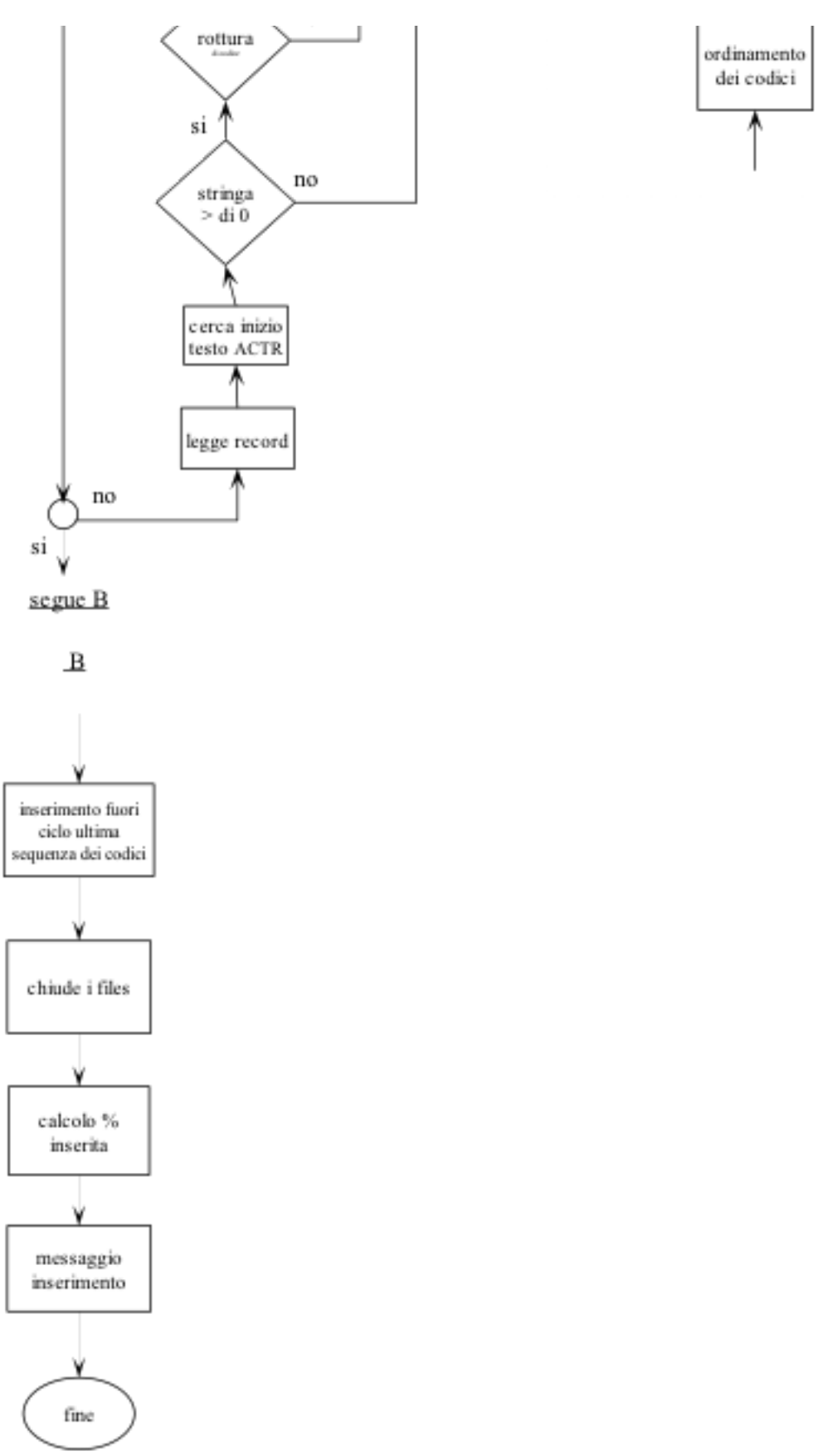


segue A



BLOCCO X




rottura
si
no
stringa
> di 0
cerca inizio
testo ACTR
legge record
no
si
segue B
B
ordinamento
dei codici
inserimento fuori
ciclo ultima
sequenza dei codici
chiude i files
calcolo %
inserita
messaggio
inserimento
fine

ALLEGATO 4

INCOERENZE ATECO-PRODCOM

Nel confronto Ateco-Prodcom sono state riscontrate delle incongruenze, alcune delle quali già segnalate a suo tempo, ma rimaste senza soluzione.

Il caso della **17.1**, relativo al settore tessile ***('Preparazione e filatura di fibre tessili')***, risulta eclatante. Infatti, secondo l'ATECO, si attribuisce l'attività economica dell'impresa in base alle sottoclassi:

- 1711 à fibre tipo cotone
- 1712 à fibre tipo lana cardata
- 1713 à fibre tipo lana pettinata
- 1714 à fibre tipo lino
- 1715 à fibre tipo seta e sintetici o artificiali
- 1716 à filati tipo cucirini
- 1717 à altre fibre tessili

Per PRODCOM, invece, (forse più realisticamente) non si prevede questa suddivisione in quanto le imprese non hanno più la netta distinzione o prevalenza a seconda del tipo di fibra trattata, come vorrebbe individuare l'ATECO; in base a tale logica, la PRODCOM prevede una più generica classe 1710.

