Quaderni Taltac

Sergio Bolasco

TaLTaC$^{2.10}$

SVILUPPI, ESPERIENZE
ED ELEMENTI ESSENZIALI
DI ANALISI AUTOMATICA DEI TESTI

LED
Edizioni Universitarie di Lettere Economia Diritto

ISBN 978-88-7916-459-7

LED Edizioni Universitarie di Lettere Economia Diritto
Via Cervignano 4 – 20137 Milano
Catalogo: www.lededizioni.com – E-mail: led@lededizioni.com

Testo on line: www.ledonline.it/taltac.html

Stampa: Digital Print Service

INDICE

INTRODUZIONE

L'idea dei ***Quaderni TaLTaC*** nasce dal voler raccogliere una serie di contributi di taglio differente, dal teorico al metodologico, dall'applicativo al caso di studio, su *come fare ricerca su dati non strutturati* con un impianto di tipo metrico, utilizzando strumenti della statistica e della linguistica computazionale. È questa infatti la logica e il tratto che accomunano i lavori della comunità di centinaia di ricercatori che ormai da una ventina di anni si riuniscono nei convegni JADT, le giornate internazionali di analisi statistica dei dati testuali. Nel seguito per riferirci a questo tipo di studi useremo l'acronimo ADT. Lo strumento intorno al quale ruotano i contributi dei ricercatori italiani è, fra gli altri, il software **TaLTaC^2**. Acronimo [1] che sta per **T**rattamento **A**utomatico **L**essicale e **T**estuale per l'**A**nalisi del **C**ontenuto / di un **C**orpus. Questi quaderni potranno raccogliere tali esperienze più ampiamente di quanto non consentano delle comunicazioni congressuali.

Questo primo contributo, come *numero zero* della serie, vuole colmare un vuoto, in quanto è mancata finora una presentazione autentica ed estesa del software TaLTaC^2, soprattutto aggiornata alla sua più recente versione, la 2.10 del giugno 2010 (www.taltac.it).

In una prima sezione, si ricostruiscono alcuni elementi di storia dei primi dieci anni di TaLTaC, si propongono in forma sintetica le principali funzionalità del programma e si dà cenni su una strategia di analisi. Seguono gli ambienti e le logiche d'uso del programma nonché il modello generale di costruzione della collezione dei testi. In una seconda sezione, si illustrano alcuni elementi delle fasi essenziali di lavoro sui dati testuali con dettagli su step e funzionalità, rimandando alla *Guida online* di TaLTaC^2 per la descrizione analitica dei comandi e delle pro-

[1] La pronuncia corretta del nome del software è Tàltac, con l'accento sulla prima sillaba.

cedure. L'esposizione seguirà nel suo sviluppo logico gli argomenti dei menu che a loro volta sono stati improntati ad una pratica didattica delle funzioni dell'analisi automatica dei testi. Ogni argomento contiene una breve presentazione delle principali caratteristiche e, laddove utile, un esempio che ne illustra il loro funzionamento o il risultato del corrispondente passo di analisi.

1.
ELEMENTI DI QUADRO GENERALE

1.1. UN PO' DI STORIA: SVILUPPI ED ESPERIENZE

Taltac, ideato nel 1999 da Sergio Bolasco e sviluppato grazie alla collaborazione di Francesco Baiocchi e Adolfo Morrone presso l'Università degli studi di Roma "La Sapienza", intendeva rispondere all'esigenza di ridurre, se non eliminare, le onerose operazioni di preparazione del testo prima di ogni analisi statistica su dati testuali. Nella sua prima release (versioni 1.0-1.5 sviluppate nel periodo 2000-2005) era un programma prevalentemente orientato al livello lessicale di analisi del contenuto e dotato di alcune prime risorse disponibili. Nella seconda release (2005-2010), con l'ulteriore apporto di Alessio Canzonetti, il software ha progressivamente implementato anche funzionalità di tipo testuale fino a trasformarsi in una piattaforma, dotata strumenti di text mining e arricchita nelle risorse statistico-linguistiche. In questi anni le collaborazioni alla realizzazione di questa piattaforma hanno visto la partecipazione di molti laureandi, dottorandi, assegnisti, ricercatori e tecnici informatici fra i quali: Franca Basilotta, Claudia Brunini, Federico Capo, Simona Carbone, Marco Castagna, Isabella Chiari, Elisabetta Davino, Francesca Della Ratta, Marina De Palo, Michelangelo Misuraca, Matteo Morganti, Pasquale Pavone, Bhupesh Singh, Arjuna Tuzzi. A vario titolo, essi hanno contribuito a mettere a punto algoritmi, nuove funzionalità, risorse statistico-linguistiche, grammatiche locali, dizionari tematici, l'help on-line, nonché l'interfaccia in inglese.

A partire dal 2000, in questi dieci anni, sono stati svolte varie attività intorno a Taltac. In particolare; corsi presso la Società Italiana di Statistica, il Censis, l'Inea, il Mides, l'Autorità garante per la concorrenza e il mercato (antitrust), la Faculté de Lettres di Besançon; in due scuole

estive di Bertinoro organizzate dalla Società Italiana di Psicologia; nonché per i dottorandi della Facoltà di Psicologia2 della Sapienza di Roma, per quelli dei Dipartimenti di Scienze Sociali delle Università di Torino e di Firenze; per il corso Madit organizzato presso la Facoltà di Psicologia dell'Università di Padova, oltre che per gli specializzandi della Facoltà di Statistica della stessa università. Dal novembre 2005 si organizzano periodicamente, tre volte l'anno, corsi di formazione di base e avanzati sull'analisi automatica dei testi, fondata sul software Taltac. Questi corsi, della durata di 2-3 giorni, sono stati svolti alla SAPIENZA fino al 2007 presso il CITICoRD e dal 2008 presso il Dipartimento di Studi geoeconomici linguistici statistici storici per l'analisi regionale [1]. In tutti i corsi effettuati finora sono passati oltre 400 fra studenti, dottorandi, ricercatori e docenti. Ad essi vanno aggiunti i corsi svolti presso il Master "Meters" di Fonti, Strumenti e Metodi per la Ricerca Sociale nella Facoltà di Statistica dell'Università di Roma "La Sapienza", e i corsi tenuti successivamente da altri docenti presso le loro università fra cui la Facoltà di Scienze della Comunicazione della Sapienza, le Facoltà di Psicologia e di Statistica dell'Università degli studi di Padova, le Università di Macerata e di Viterbo.

Oltre ai corsi, da ricordare tre seminari organizzati presso la facoltà di Economia della Sapienza in forma di "Giornate Taltac" rispettivamente nel gennaio 2002, nel dicembre 2003 in occasione della presentazione delle versioni Taltac 1.5 e 1.6 e, successivamente, nel giugno 2007 intorno alla versione 2.5. Mentre i primi due avevano carattere prevalentemente di *tutorial* sul programma, ed erano indirizzati ad una cinquantina di partecipanti ciascuno, il terzo ebbe il carattere di workshop per uno scambio di esperienze fra ricercatori e analisti con interessanti relazioni non pubblicate, sfociate poi in presentazioni al JADT. Questi quaderni potrebbero essere il luogo per raccogliere tali contributi in futuro.

Attualmente **TaLTaC**[2] è presente in oltre 120 fra dipartimenti e centri di ricerca in Italia e all'estero, per un insieme di oltre mille istallazioni in dieci anni.

Molte ricerche svolte presso università o centri di ricerca esterni testimoniano l'uso di TaLTaC nel produrre risultati nelle analisi su dati testuali. Alcune fra queste ricerche reciprocamente hanno generato problemi e offerto l'occasione per sviluppare algoritmi successivamente implementati in TaLTaC. Qui nel seguito se ne richiamano gli esempi dei

[1] Si veda la pagina: http://geostasto.eco.uniroma1.it/corsi/?pagina=corsi&tipo=3

gruppi di lavoro originati principalmente nella facoltà di Economia della Sapienza, aprendo quindi una vista su riferimenti spesso frammentati e dispersi. Le pubblicazioni prodotte in questi anni sono riunite qui per tema o tipo di contributo. Fra i titoli citati, ve ne sono anche di letteratura "grigia" (lavori non pubblicati, citati in bibliografia con l'asterisco), alcuni dei quali disponibili sul sito di Taltac (*www.taltac.it*).

Le raccolte in Italia di articoli sull'ADT secondo l'approccio "metrico", che a noi più interessa, riguardano gli atti del convegno del 1992 (Cipriani, Bolasco, eds. 1995), del 3^ JADT svoltosi presso la sede del CNR a Roma (Bolasco, Lebart e Salem, eds. 1995), della Giornata di studi sulle applicazioni svoltasi presso la facoltà di Statistica (Aureli, Bolasco, eds. 2004), del 10^ JADT del 2010 presso la facoltà di Economia della Sapienza (Bolasco, Chiari e Giuliano, eds. 2010). L'insieme invece degli 800 contributi presentati nelle dieci edizioni del JADT, sono disponibili online alla pagina: http://jadt.org (a partire dall'edizione del 1998).

I primi impianti di strategie di analisi che prenderanno forma poi nell'architettura di Taltac risalgono alla presentazione fatta nel 1997 alla riunione Cladag di Pescara (Bolasco, Morrone e Baiocchi 1999), al 5^ JADT (Bolasco 2000a), al convegno sulla Ricerca qualitativa nella psicologia sociale (Bolasco, Giovannini 2002).

Alcune sintesi sui principi epistemologici dell'analisi dei dati testuali (Bolasco 2004a) e sui fondamenti di analisi lessicale e di analisi testuale in Taltac si trovano su Quaderni di Statistica (Bolasco 2005b), nonché nella relazione al convegno Giscel su applicazioni alla linguistica dei corpora (Bolasco 2008a).

Per quanto riguarda i contributi di carattere metodologico, una discussione sul trattamento di forme testuali si trova nella relazione alle prime giornate JADT di Barcellona (Bolasco, 1990), poi ripreso e approfondito nel convegno su "Ricerca qualitativa e computer" (Bolasco, 1992) e nel 2^ JADT (Bolasco 1993). L'identificazione del linguaggio peculiare, la nozione di isofrequenza e il suo uso per individuare locuzioni grammatical, nonché una simulazione con il metodo *bootstrap* per valutare le scelte di lemmatizzazione sono oggetto nel 1996 di una relazione invitata al convegno dell'IFCS a Kobe (Bolasco, 1998). Alcuni articoli sulle risorse statistico-linguistiche quali i dizionari di frequenza su vari campioni di italiano sia per rilevare *multiwords* (Bolasco, Morrone 1998), sia per ricercare neologismi ed obsolescenti (Bolasco, Canzonetti 2005; Bolasco 2005a) sono testimonianze utili di analisi secondo i principi della linguistica dei corpora. Un confronto di applicazioni del

dizionario positivo/negativo è presentata al JADT di Lovanio (Bolasco, della Ratta-Rinaldi 2004). Un'idea di come sfruttare lessici di frequenza per individuare strutture ad elementi variabili, in un quadro di formalizzazione dei dati simbolici, è presente nei lavori con Rosanna Verde e Simona Balbi al 6^ JADT (Bolasco et al. 2002a; Balbi et al, 2002b). Negli anni più recenti le proposte metodologiche si sono concentrate sulla formulazione di modelli di grammatiche locali per la costruzione di risorse, rispettivamente su locuzioni di luogo (Bolasco, Pavone 2010) e verbi idiomatici (Bolasco 2010). Infine un altro lavoro riguarda l'integrazione fra classificazione automatica di tipo non supervisionato e uso del TFIDF per giungere ad una categorizzazione automatica di tipo fuzzy (Bolasco, Pavone 2008).

Per quanto riguarda i contributi sul linguaggio politico, è del 1996 un'analisi del discorso programmatico di governo, risultato dello studio di un corpus dei discorsi tenuti in Parlamento dai presidenti del Consiglio incaricati per presentare il loro programma di governo. Riguardano i 48 governi della prima Repubblica, quindi dal 1947 al 1994 (Bolasco, 1996). Fa seguito, nel 2000, un approfondimento sul corpus delle cosiddette "Repliche" tenute da ciascun presidente incaricato, alla fine del dibattito parlamentare sul programma, appena prima del voto di fiducia delle Camere (Bolasco 2000b). Successivamente nel 2006, in collaborazione con Nora Galli de' Paratesi, linguista, e Luca Giuliano, sociologo, uno studio su un corpus altamente rappresentativo dei discorsi di Silvio Berlusconi (Bolasco et al. 2006), sia nei periodi in cui era al governo, sia in quelli passati all'opposizione, ha "misurato" sotto varie prospettive l'esperienza politica del premier dalla discesa in campo del 1994 al settembre 2005 (terzo governo).

Altri esempi di applicazioni e casi di studio (Bolasco 1999) riguardano l'analisi di: i) risposte libere a domande aperte (indagine Censis, più approfondita in: Bolasco, 2001a); ii) scambi epistolari dalla ricerca sul "Contadino polacco"; iii) interviste sulla customer satisfaction bancaria. E ancora, una ricerca Irrsae su un sito di insegnanti "Funzione Obiettivo" con relativo studio dei vari forum (Bolasco 2001b); una ricerca Irre con un'analisi di focus group sull'educazione alimentare (Bolasco 2004b); un'analisi lessicale e testuale nell'indagine TUS dell'Istat relativa ai diari del tempo su un campione di 50.000 individui (Bolasco et al. 2007) con esempi di analisi per concetti, di processi di tipo ETL per il calcolo di statistiche tradizionali, di costruzione di risorse. Infine, fra i lavori non pubblicati compiutamente ma spesso riportati in altri articoli, un'analisi sul lessico enogastronomico delle guide del Gambero Rosso (Bolasco S.,

Bolasco M. 2004), nonché su tipologie della ristorazione italiana nelle varie regioni (Bolasco 2008b).

Infine vari sono i contributi specifici sul Text Mining (TM), rispettivamente provenienti da: i) i lavori del progetto europeo Nemis [2] (Bolasco et al. 2004a; Bolasco et al. 2005c), ii) una rassegna con oltre 100 schede sulle applicazioni del TM a livello internazionale (Bolasco, Canzonetti e Capo 2005b); iii) un'attività presso l'Autority italiana dell'antitrust per l'individuazione di entità nominate con il riconoscimento automatico dei nomi di impresa (Baiocchi et al. 2005). Da segnalare inoltre una rassegna non pubblicata sulle applicazioni del TM in ambito bancario (Bolasco 2004c).

1.2. Le principali funzionalità di TaLTaC^2 in sintesi

TaLTaC^2 serve ad analizzare documenti o collezioni di dati espressi in linguaggio naturale, e – più in generale – corpus testuali, sfruttando *risorse* sia di tipo *statistico*, sia di tipo *linguistico*, fortemente integrate fra loro. L'acronimo individua le finalità stesse del software: sviluppare un trattamento del testo in modo automatico, a livello sia lessicale che testuale, finalizzato all'analisi del contenuto o all'analisi del testo a prescindere dal suo contenuto, ovvero gli elementi strutturali del corpus.

Alcune fasi di tale trattamento costituiscono una **preparazione** indispensabile del testo per le successive analisi svolte in TaLTaC^2 (o in altri software) sia nella logica di *Text Analysis* (TA) che di *Text Mining* (TM).

Tali analisi consentono di **dare rappresentazioni** del fenomeno studiato sia a livello di unità di testo ("parole" in senso lato) sia a livello di unità di contesto ("documenti" o frammenti di testo). L'approccio seguito consente di **non leggere** materialmente la collezione di testi e quindi di analizzare il corpus indipendentemente dalla sua dimensione, che può essere anche vastissima (milioni di parole). TaLTaC^2 elabora attualmente files fino a 130 MegaBytes, equivalenti a 40.000 pagine di testo e decine di milioni di parole.

Le diverse funzioni di TaLTaC^2 nel loro insieme costituiscono una "*cassetta degli attrezzi*" per svolgere le operazioni fondamentali di trat-

[2] http://cordis.europa.eu/fetch?CALLER=PROJ_ICT&ACTION=D&CAT=PROJ&RCN=64001

tamento del linguaggio naturale, di ricerca ed estrazione d'informazioni dal testo, nonché per fare **annotazioni del lessico** (vocabolario del corpus) e per la **categorizzazione automatica** dei documenti. Ciò avviene sia facendo ricorso ad elementi esogeni al corpus (risorse disponibili in TaLTaC^2), sia facendo ricorso ad algoritmi basati su caratteristiche endogene della collezione di testi.

L'analisi svolta in TaLTaC^2 *permette di selezionare ed estrarre l'informazione più significativa* dal corpus oggetto di studio, sia a livello **lessicale** in termini di differenti tipi di linguaggio: *peculiare*, *caratteristico*, *rilevante*, sia a livello **testuale** in termini di *entità di interesse*, di strutture linguistiche o formali, di concetti o tratti semantici, di *multi-word expressions*. Queste ultime sono "parole di più parole", che spesso hanno significato polirematico e quindi catturano il vero senso nel testo.