Come ben si intuisce questa diversità interpretativa implica una serie di problemi da parte dell'archivio che si riflettono poi nelle altre fasi, come i piani di compatibilità e definizioni di universo statistico.

Se dovesse essere confermato il criterio ATECO, occorrerebbe allora segnalare quali prodotti PRODCOM debbono confluire nelle varie sottoclassi ATECO e, secondo la prevalenza, attribuire alle imprese una ATECO che potrebbe comunque essere aggiornata con una certa frequenza, data la variabilità del fenomeno produttivo.

Le altre incoerenze finora riscontrate sono riportate nella seguente Tabella:

ATECO		**PRODCOM**	
Produzione di estratti e sughi di carne *Produzione di prodotti a base di pesce*	15130 15202	15130 1270	*Estratti di sughi di carne, peschi e invertebrati acquatici*
Fabbricazione di prodotti per l'alimentazione degli animali da allevamento	1571	15330 3000	*Materie vegetali e cascami vegetali per l'alimentazione degli animali*
Produzione di paste di cioccolato da spalmare	15893	15840 1100	*Pasta da spalmare contenente cacao*
Produzione di fecola di patate	1531	15620 2215	*Fecola di patate*
Produzione di oli vegetali grezzi… granturco	15412	15621 1030	*Olio greggio di granturco e sue frazioni*
Produzione di aceto	15893	15870 1130 15870 1190	*Aceto di vino e suoi succedanei* *Aceto di vino e suoi succedanei*
Produzione di idromele	15950	15940 1000	*Altre bevande fermentate, es. idromele*
Fabbricazione di altre bevande analcoliche:… nettari	1599	15980 1255	
ATECO non prevede questa voce		16000 1170	*Sigari, sigaretti, sigarette, non contenenti*

Fabbricazione di altre bevande analcoliche:… nettari	1599	15980 1255	
ATECO non prevede questa voce		16000 1170	*Sigari sigaretti sigarette non contenenti tabacco*
Confezione su misura di vestiario	18222		*PRODCOM non prevede la distinzione 'su misura'*
Produzione di lame di seghe	28621	28622 2020	*Lame di seghe*
Produzione di cambi	29141	34300 2033	*Cambi*
Produzione installazione di fornaci e bruciatori	29211	29212 9100	*Installazione di fornaci e bruciatori*
Produzione di scambiatori di calore	29243	29231 1132 29231 1133 29231 1135 29231 1137	*Scambiatori di calore*
Produzione di laminatoi	29243	29510 1153 29510 1155 29510 1157	*Laminatoi*
Produzione di fibre ottiche e cavi di fibre ottiche	31300	33403 2119	*Altre fibre ottiche, fasci e cavi di fibre ottiche*
Produzione di condensatori	32100	29231 3050	*Evaporatori e condensatori, diversi da quelli per gli apparecchi del tipo domestico*
Produzione di lenti	33403	33404 1130 33404 1153 33404 1155 33404 1159 33404 1170	*Fabbricazione di lenti e strumenti ottici di precisione*
Produzione di mobili di metallo	36121	36141 1101 36141 1102	*Fabbricazione di mobili di metallo*
Produzione di parti di mobili	36141	36142 1530 36142 1550 36142 1599 36142 1595	*Parti di mobili*
Produzione di puzzle	36501	36502 3250 36502 3290	*Puzzle*

Per quanto attiene alle attività del gruppo di lavoro, si auspica un sollecito intervento dei soggetti preposti a queste verifiche per chiarire quale delle due fonti sia da prendere come riferimento nella costruzione del dizionario informatizzato per la codifica dell'ATECO; si ritiene comunque che tale chiarimento sia anche di fondamentale importanza per garantire la coerenza tra il sistema informativo delle imprese e l'indagine PRODCOM.

ALLEGATO 5

Parole considerate nel passaggio di correzione ortografica					
parole errate TESTO LONG-FORM E INPS					
nome1					
affitto	Ffitto	Fitto			
agente	Gente				
altri	Ltri				

affitto	Fitto	Fitto			
agente	Gente				
altri	Ltri				
ambulante	Ambilante				
amministrazione	amm.ne				
assemblaggio	Assembl	Ssemblaggio			
attività	Ttività	Attiv			
commercio	Comm	Ommercio	Comme	commer	commerc
	Comemrcio	Commecio	Commeercio	commefcio	commeriio
	Commercioal	Commercio	Commerco	Commercxio	commericio
	Commertcio	Commwercio	Ommerio	ommer	
	Com	Omm	Coomm		
confezione	Onfezione	Conf	Confez	onf	
costruzione	Costr	Costruz	Costruzi	ostruzione	costurzione
	Ostruz				
creazione	Creaz				
dettaglio	Dettalgio				
estrazione	Estraz				
fabbricazione	Fabb	Fabbr	Fabbric	fabbricaz	abbr
	Abbricazione	Fab	Fabbrc	fabbricazone	fabbriz
	Abbri	Abbric	Fabbri		
gestione	Gest				
imbottigliatore	Imbottilgiatore				
ingrosso	Igrosso	Inngrosso	Ingorosso	ngrosso	
installazione	Inst	Instal	Install	installaz	istallazione
	Istallatore				
laboratorio	Aboratorio				
lavorazione	Lavor	Avor	Lavoraz	lavo	lavorazi
	Avorazione	Lavortazione	Avorazioni		
lavori	Avori				
locazione	Ocazione	Locaz			
manutenzione	Manut	manute	Manutenz		
montaggio	Monteggio				
nettezza	Nertezza				
noleggio	Oleggio				
officina	Fficina				
produzione	Produ	Produz	Produzio	roduz	prod
	Produzone	Roduzione	Produzionew	pro.ne	rod
	Oroduzione				
progettazione	Prog	Progettaz	Progett		
realizzazione	Realizz				
restauro	Estauro				
rifinizione	Ifinizione				
riparazione	Iparazione				
	Rip	Ripar	Riparaz	inparazione	