L'informazione estratta grazie a TaLTaC^2 può essere **esportata** in forma di **corpus annotato** (anche sub-corpus) o di **matrici**. Queste sono: i) tabelle di frequenza [parole x testi], ii) tabelle sparse [3] prevalentemente booleane, [frammenti x parole selezionate], spesso trasposte in forma di [termini x documento]. Sia il corpus annotato, sia le matrici sono esportate per essere analizzate con altri software statistici di analisi testuale (quali, ad esempio: Lexico, Alceste, Hyperbase, Sphinx, Wordmapper, Wordstat) o di analisi qualitativa (NVivo, Atlas) e quantitativa (Sas, R, Spad, Spss).

Le numerose funzionalità del programma consentono le operazioni classiche nella tradizione lessicometrica: dall'analisi dei **segmenti ripetuti** all'analisi delle **concordanze**, dal calcolo dei parametri della **legge di Zipf** a quello della **dispersione e uso** delle parole secondo differenti fonti, dall'analisi delle **specificità** al calcolo delle **co-occorrenze**.

Inoltre TaLTaC^2 offre sofisticati strumenti di **ricerca di informazioni** tipici del ***text mining*** sia sulle parole indicizzate, sia sulle unità di contesto definite dalla frammentazione del corpus o dalla collezione dei testi. Ogni elemento della collezione può essere anche diviso in **sezioni**. Tali strumenti si fondano su **query** singole o multiple, *predefinite* (quindi rapidamente riapplicabili) o *personalizzabili* per ogni singolo corpus.

In particolare in TaLTaC^{2,} nelle fasi esplorative del testo e in quelle di training per l'approntamento di modelli (ricerca di entità, costruzione di grammatiche locali), è possibile mettere a punto specifici **piani di**

[3] Si dicono sparse le matrici con oltre il 95% di zeri.

lavoro. Questi consistono in filiere di istruzioni (files *batch*) anche complesse, per le annotazioni delle entrate del vocabolario o per lanciare processi di categorizzazione dei documenti o di altre unità di contesto. Così facendo si svolgono funzioni tipiche del Text Mining come operazioni di **ETL** (Extraction, Transformation, Loading) ossia l'estrazione dal testo di informazione non strutturata, la sua trasformazione in dato codificato e l'archiviazione di quest'ultimo in tabelle strutturate. Questo processo consente successivamente la **categorizzazione automatica supervisionata** dei documenti.

Come ogni altro software per l'analisi di dati testuali, TaLTaC^2 consente una tokenizzazione automatica delle parole del testo e quindi è **indipendente dalla lingua,** per cui è possibile di fatto analizzare in egual modo testi in italiano o in inglese, così come in qualsiasi altra lingua o alfabeto, incluso l'arabo, il cinese o altro. Tuttavia utilizzando **risorse linguistiche** – dizionari di singole lingue – TaLTaC^2 è in grado di effettuare il **tagging grammaticale**, ossia l'annotazione di informazioni linguistiche alle parole del testo. Tutte le funzionalità di TaLTaC^2 che non dipendono dal tagging grammaticale, restano indipendenti dalla lingua. Il dizionario di Taltac è costituito da circa 74.000 lemmi, pari a oltre 530.000 forme flesse. Per completare i riconoscimenti di parole non contenute nel dizionario, sono disponibili centinaia di **algoritmi** in grado di individuare numeri scritti in lettere (centodiecimila), enclitiche verbali (dammi, portategli), derivati (buonista, craxiano) e alterati (tavolinetto), forme con prefissi (neoformazioni), forme complesse (iperberlusconismo).

Dalla versione 2.8 è possibile sfruttare **corpus lemmatizzati da Tree Tagger** [4] non solo in italiano, ma anche in altre quattro lingue europee (francese, inglese, spagnolo e tedesco). Si rimanda alla Guida per i particolari.

In TaLTaC^2 è possibile scegliere l'interfaccia dei **menu** e della **diagnostica** anche **in inglese**, così da facilitarne l'uso a livello internazionale. Ogni operazione svolta in TaLTaC2 è descritta in un **giornale della sessione di lavoro** e pertanto è ricostruibile per essere ripetuta successivamente.

È possibile operare **annotazioni** del vocabolario non soltanto di natura **linguistica** (morfo-grammaticale), ma anche di tipo **semantico**. Ciò con-

[4] Software sviluppato da Helmut Schmid all'istituto di Linguistica computazionale della Università di Stoccarda: http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

sente l'**analisi di concetti**, di entità "nominate" (riconoscimento automatico di nomi, toponimi, sigle ecc.) o lo studio sistematico del lessico per classi di parole, ove queste ultime possono comprendere allo stesso tempo sia singole forme o lemmi, sia *multiword expressions.* Una volta definite nel vocabolario del corpus, ossia a livello lessicale, queste entità possono essere ricercate nel testo, ossia a livello testuale. Fra le risorse semantiche disponibili da citare un dizionario di 6.000 forme flesse di aggettivi classificati come positivi o negativi (Bolasco, della Ratta-Rinaldi 2004). Il confronto tra il vocabolario del corpus e questo dizionario consente di apprezzare la tonalità negativa o positiva del corpus: il cosiddetto "***sentiment***" di un testo. Fra le altre risorse disponibili in TaLTaC$^{2.10}$: una query predefinita sulle nazionalità ed etnie, un dizionario e una query sulle figure parentali, e altri dizionari rispettivamente sui crononimi (nomi del tempo), sulle locuzioni di luogo (delle attività quotidiane), sull'enogastronomia.

Sempre nella versione 2.10 di TaLTaC^2 è presente un algoritmo (Bolasco, 2010) per individuare in maniera automatica le locuzioni verbali a partire da una risorsa di 5000 **verbi idiomatici** derivata dal Gradit (De Mauro, 2007), in grado di riconoscere e raggruppare a lemma locuzioni verbali anche con inserti fra verbo e collocato.

Tali operazioni avvengono in fase di **normalizzazione** nel pre-trattamento del testo o nelle fasi di **tagging semantico**. Questi step consentono, fra le altre funzionalità, di **lessicalizzare** locuzioni **polirematiche** (*multiword expressions* di significato idiomatico) o **collocazioni** di interesse (forme cristallizzate).

Nel corso dell'analisi lessicale, TaLTaC^2 permette di **selezionare parti di lessico** (Vocabolario del corpus) con differenti criteri.

In primo luogo il **linguaggio peculiare** (LIPE), ottenibile in modo *contrastivo* per confronto della frequenza delle unità lessicali nel corpus con la frequenza che queste hanno in un lessico di riferimento (Bolasco, 1999, pp. 2232-234). In tal modo, sulla base del sopra-/sotto-uso rispetto ad un valore atteso di frequenza, si concentra l'attenzione sul nucleo ("core") del vocabolario, in quanto tale confronto fornisce la parte essenziale del linguaggio, quella inerente i principali contenuti del testo analizzato. Ciò è possibile esistendo in TaLTaC^2, fra le risorse statistico-linguistiche, due **dizionari di frequenza** in forme grafiche, rispettivamente: i) una lista dell'***italiano standard*** ossia una mistura di differenti generi di scritto e parlato come campione di media dimensione dell'italiano contemporaneo (4 mln. di occorrenze, in seguito detto

corpus *Polif*, [Bolasco, Morrone 1998]); ii) una lista di ***linguaggio comune***, proveniente da un vasto campione di italiano scritto come lessico della stampa (in seguito il corpus è denominato *Rep90*) ricavato da dieci annate del quotidiano "*La Repubblica*" (oltre 230 mln. di occorrenze; vedi dettagli in Bolasco 2005a). A questi si aggiunge un dizionario di 15.000 espressioni poliematiche e poliformi tecnico-specialistici del ***linguaggio economico-finanziario*** (in seguito *Lef*: Canzonetti, 2001), derivato da un ampio campione di testi del settore (4 mln. di occorrenze).

La selezione del **linguaggio rilevante** consente invece di esplicitare le principali **parole chiave**, ossia forme grafiche capaci di discriminare al meglio fra loro le unità di contesto. Per ottenerlo si ricorre all'**indice** TFIDF (Salton, 1989) che pondera i termini in funzione della loro capacità discriminante. In particolare, facendo uso del TFIDF, a partire **da una classificazione non supervisionata** di un corpus di apprendimento è possibile produrre una **categorizzazione fuzzy dei documenti** (Bolasco, Pavone, 2008).