restauro	Estauro				
rifinizione	Ifinizione				
riparazione	Iparazione				
	Rip	Ripar	Riparaz	iparazione	
servizi	Ervizi	Serv			
tessitura	Tessit	Essitura	Tess	tessirura	
trasformazione	Trasf				
vendita	Vend				
nome2,nome3,nome4,nome5,nome6,nome7,nome8					
ambulante	Abul	Ambilante			
assicurazione	Assicurazoni	Asssicurazioni			
autocaravan	Autocaravam				
biliardi	Gigliardi				
calzature	Calzaure				
certificazione	Cert				
cinematografica	Sinematografica				
commercio	Comm	Ommercio	Comme	commer	commerc
	Coomm				
costruzione	costr	Costruz	Costruzi		
dentarie	dentarier				
dettaglio	dettalgio				
fabbricazione	fabbricaz	Fabbriz			
farmaceutici	farm.ci				
gioielleria	gioelleria				
hardware	harware				
imbarcazione	imbarc				
imbottigliatore	imbottilgiatore				
immobiliare	immobhiliari				
ingrosso	igrosso	Inngrosso	Ingorosso	ngrosso	
installazione	inst	Instal	Install	installaz	
investigazione	inverstcazioni	Investicazioni			
laterizi	latterizi	Terizi			
lavorazione	lavor	Avor	Lavoraz	lavo	lavorazi
macchine	macchone				
manutenzione	manut	manute	Manutenz		
medic.li	medicinali				
metalliche	matalleche				
minerali	mineri				
nettezza	nertezza				
preziosi	prez				
produzione	produ	Produz	Produzio	roduzione	
riparazione	rip	Ripar	Riparaz		
software	softwuere	Stoftware			
specializzato	speci				
termoacustici	termoa				
trafilatura	trafil				
vendita	vend				

ALLEGATO 6

vendita	vend				

ALLEGATO 6

Specifiche tecniche nella definizione del Parsing dell'applicazione Ateco

Premessa di carattere generale

Come già detto nel cap. 2 di questo documento, la fase di *standardizzazione* dei testi, detta '*parsing*' , ha lo scopo di eliminare dal testo la variabilità grammaticale o sintattica che non incide sul *significato semantico* ma soltanto sulla *forma* e che pertanto è irrilevante ai fini dell'abbinamento con le voci del dizionario della classificazione. La fase di standardizzazione è completamente controllata dall'utente che ha il compito di adattarla al particolare contesto applicativo (lingua, classificazione, tipologia di rispondente). La peculiarità di ACTR rispetto ad altri sistemi è che, mettendo a disposizione fino a 14 diverse funzioni di *parsing* (mappatura dei caratteri, eliminazione di parole ininfluenti, definizione di sinonimi per parole singole o per gruppi di parole, rimozione di suffissi, prefissi, etc…), consente una notevole flessibilità e possibilità di personalizzazione del processo di standardizzazione.

Le fasi del processo di *standardizzazione (parsing)* dei testi in ACTR si possono suddividere in cinque parti:

- **Pre-processing** (pre-elaborazione à trattamento dei caratteri)

STRIMM TRIMMING (pulizia dei caratteri)
Consente la standardizzazione del testo tramite l'eliminazione di caratteri estranei, quali spazi multipli, caratteri di tabulazione, ecc.. In particolare:

AUTOTRIM
Consente la cancellazione dei carattreri spuri, quali doppi blank, tabulazioni, andata a capo
TRIMLEF
Consente la cancellazione dei blank a sinistra del simbolo specificato ad es)]}-/\!?,
TRIMRIGHT
Consente la cancellazione dei blank a destra del simbolo specificato ad es. ([{-/$
WCHR (Word Characters)
Consente la definizione dei caratteri validi di riferimento e la relativa traduzione ad es. da minuscolo in maiuscolo ovvero, la mappatura dei caratteri che compongono la parola (solitamente si passa da minuscolo a maiuscolo, si eliminano le vocali accentate)

- **Phrase processing** (trattamento delle stringhe)

1. DCLS (Deletion Clauses)

Consente la cancellazione di tutto ciò che è compreso tra incisi o clausole ad es. *[ ]*

2. DSTR (Deletion Strings)

Consente la cancellazione di stringhe ritenute inutili nella fase di codifica, ad es. NELL'

3. RSTR (Replacement Strings)

Consente la sostituzione di stringhe ad es. C/TERZI à.CONTOTERZO

- **Separazione in parole**

4. WCHR (Word Characters)

La separazione della stringa in parole avviene in questa fase, facendo riferimento al WCHR, in quanto i caratteri non contenuti in tale lista sono considerati come delimitatori delle parole.