Una ulteriore importante selezione di linguaggio si ottiene dalla ricerca delle ***parole caratteristiche*** che si compie attraverso l'**analisi delle specificità** (Lafon, 1980). Essa consente di evidenziare quali termini sono specifici e caratteristici di una parte dei testi, rispetto ad un'altra, sulla base di un test statistico. Questo calcolo, soprattutto in corpus di dimensioni ridotte, può giovarsi dei processi di lemmatizzazione e/o di categorizzazione semantica messi precedentemente in atto. In tal caso, per calcolare le specificità non sulle singole entrate del vocabolario ma su loro gruppi o tipi è possibile produrre le tabelle per classi (lemmi o altre categorie di interesse), calcolarvi le sub-occorrenze e quindi le specificità. Il caso più comune è quello dei verbi per i quali la riconduzione al lemma consente analisi più efficaci sulla loro presenza nei testi. Vedi un esempio sui verbi peculiari in Bolasco (2005b).

Intrecciando analisi lessicale e testuale, TaLTaC^2 consente di individuare e di formalizzare "**grammatiche locali**", ovvero insiemi di regole proprie di un caso di studio, per poi applicarle al testo attraverso **modelli ibridi** (dizionari + regole) ad esempio per la costruzione di **risorse** statistico-linguistiche (Bolasco, Pavone e D'Avino 2007; Bolasco, Pavone 2010). Questo genere di attività è valorizzato e reso possibile, con interessanti guadagni dei tempi di esecuzione, dalla predisposizione in TaLTaC^2 di **metaliste** e **metaquery**, ovvero set di istruzioni per il lancio in modalità "batch" di un insieme di queries, precedentemente validate in fase di *training* dei modelli.

Nel corso dell'analisi testuale, è possibile indirizzare le **ricerche su sottoinsiemi di testo** o parti del corpus. Si consideri una parte delle unità di contesto, filtrate attraverso le modalità di variabili codificate: ad esempio considerando la variabile "tempo" i soli documenti relativi ad un dato anno. Oppure singole sezioni di tutti i frammenti della collezione: ad esempio in un corpus di articoli scientifici, nei quali ogni documento è diviso in tre sezioni – titolo, abstract, testo dell'articolo – si limita la ricerca ai soli abstract della collezione.

Un'altra funzione di base – classica per l'ADT – è costituita dalle **concordanze** possibili sia per parole che per segmenti. In TaLTaC^2 questa funzione è particolarmente ampia, potendo visualizzare sia i contesti locali di una singola unità lessicale, sia quelli di classi grammaticali, di tratti semantici o di lemmi. È possibile estendere l'ampiezza dei co-testi destro e sinistro, ordinarli in senso crescente/decrescente, nonché esportare il sub-corpus costituito dalle concordanze visualizzate. Quando la collezione di testi è molto numerosa, una concordanza di una parola "tema" (quindi molto frequente), o di un concetto centrale per lo studio, corrisponde ad una quantità di testo notevole, che può giustificare uno studio a sé È il caso di uno studio sulla guerra in Iraq (Giuliano, 2004).

Una particolarità di TaLTaC^2 è quella di poter lavorare "*senza corpus*", ovvero di poter eseguire alcune procedure a partire da **liste o tabelle importate**. È il caso del tagging grammaticale che può operarsi su un vocabolario ottenuto con altri software e importato in TaLTaC^2. Questo è particolarmente utile quando si volesse creare un dizionario di frequenza da inserire nelle proprie risorse di sistema per estrarre il linguaggio peculiare. Infatti è di primaria importanza che le annotazioni grammaticali siano fatte con lo stesso metodo con il quale viene analizzato il corpus oggetto di studio. Anche l'analisi di specificità può applicarsi a tabelle diverse dal vocabolario e quindi anche a tabelle importate, purché contenenti delle sub-occorrenze per parti.

Una prima guida di TaLTaC versione 1.0 è pubblicata da Cisu nel 2000 (Bolasco, Baiocchi e Morrone 2000); successivamente la Guida, a causa dei continui aggiornamenti, viene sviluppata solo *on-line* nel programma, con consultazione ipertestuale e sensibile "al contesto", ossia con apertura diretta sull'argomento trattato a quell'istante dall'analista [5].

Al fine di dare esempi di applicazione di funzionalità di TaLTaC^2, nel

[5] Presente anche sul sito di TaLTaC nell'ultima versione: http://www.taltac.it/it/supporto3.shtml

seguito del quaderno, si considerano due diversi corpus:

– il primo è uno stralcio di recensioni tratte dalla guida dei ristoranti del GamberoRosso Editore: 277 documenti per un totale di 36.000 occorrenze (nel seguito indicato con GRS);

– il secondo è un piccolo corpus che raccoglie i principali discorsi di Obama (nel seguito OBM) nel suo primo anno di attività prima come candidato presidenziale poi come Presidente degli Stati Uniti: 6 discorsi per un totale di 27.450 occorrenze.

1.3. Una strategia per l'analisi del testo

Com'è noto, nella ricerca qualitativa non è possibile individuare un unico percorso d'indagine, tuttavia alcuni passi di una strategia d'analisi assumono un carattere generale e possono essere considerati validi in molti casi.

La strategia tipo – che può essere condotta con TaLTaC^2 – presuppone la creazione di una Sessione di lavoro, ovvero dell'ambiente che conterrà via via tutti i materiali dell'analisi che si va ad effettuare. Tale ambiente comprende in una stessa cartella il file di un database di sessione, una sub-cartella che conterrà i files temporanei di lavoro, ogni tabella/lista esportata o creata a partire dal trattamento in TaLTaC^2.

La fase di inizializzazione del processo è l'acquisizione del Corpus oggetto di studio attraverso la tokenizzazione (*parsing*) del testo, previa definizione del set di caratteri alfabeto/separatori.

Il primo step di analisi consiste nel pre-trattamento del testo attraverso la Normalizzazione, volta ad eliminare le possibili fonti di sdoppiamento del dato. Ad esempio, abbassando le maiuscole non rilevanti, uniformando la grafia di nomi propri, sigle, numeri e date che generalmente comportano una scarsa stabilità. Tali entità, una volta uniformate graficamente, vengono etichettate così da essere recuperabili successivamente. In questa fase di normalizzazione si possono anche individuare espressioni di senso compiuto o locuzioni utili a disambiguare fin dall'inizio occorrenze di "parole di più parole" (*multiwords*) molto frequenti (es. *'dato di fatto'*, *'Unione Europea'*, *'mercato nero'*).

Successivamente è possibile visualizzare ed analizzare il vocabolario del

corpus ed effettuarne alcune misurazioni lessicometriche: calcolo dei parametri della legge di Zipf, misure di ricchezza lessicale, gamme di frequenza e relativo tasso di copertura del vocabolario.

Un secondo step prevede d'individuare sequenze di parole, ovvero ciò che nella letteratura ADT è conosciuto sotto il nome di Segmenti ripetuti (Salem, 1987). Dal loro inventario è possibile visualizzare i segmenti più significativi grazie all'indice IS (Morrone, 1993). In tal modo si può selezionare un insieme di espressioni da lessicalizzare, per trasformare nel testo le sequenze d'interesse in nuove occorrenze, come unità minimali di significato. Ad esempio, nel caso del corpus GRS, *'fiori di zucca'* diventa un'unica "parola" con un contenuto semantico superiore rispetto alle parole semplici che lo compongono.

TaLTaC^2 dispone, come noto, di risorse linguistiche che consentono di annotare le entrate del vocabolario del corpus con etichette (*tagging*) di tipo grammaticale e semantico. Queste meta-informazioni si sfruttano in seguito per selezionare sottoinsiemi o tipi di unità lessicali a fini diversi: dalla ricerca di entità alla compilazione di un report, dalla costruzione di una grammatica locale alla messa a punto di un piano di lavoro.

Lo step del Tagging grammaticale confronta il vocabolario del corpus con il dizionario di TaLTaC^2. In questo modo è possibile etichettare grammaticalmente le forme grafiche non ambigue presenti nel vocabolario. Fra le possibilità di estrazione di classi di parole sono ad esempio estraibili i derivati o alterati: nel GRS sono numerosissime le citazioni di diminutivi e vezzeggiativi nella descrizione di piatti o presentazione di menu: *sformatino*, *lasagnette*, *raviolini*, *scamponi*, *quaglietta*, *caponatina*, *fritturina*. Analogamente estraendo le forme verbali è possibile passare a raggruppare le flessioni di uno stesso lemma per avere una statistica valida dei verbi.