- **Word Processing** (trattamento delle parole)

5. RWRD (Replacement Words)

Consente la vera e propria gestione dei sinonimi: sostituzione di parole, ad es. *MINUTO*à*DETTAGLIO,* così come la sostituzione con blank di parole ininfluenti, ad es. preposizioni, articoli, etc.

6. DWRD (Double Words)

Realizza la gestione dei sinonimi a livello di coppie di parole in sequenza. Consente infatti la sostituzione di due parole, che in sequenza acquistano un significato preciso, o con una parola o con due; ad es *ABBIGLIAMENTO FIRMATO*à *ABBIGLIAMENTO* oppure *ABBIGLIAMENTO MATRIMONIO* à *ABBIGLIAMENTO ADULTO*

7. HWRD (Hyphenated Words)

Consente il trattamento di parole separate dal trattino ad es. *BABY-SITTER* -à *BABYSITTER*

8. IWRD (Illegal Words)

Consente la cancellazione dei caratteri illegali o spuri all'interno delle parole ad es. *I NUMERI*

9. EXCP (Excepion Words)

Consente la definizione di parole che non debbono essere sottoposte all'eliminazione di suffissi, prefissi, ad es. *BORSA*à*BORSA*

10. MCHR (Multiple characters)

Consente la definizione di caratteri che se doppi o tripli all'interno di una parola vengono ridotti ad uno ad es. *ABBIGLIAMENTO* à *ABIGLIAMENTO;* può pertanto limitare l'effetto di errori d'ortografia

11. PRFX (Prefixes)

Consente di definire dei prefissi da eliminare; viene effettuato solo se la parola troncata rimane almeno di 4 lettere

12. SUFX (Suffixes)

Consente l'eliminazione dei suffissi delle parole ad es. *OA, OE, OI;* viene effettuato solo se la parola troncata rimane almeno di 4 lettere

Consente di definire dei prefissi da eliminare, viene effettuato solo se la parola troncata rimane almeno di 4 lettere

12. SUFX (Suffixes)

Consente l'eliminazione dei suffissi delle parole ad es. *OA, OE, OI;* viene effettuato solo se la parola troncata rimane almeno di 4 lettere

- **Post processing** (ulteriore trattamento delle parole)

13. RDUP (Remove Duplicates)

Consente la rimozione di parole doppie ossia duplicate.

14. SORT (Ordinamento delle parole)

La scelta è comunque dettata dal tipo di contesto

I processi descritti, dei quali l'unico obbligatorio è il WCHR, nonché l'ordine con cui eseguirli nell'ambito di ciascuna fase, vengono elencati nel '***Parsing Strategy File***' (Parsing Strategy File).

La strategia varia ovviamente in funzione del contesto applicativo di codifica. I processi inclusi nella strategia sono individuati da parole chiave (Process Keyword), mentre i parsing data file sono referenziati tramite delle parole chiave dei dati (Filedata Keyword).

In linea generale è bene, nella strategia di parsing, anteporre sempre il processo che definisce le parole eccezione (EXCP) a quello dei SUFX, dei MCHR e PRFX.

Il Parsing sviluppato per l'Ateco

Nella stesura dei file di parsing per la codifica automatica dell'Attività Economica si sono seguiti i seguenti criteri di carattere generale:

- Nell'ambito del **WORD PROCESSING,** ovvero del trattamento che riguarda le parole:

1) Tutte le parole che subiscono una trasformazione nelle DWRD sono state convertite **al maschile singolare** nella fase delle RWRD come per es. *Abrasiva*à*Abrasivo, Abrasive*à*Abrasivo, Abrasivi-*à*Abrasivo*. Ciò è stato ritenuto necessario per permettere di gestire nelle DWRD un numero minore di sinonimi con le relative schede.

E' stato necessario prevedere alcune **eccezioni** per dei sostantivi che assumono un significato diverso se usati al femminile/maschile, plurale/singolare, come per es. *maniche* à *manica*, mentre *manici* à *manico*.

Le ***gomme*** non sono state trasformate nel singolare (***gomma),*** perché, dall'esame dei file utilizzati per l'addestramento di ACTR, è emerso che il plurale della parola viene spesso usato come sinonimo di *pneumatici*. Non è stato creato neanche il sinonimo *gomme = pneumatici,* e si è convenuto invece che, quando si inserisce un'empirica che tratta questo prodotto, deve essere inserita sia l'empirica con *gomme* che quella corrispondente con *pneumatici*.

2) In caso di **parole che esprimono categorie** (es. *cereali,*), sono stati definiti sinonimi nelle RWRD, laddove tutte le specificazioni che in Italiano fanno parte della stessa categoria vengono trattate dall'ATECO nelle stesse categorie. Quindi per es. *avena, frumento, grano, granoturco, mais, orzo, riso, saggina, segale, sorgo-*à *cereale*.

Non è stato possibile invece ricondurre alla categoria degli *elettrodomestici* i prodotti che ne fanno parte, perché la classificazione Ateco distingue per es. in categorie diverse la produzione di *frigoriferi per uso domestico* da quelli per *uso industriale*

3) Abbreviazioni problematiche, ovvero quelle a cui non corrisponde univocamente una singola parola. Dall'esame del file utilizzato per l'addestramento, è stato stabilito di esplicitare l'abbreviazione in questo step (RWRD), soltanto qualora questa sottintenda una parola piuttosto che un'altra **in una percentuale di casi molto elevata** (criterio della *'massima frequenza'*).

Per esempio, è molto più frequente che

Inf à *Informatico*

piuttosto che

Inf à *Infisso*

oppure

Bib à *Bevanda (bibita, quindi Bevanda)*

piuttosto che

*Bib*à *Biblioteca*

La gestione delle abbreviazioni che sottintendono parole più rare è stata risolta nelle RSTR, prima che vengano esplose nella RWRD nelle parole più frequenti.

Per esempio:

Inf. alluminio à*Infissi in alluminio*

*Inf. in alluminio-*à*Infissi in alluminio*

*Inf. in all.-*à *Infissi in alluminio*

*Inf. in legno--*à*Infissi in legno*

*Inf. in vetroresina --*à *Infissi in vetroresina*

4) Le abbreviazioni problematiche a cui **non** è stato possibile applicare il criterio della '*massima frequenza'* (di cui al punto precedente) sono state risolte nelle DWRD, esplicitandole a seconda delle parole con cui sono abbinate. Il troncamento ***Cons,*** per esempio, può essere ricondotto sia a *Consulenza* che a *Conservazione*. Il trattamento di tali abbreviazioni è stato rimandato, come già detto, nella fase delle DWRD, utilizzando più schede che hanno permesso di trattare le parole in sequenza

*Cons Carne-*à*Conservazione Carne*

*Cons Finanza-*à*Consulenza Finanza*

Altri troncamenti che hanno comportato un lavoro complesso e non sempre, però, definitivamente risolutivo, vista la

trattare le parole in sequenza

*Cons Carne-*à*Conservazione Carne*

*Cons Finanza-*à*Consulenza Finanza*

Altri troncamenti che hanno comportato un lavoro complesso e non sempre, però, definitivamente risolutivo, vista la grande variabilità di risposta, sono stati sia l'abbreviazione ***Acq,*** che può essere per es. ricondotta ad *Acquisto, Acqua o Acquedotto,* che l'abbreviazione ***Conf*** che può essere ricondotta sia a *Confezionamento* che a *Confezioni*. Il criterio utilizzato anche in quest'ultimi casi è stato quello di riportare il significato dell'abbreviazione ad una o all'altra in funzione della seconda parola, ultilizzando sempre più schede nelle DWRD.

5) Abbreviazioni che contengono il punto internamente alla parola come per esempio *AMBUL.TE* (che sta per *AMBULANTE).*

Sono state risolte, quando possibile, nelle DWRD perché l'uso della trasformazione nelle RSTR (necessario invece per quelle che terminano per '*LA'* e '*LE'*) avrebbe potuto generare trasformazioni sbagliate come per esempio nel caso già specificato:

AMBUL.TESSILI che sarebbe diventato ***.****AMBULANTE.SSILI*

6) Non è stato possibile eliminare la ***congiunzione 'ed',*** come invece è stata eliminata la *congiunzione 'e',* in quanto spesso '*ed'* è utilizzato come abbreviazione di *edilizia*. Per **risolvere questa *abbreviazione 'ed'* con *edilizia***, si è proceduto ad inserire nel file di parsing, relativo alle DWRD, le possibili combinazioni di parole usate con l'abbreviazione di edilizia. Le schede che sono state previste sono le seguenti:

COSTRUZIONE	*ED*	*COSTRUZIONE*	*EDILIZIA*
FALEGNAMERIA	*ED*	*FALEGNAMERIA*	*EDILIZIA*
MATERIA	*ED*	*MATERIA*	*EDILIZIA*
PIETRA	*ED*	*PIETRA*	*EDILIZIA*
PONTEGGI/0	*ED*	*PONTEGGI/0*	*EDILIZIA*
ACCESSORIO	*ED*	*ACCESSORIO*	*EDILIZIA*
PREFABBRICATO	*ED*	*PREFABBRICATO*	*EDILIZIA*
CANTIERE	*ED*	*CANTIERE*	*EDILIZIA*
DEMOLIZIONI	*ED*	*DEMOLIZIONI*	*EDILIZIA*
SCAVO	*ED*	*SCAVO*	*EDILIZIA*
ARTIGIANO	*ED*	*ARTIGIANO*	*EDILIZIA*
ATTIVITA'	*ED*	*ATTIVITA'*	*EDILIZIA*
IMPRESA	*ED*	*IMPRESA*	*EDILIZIA*
CONSOLIDAMENTO	*ED*	*CONSOLIDAMENTO*	*EDILIZIA*
RESTAURO	*ED*	*RESTAURO*	*EDILIZIA*
RESIDENZIALE	*ED*	*RESIDENZIALE*	*EDILIZIA*
MANUTENZIONE	*ED*	*MANUTENZIONE*	*EDILIZIA*
MURATORE	*ED*	*MURATORE*	*EDILIZIA*
OPERA	*ED*	*OPERA*	*EDILIZIA*
STRUTTURA	*ED*	*STRUTTURA*	*EDILIZIA*
RECUPERO	*ED*	*RECUPERO*	*EDILIZIA*
RISTRUTTURAZIONE	*ED*	*RISTRUTTURAZIONE*	*EDILIZIA*
IMPRENDITORE	*ED*	*IMPRENDITORE*	*EDILIZIA*
SERVIZIO	*ED*	*SERVIZIO*	*EDILIZIA*
APPALTI	*ED*	*APPALTI*	*EDILIZIA*
CARPENTIERE	*ED*	*CARPENTIERE*	*EDILIZIA*
MATERIA	*ED*	*MATERIA*	*EDILIZIA*
ISOLANTE	*ED*	*ISOLANTE*	*EDILIZIA*
PAVIMENTO	*ED*	*PAVIMENTO*	*EDILIZIA*
DECORAZIONE	*ED*	*DECORAZIONE*	*EDILIZIA*
PITTORE	*ED*	*PITTORE*	*EDILIZIA*
COMPONENTE	*ED*	*COMPONENTE*	*EDILIZIA*
PRODOTTO	*ED*	*PRODOTTO*	*EDILIZIA*
ATTREZZATURA	*ED*	*ATTREZZATURA*	*EDILIZIA*
COOPERATIVA	*ED*	*COOPERATIVA*	*EDILIZIA*
PROGETTAZIONE	*ED*	*PROGETTAZIONE*	*EDILIZIA*
PERITO	*ED*	*PERITO*	*EDILIZIA*