Lo step del Tagging semantico può essere effettuato grazie alla definizione di dizionari tematici predisposti dall'utente anche tramite le funzionalità messe a disposizione da TaLTaC^2 stesso (vedi par. 2.6.2). Nel caso del GRS un primo interesse è riconoscere ogni citazione inerente cibi, piatti o elementi gastronomici dalla preparazione alla cottura. Un dizionario sul web di oltre 8000 entrate, utilizzato a riferimento, consente una prima presa di contatto con questi contenuti. A partire dal quale con ricerche mirate si migliora la cattura di tutte le

unità lessicali del testo [6].

In step successivi, una delle funzioni essenziali di TaLTaC^2 è l'Estrazione di informazione significativa dal vocabolario, in una logica tipica del *Text Mining*. Gli step di analisi per eseguire queste funzionalità sono nel menu Analisi lessicale/Selezione del vocabolario. L'estrazione di sottoinsiemi del vocabolario si ottiene utilizzando o risorse endogene o risorse esogene rispetto al testo in analisi.

Le risorse esogene sono le statistiche di riferimento (lessici di frequenza) contenute nelle Risorse statistico-linguistiche di TaLTaC^2. Confrontando il vocabolario del corpus con il lessico di frequenza più adeguato, è possibile individuare il *Linguaggio peculiare* del testo nel suo complesso, nei termini sia delle unità lessicali sovra/sotto rappresentate (quelle cioè che presentano maggiori o minori scarti d'uso in valore assoluto), che di quelle originali del testo (cioè non presenti nel lessico di riferimento utilizzato. In GRS: *dessert*, *ricotta*, *filetto*, *semifreddo*, *antipasti*, *menu*, *degustazione*, …. Anche i segmenti ripetuti individuati nel corpus possono essere confrontati con un lessico di poliformi, laddove esista un riferimento pertinente, come nel caso di linguaggi specialistici (vedi il *Lef)*.

Le risorse endogene sono invece determinate dalla quantità di frammenti e dalle variabili categoriali che è possibile associare al testo, grazie alle quali è possibile partizionare il corpus, suddividendolo in senso logico. È possibile individuare le parole *specifiche* (*Linguaggio caratteristico*) delle varie parti o sub-testi attraverso l'analisi delle specificità. È il caso di analisi delle risposte aperte in un questionario in cui si vogliano rilevare le differenze tra il linguaggio dei maschi rispetto a quello delle femmine, oppure si vogliano rilevare le caratteristiche dei linguaggi utilizzati a seconda della professione o del titolo di studio. In GRS rispetto alla variabile regione Piemonte/Sicilia si trovano termini come: *agnolotti*, *vitello*, *tajarin*, *formaggi*, *ambiente*, *premuroso*, *bonus / pesce*, *spada*, *ricotta*, *mandorle*, *mare*, *finocchietto*, *pistacchi*, *cassata*, *terrazza*, …. Oltre all'analisi delle specificità, TaLTaC^2 mette a disposizione il calcolo dell'indice TFIDF per l'estrazione delle parole discriminanti (*Linguaggio rilevante*) dai frammenti e per estrarre i frammenti più significativi in termini del linguaggio così individuato. In GRS rispetto ad una query sulla cucina tradizionale di territorio tipica del Piemonte si

[6] Sul corpus del GRS inerente tutte le regioni, il primo tag semantico estrae 3500 elementi dal dizionario web, scaricato dalla pagina http://www.culturagastronomica.it/pagine/le_parole.htm

trovano come parole chiave: *carne*, *tradizione*, *piemontese*, *sugo*, *erbe*, *gnocchi cruda*, *salame*, ….

Al termine delle varie operazioni effettuate negli step finora descritti, è possibile esportare la tabella vocabolario con tutte le annotazioni, da sfruttare nelle successive fasi di analisi del contenuto grazie all'uso delle tecniche statistiche multidimensionali. In genere si tratta di matrici del tipo "parole x testi" dove i termini selezionati sono di volta in volta, sia l'intero vocabolario a soglia di occorrenze, sia differenti "*start list*" di unità lessicali selezionate secondo criteri di interesse (di tipo tematico o sulla base di un "peso" dato alle parole), sia insiemi di lemmi per categoria grammaticale, ad esempio i lemmi dei verbi.

Oltre a tutto questo, l'analisi può proseguire operando del text mining a livello testuale, attraverso la Ricerca di Entità (concetti, strutture linguistiche, locuzioni, entità "nominate") sul testo mediante interrogazioni, ossia query più o meno complesse tradotte in espressioni regolari. Esse consentono di individuare: i) tutti i frammenti che rispondano alla query (ossia che presentino una o più parole, sequenze di parole o quasi-sequenze, nonché alcune relazioni fra tipi, classi o gruppi di parole derivanti dal tagging svolto a livello lessicale), ii) tutte le stringhe di testo trovate con la loro posizione nella collezione, iii) il numero di volte in cui una stessa stringa trovata è presente nel corpus. Nel corpus GRS alla query che voglia estrarre utili riferimenti al territorio – ovvero sostantivi legati a toponimi o simili – corrisponde una espressione regolare del tipo "CATGR(N) LAG2 CATGR(NM)" che estrarrà gruppi nominali quali: *menu di Langa*, *robiola di Roccaverano*, *gnocchetti con Castelmagno*, *coniglio all'Arneis*, *calamaretti saltati al Marsala*, *pistacchi di Bronte*, *bucatini alla Norma*, …; ciascuna di queste espressioni è localizzata nel testo e se ne conteggiano le occorrenze complessivamente nel corpus.

Accanto a questa possibilità di recupero di informazione, TaLTaC^2 offre l'opportunità di creare nuove variabili grazie alle quali categorizzare quei frammenti che soddidisfano la query. Così facendo la tabella dei frammenti presente nel DB di sessione si arricchisce di ulteriori informazioni ricavate direttamente dal testo. Una matrice così popolata può produrre risultati molto più interessanti e precisi in sede di analisi multidimensionale effettuata con software statistici. Si pensi ad un questionario sulla soddisfazione della clientela a cui si può aggiungere una variabile "non soddisfatto" per tutti quei record che presentino, nel testo di una risposta ad una specifica domanda aperta, le espressioni

"non sono soddisfatta" o "mi hanno trattato male" o "scarsa professionalità". La creazione di nuove variabili ex-post frutto dell'analisi testuale, corrisponde alla classica operazione di text mining che consiste nell'estrarre un "concetto" da un testo non strutturato, per trasformarlo in un dato codificato in un database strutturato (processo di ETL), pronto per essere analizzato.

A questo punto, è possibile esportare da TaLTaC^2 anche la matrice denominata [FRxFS] (leggi: frammenti x forme selezionate), frutto dell'analisi testuale [7]. Ad esempio nel caso del corpus GRS come forme selezionate si sceglie il linguaggio peculiare, o quelli caratteristico/rilevante: in tutti i casi poche centinaia di elementi, indipendentemente dalla loro frequenza nel corpus, ma tutte salienti rispetto agli scopi della ricerca [8]. Su questa matrice, è possibile effettuare analisi multidimensionali – con software quali SPAD, SAS, WordMapper o altri – sulle unità originarie della collezione di testi. In questa matrice sono presenti oltre alle parole, anche i concetti, in forma di codifiche nelle variabili testuali create precedentemente, nonché le variabili categoriali, note fin dall'inizio e associate a ciascun record della collezione dei testi. Come è evidente è possibile sviluppare su questa matrice numerose analisi statistiche di correlazione fra parole, concetti e variabili strutturali individuali.

Infine, con TaLTaC^2, è possibile ricostruire il testo originario in forme etichettate, in lemmi o in categorie semantiche oppure ricostruire un corpus "pulito" con la correzione/cancellazione di errori ortografici. Nel caso della ricostruzione del Corpus con lemmi la procedura ricostruisce il corpus originario sostituendo i lemmi alle forme grafiche del testo. In questo caso il lemma è seguito dalla categoria grammaticale. Laddove la categoria non è specificata significa che la parola è grammaticalmente ambigua.

Esempio di testo lemmatizzato: "*signor_presidente_N*, *onorevoli_deputati_N*, *il_DET Governo_N che_ si_ presentare_V oggi_ alle_ Camere_N chiedere_V ….. ".*

[7] Se fosse utile per problemi dimensionali, questa può essere trasposta in [FSxFR]. Così facendo corrisponde a quella che viene chiamata in letteratura: matrice [termini x documenti]. Ciò avviene quando i termini selezionati sono più dei documenti e questi ultimi sono in numero delle centinaia, o di un migliaio.