7) Per le **parole singole che sottintendono la produzione di un prodotto** si è scissa la parola nella *Produzione + Prodotto* nelle RWRD, per es. *Calzaturificio-*à *Produzione Calzatura* mentre *Prosciuttificio-*à *Produzione Prosciutto*

Un'**eccezione** a questo criterio generale è rappresentata dalle parole '***Panificio** e sinonimi'* che sono stati trasformati in una parola unica '*Produzionepane'*. Ciò è stato necessario per gestire la codifica dell'attività di ' *produzione forni per panifici'* che, in assenza di tale eccezione, sarebbe stata trasformata in '*produzione forno produzione pane'* (quindi soppressa la parola doppia '*produzione'*) e attribuita al codice relativo al '*forno produzione di pane'*. Per le parole

Un'**eccezione** a questo criterio generale è rappresentata dalle parole '***Panificio** e sinonimi'* che sono stati trasformati in una parola unica '*Produzionepane'*. Ciò è stato necessario per gestire la codifica dell'attività di ' *produzione forni per panifici'* che, in assenza di tale eccezione, sarebbe stata trasformata in '*produzione forno produzione pane'* (quindi soppressa la parola doppia *'produzione')* e attribuita al codice relativo al '*forno produzione di pane'*. Per la parola *produzionepane* sono state previste poi nelle DWRD le stesse regole adottate per le doppie attività di *produzione* e *commercio, dettaglio e ingrosso*.
La parola ***panetteria,*** invece, in quanto può essere usata indifferentemente sia per indicare la *produzione del pane* che la *vendita del pane,* non ha subito nessuna trasformazione.

8) Per le **parole singole che sottintendono** già **una attività,** diversa dalla produzione, si è operato trasformando per es. *Autotrasportatore-*à*Auto Trasporto;* si è scissa cioè la parola nell'A*ttività + Mezzo* nelle RWRD

9) Un **caso particolare** è rappresentato da *Automercato,* sia esso scritto unito, separato o con il trattino. Questo è stato riportato alla parola unica *Commercioauto*. Ciò ha permesso di poter gestire come coppie di parole, al livello di DWRD le regole di prevalenza rispetto per esempio alla vendita d*i Autoricambi*

10) Per le **parole** che possono presentarsi unite, separate dal trattino o doppie, per es. quelle che **indicano un mezzo di trasporto** come *Auto Carro*, *Auto Articolato,* per cercare di uniformarne il trattamento, si sono riportate, nelle DWRD, ad un'unica parola *Autocarro*. Lo stesso procedimento si è avuto per il *Super mercato*à*Supermercato (*comprese le relative abbreviazioni)

11) Sono state **unificate** nelle DWRD **parole** che singolarmente potevano generare falsi match, mentre accoppiate portano necessariamente ad abbinamenti esatti. Es. *Polizia Stato* à*Poliziastato, Pubblica Sicurezza* à*Poliziastato*.

12) Si è tentato di **evitare ridondanze**, pertanto i casi di parole doppie come per es. *Commercio Ingrosso*à*Ingrosso e Commercio Dettaglio*à*Dettaglio* sono stati risolti nelle DWRD, come si nota, con un'unica parola. Per il *Commercio Ambulante* non si è proceduto così perché la parola ambulante non è esclusiva del settore del commercio ed inoltre la parola commercio è stata ritenuta importante per l'assegnazione del filtro (commercio ambulante filtro 50)

13) Un lavoro maggiore e un po' più complesso hanno invece richiesto le **doppie attività** con la definizione della rispettiva **prevalenza**. Su indicazione del servizio CUE sono state inserite nei file di parsing le seguenti regole di carattere generale:
Produzione >commercio
Produzione > ingrosso
Produzione > dettaglio
Trasformazione >commercio
Trasformazione >ingrosso
Trasformazione >dettaglio
Produzione >commercializzazione
Trasformazione >commercializzazione
Lavorazione>commercio
Lavorazione>ingrosso
Lavorazione>dettaglio
Produzione >riparazione
Commercio >riparazione
Tali regole, a meno dei sinonimi tipo *Vendita-*à*Commercio* e *Fabbricazione-*à*Produzione* (gestiti nelle RWRD), sono state tutte inserite nelle DWRD e pertanto sono considerate ***regole generali*** che vengono sempre rispettate nel passaggio di codifica automatica. Per far sì che tali regole siano sempre rispettate, indipendentemente dalla sequenza delle parole, sono state previste per ogni regola due schede per esempio:
*Produzione commercio*à*produzione*
Commercio produzione à*produzione*

14) I casi **eccezione rispetto alle regole generali** devono essere risolti nelle RSTR, prima che le parole vengano trasformate nelle DWRD in funzione delle regole.
Per esempio nel caso del *Commercio e riparazione di pneumatici* (purché non all'ingrosso) prevale l'attività di *riparazione,* al contrario della regola generale. Quindi si è deciso di trasformare nelle RSTR le coppie di parole '*commercio pneumatici'* e *'riparazione pneumatici'* in parole uniche, (*commerciopneumatici e riparazionepneumatici)* in modo che successivamente, nelle DWRD, *commercio riparazionepneumatici* à r*iparazionepneumatici*.
Nella RSTR è stata prevista tutta la casistica (gomme e pneumatici, abbreviazioni possibili ed eventuale presenza della preposizione 'di' tra le parole).
Si sono volutamente tenute distinte le parole *Gomma* e *Gomme* (intese nel senso di pneumatici). Quest'ultima parola è stata resa parola eccezione (nelle EXCP).

15) Nelle RWRD sono state anche gestite numerose **parole** che sono state ritenute **ininfluenti** ai fini della codifica come per es. *Affine, Annesso…..* che sono state riportate a blank, come le preposizioni semplici e articolate.

16) Nelle EXCP sono state gestite numerose **parole** che, per il loro **significato non unico**, potevano comportare dei match non corretti e non univoci come per es. *Confezioni* che volutamente è rimasta distinta dalla *Confezione-*à*Confezionamento,* in quanto è stata equiparata all'*Abbigliamento*.

- Nell'ambito del **PHRASE PROCESSING**, ovvero nel processo che riguarda il trattamento delle stringhe:

1) Le abbreviazioni problematiche (vedi terzo punto del *Word Processing*), che sottintendono parole più rare, sono state risolte nelle RSTR, prima che vengano esplose nella RWRD nelle parole più frequenti.
Per esempio la stringa *Inf. in alluminio-*à *Infissi in alluminio* prima che venga effettuata la trasformazione *Inf=Informatico* nelle RWRD

2) Si è ricorso all'uso delle RSTR nel **caso di preposizioni semplici che risultavano determinanti** al fine dell'attribuzione del codice. L'alternativa di considerarle sempre parole determinanti ai fini della codifica, ovvero di non sostituirle con blank nella RWRD, è stata scartata, perché si è ritenuto che ciò avrebbe pesato negativamente sul tasso di codifica.
Per esempio nel caso della preposizione 'IN' abbinata a 'TESSUTO' c'era la necessità di differenziare i due testi *'Commercio dettaglio abbigliamento in tessuto'* e *'Commercio al dettaglio di tessuti per abbigliamento'*. A tal fine nelle RSTR:

IN TESSUTO à*.INTESSUTO*.

Ciò è stato fatto anche per *IN STOFFA*à*.INTESSUTO*.

3) Relativamente all **congiunzione 'e'** è stato adottato un trattamento diverso a seconda che sia espressa **'e'** oppure '**ed**'. La prima, infatti, è stata sostituita con blank nelle RWRD, la seconda invece, potendo rappresentare anche l'abbreviazione di 'edilizia', come già descritto, ha subito un doppio trattamento. Per l'uso della *preposizione ed* davanti alla parole che iniziano per vocale sono state realizzate nelle RSTR le seguenti trasformazioni:

ED A	*. E A*
ED E	*. E E*
ED I	*. E I*
ED O	*. E O*
ED U	*. E U*

Queste schede hanno permesso di uniformare, nel caso della congiunzione con parole che iniziano per vocale, la **e** con la **ed.** Successivamente, infatti, nello step delle RWRD la **e** verrà sostituita con blank.
Anche per **l'articolo indeterminativo 'un'** si è reso necessario prevedere più schede, per uniformarlo alla lettera **'u'** che nel passaggio successivo delle RWRD viene sostituita con blank.
Sono state per esempio inserite le seguenti trasformazioni

UN B	*.U B*
UN C	*. U C*

4) L'uso delle RSTR si rileva ancora molto utile nella **gestione delle sigle** come ad es. *P.S.*à*Poliziastato* oppure *V.F.*-à *Vigile fuoco*. Per le sigle molto brevi e problematiche per esempio *N.U.*à*Nettezza Urbana* bisogna ricordare di lasciare un blank prima di scrivere la stringa nella prima colonna per evitare che la trasformazione sia effettuata su altre sigle che la contengono esempio:

*O.N.U.*sarebbe stato trasformato in à*O.Nettezza urbana* se non fosse stato lasciato il blank

5) Nel caso di stringhe che contengono delle abbreviazioni sulla parola finale, gestite nella RSTR, **bisogna sempre** completare l'abbreviazione con un **punto,** altrimenti la trasformazione potrebbe dare adito a match non corretti ad es. APPALTI ED. à APPALTI EDILI se non fosse chiusa dal punto trasformerebbe APPALTI EDILI-àAPPALTI EDILI.ILI

6) Nelle RSTR è molto importante anche l'**ordine di successione delle varie trasformazioni** da eseguire. Bisogna, pertanto, fare in modo che nella successione delle trasformazioni, le schede più brevi, contenute in un'altra scheda più lunga a sua volta oggetto di trasformazione, vengano sottomesse dopo quello più lunghe. Chiarendo con un esempio nella trasformazione che segue è stato stabilito il seguente ordine

C/LAVORAZIONI à*.CONTOLAVORAZIONE*.
C/LAVORAZIONE à *.CONTOLAVORAZIONE*.
C/LAVORAZION à *.CONTOLAVORAZIONE*.
C/LAVORAZ à *.CONTOLAVORAZIONE*.
C/LAV à *.CONTOLAVORAZIONE*.
C/L à *.CONTOLAVORAZIONE*.