[8] va detto che le attuali potenze di elaborazione consentono di analizzare l'intero vocabolario o tutte le parole fino a soglia molto bassa; in tal caso la lista di forme selezionate sarebbe indispensabile al momento di visualizzare la parte di risultato che più interessa.

1.4. Gli Ambienti di TaLTaC²

Ogni analisi in TaLTaC² si svolge nell'ambito di una Sessione di lavoro. Operativamente TaLTaC² presenta tre diversi ambienti accessibili dal Menu Visualizza:

– una finestra *Esplora il Corpus* (fig. 1), all'interno della quale è possibile visualizzare il Corpus in modalità *full text*, frammento per frammento e, all'interno di questi, sezione per sezione, operando anche filtri sulle sezioni o sui frammenti. In tale finestra è possibile visualizzare i valori delle variabili a priori (con la possibilità di correggerli) e i valori delle variabili a posteriori, create dall'analisi testuale;

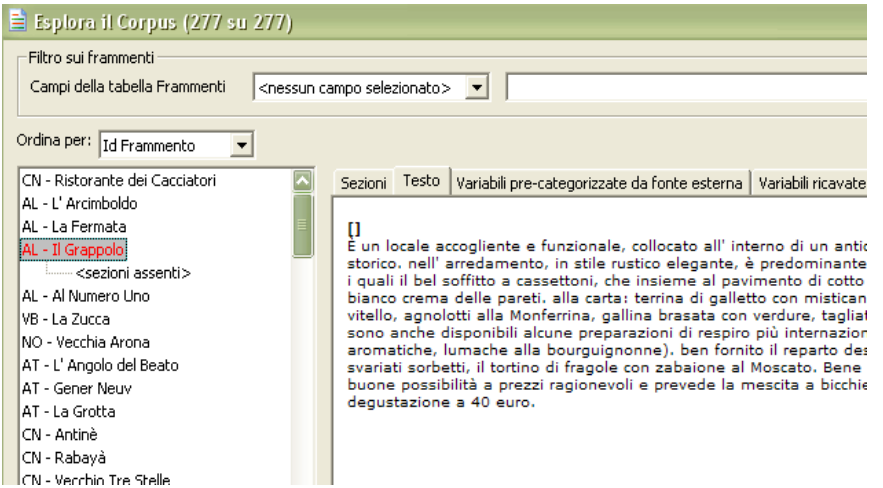

Fig. 1. Visualizzazione dei frammenti con testo e variabili associate in Taltac.

– una finestra *DB di Sessione* (fig. 2) che contiene le varie liste/tabelle prodotte nel corso della propria analisi;

– una finestra *Risorse Statistico-Linguistiche* (fig. 3) che contiene l'insieme delle risorse di riferimento disponibili in TaLTaC², ossia l'insieme delle meta-informazioni (quali lessici di frequenza, liste di frequenza e dizionari tematici) da associare ai dati testuali.

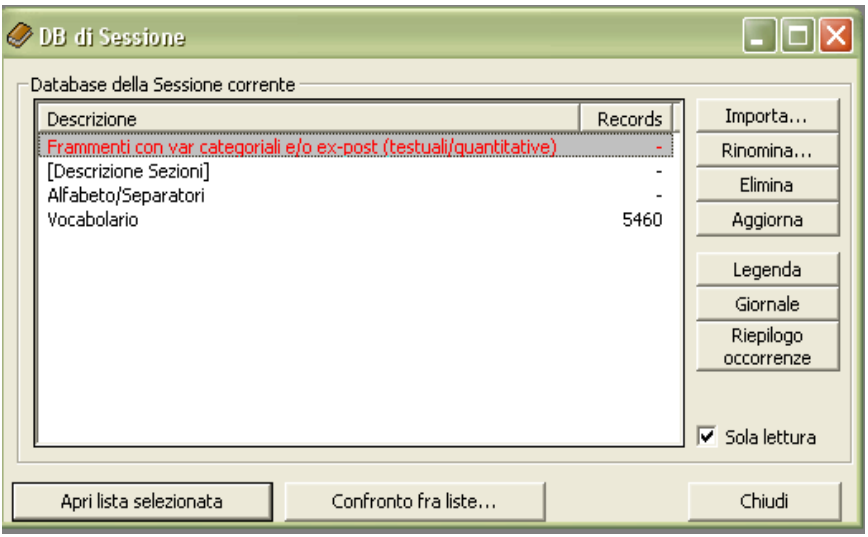

Fig. 2. Finestra del DB di Sessione con le diverse tabelle di lavoro

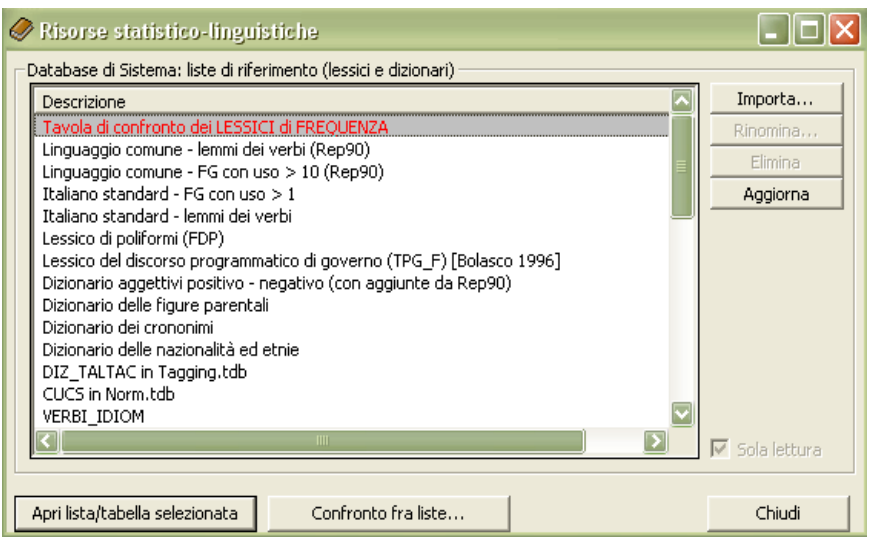

Fig. 3. Finestra delle Risorse disponibili in Taltac

In realtà, in TaLTaC^2 è presente un'altra classe di risorse linguistiche, costituita dai dizionari grammaticali e dalle liste di normalizzazione, che il programma sfrutta in occasione di alcune procedure ma che l'utente non può visualizzare né modificare. In questo ambiente l'utente ha la possibilità di importare anche risorse personali (lessici settoriali, liste di riferimento, per forme grafiche, per lemmi, liste di poliformi ecc.) utilizzabili alla stregua delle risorse predefinite di TaLTaC^2. La differenza sostanziale, nella gestione, sta nel fatto che le risorse personali sono sempre modificabili a piacere dall'utente, mentre quelle predefinite sono protette da modifiche. Gli unici campi di queste tabelle che l'utente può modificare sono i campi di *Informazioni Aggiuntive*.

Il primo ed il secondo ambiente sono legati al lavoro specifico di analisi sui dati testuali e alla diversa composizione del Corpus. Il terzo ambiente è, invece, indipendente dalla sessione e quindi identico ad ogni apertura di TaLTaC^2 e, oltre alle risorse fornite nativamente da TaLTaC^2 all'atto dell'installazione, può contenere anche risorse personalizzate messe a punto o raccolte dall'utente.

Visivamente, il contenuto del DB di Sessione e del DB delle Risorse Statistico-Linguistiche si presenterà in maniera pressoché identica (entrambi sono alimentati con le funzionalità del motore e relative tabelle di MS Access), con la sola differenza che nel secondo non è possibile scrivere nelle tabelle, né esportare records.

1.5. Il modello generale di corpus: collezione dei testi e variabili associate

In TaLTaC^2 sono state ampliate le funzioni di gestione del corpus e delle variabili categoriali associate ai testi. A questo proposito è stato messo a punto un modello generale di corpus che occorre rispettare affinché sia possibile utilizzare appieno le funzionalità del programma.

Questo modello permette l'analisi di *corpora* dalle caratteristiche molto diverse, quali: risposte libere a domande aperte provenienti da survey, raccolte di e-mail o altra messaggistica, focus group, interviste non direttive, documenti più o meno strutturati, articoli di giornale, intere opere di un Autore, basi documentali di tipo giuridico o tecnico-scientifico.

Qualsiasi corpus in quanto collezione di dati testuali, indipendente-

mente dalla sua natura, può essere considerato come un insieme di frammenti, riconducibili ad altrettante unità di rilevazione. Ad esempio se il corpus è costituito da risposte a domande aperte in un questionario, l'insieme delle risposte in testo libero di ogni intervistato costituisce un frammento; se il corpus è costituito da una collezione di documenti, i frammenti possono essere rappresentati dai singoli documenti.

L'attuale versione 2.10 di TaLTaC^2 ottimizza l'integrazione tra variabili testuali (espresse in testo libero) e variabili categoriali, in quanto a ciascun frammento possono essere associate più variabili categoriali. Ad esempio nel caso delle risposte a domande aperte, a ciascun frammento possono essere associate variabili come il sesso, l'età e l'istruzione (in pratica tutte le variabili non testuali rilevate su ogni intervistato). Nel caso di articoli di giornale, a ciascun articolo è associabile l'informazione ad esempio della testata, della data, dell'autore, della tipologia dell'articolo e così via.

I frammenti possono a loro volta essere suddivisi in più sezioni. Ad esempio, nel caso di un questionario con tre domande aperte, per ogni frammento si potranno avere tre sezioni, ciascuna delle quali corrisponde alla risposta fornita dall'intervistato ad una delle tre domande. Nel caso di documenti strutturati, le sezioni possono corrispondere ad altrettanti paragrafi o parti del documento: se il corpus fosse costituito da una raccolta di saggi o articoli scientifici, si potrebbero avere le seguenti sezioni: abstract, testo dell'articolo, testo delle note, riferimenti bibliografici.

Se si tratta di un corpus di interviste non direttive, ciascuna delle quali corposa in durata, ogni frammento della collezione rappresenta la domanda e la risposta (o la sub-risposta come nei frammenti 3, 4 e 5 della fig. 4, in tal caso la domanda non si ripete). Ogni frammento sarà composto di 2 sezioni, rispettivamente il testo della domanda dell'intervistatore, e il testo della risposta dell'intervistato. Una intera intervista è dunque l'insieme di questi passaggi dialogici che sarà ricostruibile grazie ad una variabile codificata di partizione, che somma *ex-post* tutti i frammenti di un intervistato per studi di specificità o quant'altro (vedi schema in fig. 4).

TaLTaC^2 accetta in generale come input, file in formato <.txt>. È comunque possibile indicare come corpus anche file in formato <.doc> o <.rtf> poiché, in tal caso, TaLTaC^2 provvederà automaticamente a convertirli in un formato compatibile, il formato <.*tltcorpus*>.

label Framm	Intervistatore	Griglia della intervista	Rispondente	Var1 Rispond	Var2 Topic	testo Domanda effettiva	testo Risposta
1	Luca	dom1	Federico	m	politica	testo dom1	bla bla .. risposta alla dom1
2	Luca	dom2	Federico	m	famiglia	testo dom2	bla bla .. risposta alla dom2
3	Luca	dom3	Federico	m	religione	testo dom3	bla bla .. risposta alla dom3 -parte1
4	Luca	dom3	Federico	m	famiglia		bla bla .. risposta alla dom3 -parte2
5	Luca	dom3	Federico	m	religione		bla bla .. risposta alla dom3 -parte3
6	Luca	dom1	Maria	f	politica	testo dom1	bla bla .. risposta alla dom1-parte1
7	Luca	dom1	Maria	f	religione		bla bla .. risposta alla dom1-parte2
8	Luca	dom2	Maria	f	famiglia	testo dom2	bla bla .. risposta alla dom2
9	Luca	dom3	Maria	f	religione	testo dom3	bla bla .. risposta alla dom3 -parte1
10	Luca	dom3	Maria	f	famiglia		bla bla .. risposta alla dom3 -parte2
11	Francesca	dom1	Giovanni	m	politica	testo dom1	bla bla .. risposta alla dom1
12	Francesca	dom2	Giovanni	m	famiglia	testo dom2	bla bla .. risposta alla dom2
13	Francesca	dom3	Giovanni	m	religione	testo dom3	bla bla .. risposta alla dom3 -parte1
14	Francesca	dom4	Giovanni	m	famiglia		bla bla .. risposta alla dom3 -parte2

Figura 4. Schema del modello di corpus per interviste non direttive semi-strutturate

Per le specifiche della frammentazione e del sezionamento degli elementi della collezione si rimanda alla Guida online. È utile sapere che si è scelto di adottare un sistema di codifica delle variabili nei frammenti, compatibile con quello usato dal software Alceste e molto vicino a quello del software Spad. Piccole trasformazioni in automatico sulla riga di intestazione del frammento rendono il corpus di Taltac pronto per essere sottoposto a questi due software.

2.
ALCUNI ELEMENTI ESSENZIALI DI TaLTaC [2.10]

In questa sezione si illustrano con maggior dettaglio alcune delle funzionalità già tratteggiate in precedenza con relativi esempi d'applicazione a partire dal corpus GRS.

Attraverso il menu File è possibile in TaLTaC[2] :

§ Creare/aprire la Sessione di lavoro § Consultare la cartella in cui è stata salvata la Sessione
§ Selezionare / assemblare / strutturare la collezione dei testi da analizzare § Eseguire la tokenizzazione (parsing) del Corpus
§ Salvare i record di una tabella nel DB di Sessione § Importare una Lista o un Lessico di Riferimento nel DB di Sessione § Esportare in un file di testo i record di una tabella o una lista di elementi da lessicalizzare § Esportare una matrice Frammenti x Forme o una matrice Forme x Testi § Esportare un SubCorpus § Ricostruire un Corpus annotato
§ Creare/modificare una metalista § Creare/modificare una query predefinita

Procedendo con ordine ci soffermiamo soltanto su alcune di queste funzionalità essenziali cominciando dall'apertura di un nuovo lavoro.

2.1. Predisposizione del lavoro

2.1.1. *Sessione di lavoro*

La creazione della Sessione è la prima operazione da compiere per poter iniziare un'analisi in TaLTaC[2] (nel seguito si usa anche la sigla T2, per semplicità). La Sessione è l'ambiente di lavoro del trattamento dei dati testuali. Solo dopo aver creato una sessione, si può procedere con l'acquisizione di un corpus e l'esecuzione del parsing, oppure con l'importazione di una lista esterna (un vocabolario o un inventario di segmenti ripetuti ottenuti con altro software o una risorsa da rendere disponibile per il programma), e quindi con le operazioni che T2 consente.

Dal Menu File è possibile accedere alla cartella della Sessione in cui si trovano tutti i file generati da T2 nel corso dell'analisi.

2.1.2. *Acquisizione del corpus*

La seconda operazione è scegliere, dal menu **File,** il **Corpus** da acquisire in tre diverse modalità:

Seleziona, quando la collezione dei testi è già raccolta in un solo file nel formato di Taltac (.tltcorpus) o nel formato <.txt>.

Assembla, quando il corpus si ottiene da una collezione di file contenenti ciascuno un singolo documento o frammento. Questo è il caso in cui il corpus provenga dall'assemblaggio di una base documentale residente anche in remoto, rispetto al computer in cui si è creata la sessione di lavoro. Un file di appoggio delle informazioni codificate da associare ai testi, se esiste, funge da selettore della collezione dei documenti da assemblare (fig. 5).

Struttura, quando la collezione dei frammenti è disposta in un data base (o foglio Excel), nel quale i record sono strutturati in campi, separatamente per i dati codificati e per i testi. È il caso ad esempio di dati provenienti da survey sul campo, in cui i frammenti sono gli individui, di cui per ciascun record si conoscono le caratteristiche codificate nel questionario e le risposte libere alle domande aperte. In questo tipo di casistica è possibile far rientrare anche le risposte di una serie di interviste non direttive o il resoconto di un focus group (vedi fig. 4 nel par. 1.5).