Infatti se *C/LAV* à *.CONTOLAVORAZIONE* fosse stata la prima dell'elenco, *C/LAVORAZIONE* sarebbe stata trasformata in *CONTOLAVORAZIONE. ORAZIONE*.
Grazie a questo accorgimento, è stato possibile omettere il punto finale in alcune abbreviazioni.

Concludendo è chiaro che l'uso delle RSTR è comunque molto laborioso e non sempre può risultare risolutivo, in quanto rimane assolutamente legato alla stringa prevista, che invece può presentare una variabilità elevatissima.

Strategia utilizzata nell'applicazione di codifica

Strategia utilizzata nell'applicazione di codifica

Come già detto nel *Parsing strategy file* è possibile specificare quali fasi della standardizzazione debbono essere eseguite ed in quale ordine. Nell'applicazione sviluppata per l'ATECO, la strategia utilizzata prevede i seguenti processi così ordinati:

phraseProcess1	DCLS
PhraseProcess2	DSTR
PhraseProcess3	RSTR
WordProcess1	RWRD
WordProcess2	DWRD
WordProcess3	DWRD
WordProcess4	DWRD
WordProcess5	EXCP
WordProcess6	SUFX
WordProcess7	
WordProcess8	
PostProcess1	RDUP
PostProcess2	SORT
Autotrim	Yes
TrimLeft	)]}-/\!?,.
TrimRight	([{-/$
Hyphen	-

- Normalmente il file delle **HWRD (Hyphenated Words)** consente il trattamento di parole separate dal trattino ad es. *BABY-SITTER*à *BABYSITTER,* come già detto sopra. In questa applicazione, nel file WCHR (Word Characters), dove viene eseguita la traduzione dei caratteri, il carattere *hyphen* (-) non è stato inserito, pertanto, nella fase di separazione in parole, le parole scritte col trattino vengono trattate come due parole singole. Ne consegue che è stato sufficiente far diventare parola unica le parole originariamente scritte con trattino nel file DWRD, tenendo presente eventuali trasformazioni già avvenute a livello di singola parola nel RWRD.

- Il far girare tre volte il processo **DWRD (Double words)** è un'altra peculiarità di questa applicazione: è stato necessario, infatti, per consentire una trasformazione corretta di descrizioni che contenevano al loro interno clausole di esclusione che non potevano essere abolite con il processo DCLS, ma che, d'altro canto, se non trattate, avrebbero rischiato di generare match non corretti. Un esempio per chiarire, è l'attività economica: *'ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)'*. Era necessario, pur mantenendo la clausola di esclusione, far sì che le singole parole 'MINERALI', 'URANIO' e 'TORIO' non generassero match con altre attività che includevano l'estrazione di questi minerali. Con questa strategia di *parsing*, la stringa originaria ha subito le seguenti trasformazioni:

Original Text:	ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)
String trimming	ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)
Word Characters (Translation) ...	ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)
Deletion Clauses	ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)
Deletion Strings	ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)
Replacement Strings	ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)
Word Characters (Elimination) ...	ESTRAZIONE DI MINERALI METALLICI NON FERROSI ESCLUSO I

Replacement Strings	ESTRAZIONE DI MINERALI METALLICI NON FERROSI, (ESCLUSO I MINERALI DI URANIO E DI TORIO)
Word Characters (Elimination) ...	ESTRAZIONE DI MINERALI METALLICI NON FERROSI ESCLUSO I MINERALI DI URANIO E DI TORIO
Replacement Words	ESTRAZIONE MINERALE METALLO NON FERRO NON MINERALE URANIO TORIO
Double Words	ESTRAZIONE MINERALE METALLO NONFERRO NON URANIO TORIO
Double Words	ESTRAZIONE MINERALE METALLO NONFERRO NONURANIO TORIO
Double Words	ESTRAZIONE MINERALE METALLO NONFERRO NONURANIO NONTORIO
Exception Words	ESTRAZIONE MINERALE METALLO NONFERRO NONURANIO NONTORIO
Suffixes	ESTRAZION MINERAL METALL NONFERR NONURAN NONTOR
Duplicate Word Removal	ESTRAZION MINERAL METALL NONFERR NONURAN NONTOR
Word Sorting	ESTRAZION METALL MINERAL NONFERR NONTOR NONURAN
Parsed Text:	ESTRAZION METALL MINERAL NONFERR NONTOR NONURAN

Si evince da tale esempio la necessità del passaggio per tre volte del file delle DWRD.

- I file **MCHR (Multiple characters)** e **PRFX (Prefixes)** non sono stati utilizzati nella strategia per la codifica della variabile Ateco in quanto non ritenuti utili per questo contesto.