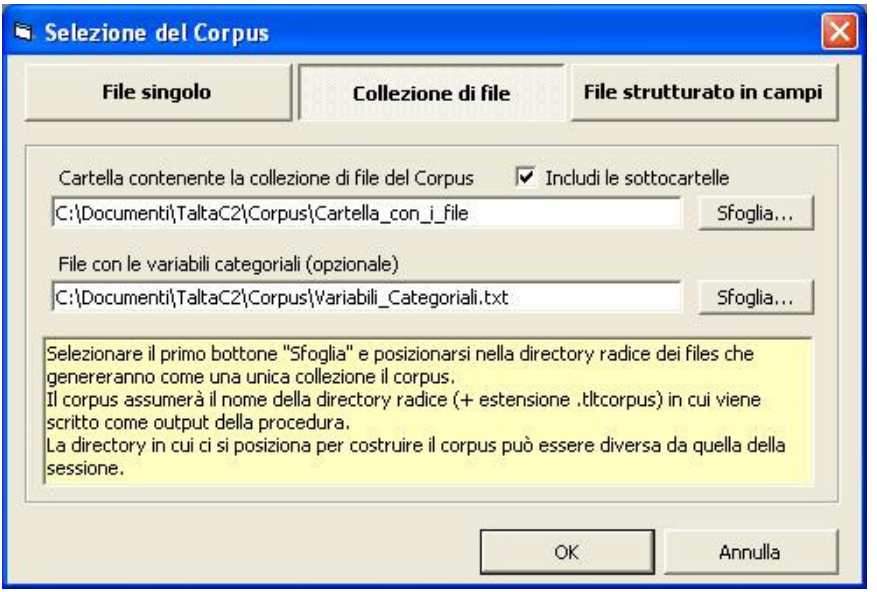

Fig. 5. Acquisizione del corpus da una collezione di file con associazione di variabili categoriali

Una volta che il corpus è nel formato Taltac lo step di ***Parsing*** effettua la tokenizzazione. A tale scopo nella procedura appare una tabella dei caratteri trovati nella collezione dei testi acquisiti e l'analista può scegliere quali caratteri sono da considerarsi come alfabeto, costituenti le parole, e quali come separatori, delimitanti un token dall'altro.

TaLTaC[2] è in grado di leggere un file fornito da **TreeTagger** ed acquisire le informazioni riguardanti categoria grammaticale e lemma. Inoltre, interrogando il proprio database di tagging grammaticale, TaLTaC[2] riesce ad attribuire l'imprinting alle varie forme, completando così il profilo degli attributi grammaticali (solo per testi in lingua italiana). TaLTaC[2] è anche in grado di mantenere l'eventuale struttura di frammenti e sezioni in cui il corpus originario era stato suddiviso, a patto che siano rispettate alcune condizioni sintattiche (vedi Guida).

2.1.3. *Tabelle del DataBase di Sessione*

Al termine del parsing vengono pubblicate nella finestra del DB di Sessione (vedi fig. 2): i) la tabella "Vocabolario" con la lista dei types diversi e relativo numero di occorrenze; ii) la tabella "Frammenti" con la lista

dei documenti inventariati nel corpus e relative informazioni strutturate ad essi associate.

La prima tabella è il luogo delle annotazioni lessicali che saranno effettuate sui types dal punto di vista linguistico, tematico, quantitativo o statistico. La seconda tabella è il luogo delle annotazioni testuali che saranno via via apportate sia come categorizzazioni dei documenti, sia come estrazione di entità di interesse dal testo e relativa creazione di nuove variabili strutturate.

Nel DB di Sessione vi è anche la tabella "Alfabeto/Separatori", che riepiloga la numerosità di ciascun carattere dell'alfabeto e dei separatori, utile per controlli e verifiche sul testo.

2.2. Operazioni di import ed export

2.2.1. *Importa lista*

In T2 è possibile importare liste o tabelle da inserire in una sessione di lavoro già esistente o in una nuova sessione in cui si pensa di lavorare con T2 senza un corpus. Infatti le funzionalità del programma, che riguardano il livello lessicale, operano su tabelle, senza agganci alle occorrenze di un testo. Peraltro in tal modo ad esempio si può lavorare su risorse acquisite altrove, applicandovi gli stessi criteri utilizzati per l'analisi delle informazioni di un corpus.

Ad esempio, se immaginiamo di importare un vocabolario ottenuto da altri software o semplicemente da altre sessioni, è possibile applicarvi il tagging grammaticale riottenendo così delle annotazioni coerenti. È il caso in cui si costruisca un lessico di riferimento di settore a partire da materiali così ampi (ad esempio dal web) le cui dimensioni esulano dalle capacità di T2. Supponete di avere elaborato un corpus di 1Gb con Lexico2 e di avere ottenuto un vocabolario di centinaia di migliaia di entrate in forme grafiche semplici. Importandolo in T2 come tabella esterna (forma, occorrenze, lunghezza) vi si applica il tagging grammaticale per costruire un dizionario di frequenza da importare poi nel DB delle Risorse statistico-linguistiche, al fine di calcolare uno scarto standardizzato per estrarre il linguaggio peculiare di corpus specialistici.

Oppure, se importate un inventario di sequenze anche di milioni di records, potete effettuare operazioni di text mining per estrarvi informazione significativa: è quanto si è fatto per elaborare la risorsa sui

verbi idiomatici (Bolasco, 2010), a partire da 19 milioni di sequenze tratte da Rep90.

Le importazioni sono possibili anche direttamente dal DB di sessione, nel caso in cui la tabella importata serva all'analisi di un corpus: ad esempio per annotare il vocabolario con dei tag semantici. In fig. 6 si riporta l'esempio di importazione di una tabella di lemmi di verbi ai fini dell'analisi di specificità.

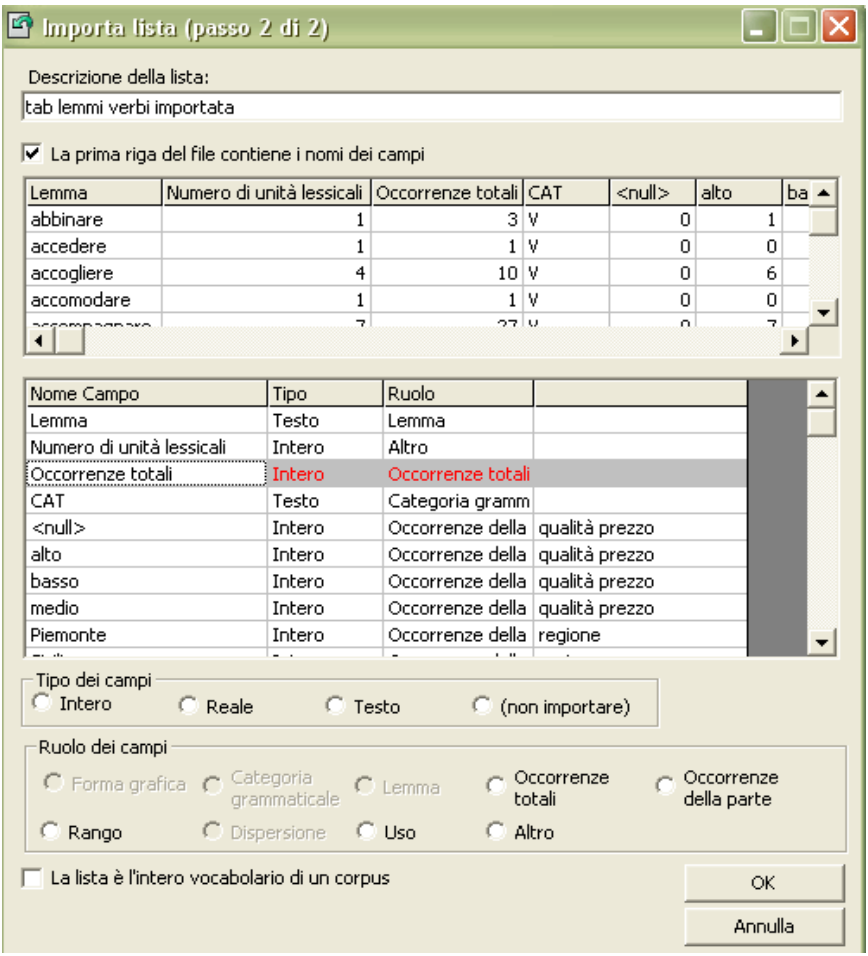

Fig. 6. Maschera di importa lista: esempio da una tabella di lemmi di verbi con sub-occorrenze

Per ogni campo occorre dichiarare il tipo e il ruolo della variabile, che garantisce la compatibilità con i calcoli ai quali è possibile sottoporre l'informazione. Nell'esempio è fondamentale l'attribuzione della variabile di partizione di appartenenza delle sub-occorrenze.

Come output dei lavori in T2, tutte le tabelle del DB di sessione sono esportabili in formato .txt sia nell'insieme dei records, sia in una loro qualsiasi selezione, per utilizzi in altre piattaforme. Si esportano liste di lessicalizzazione, liste di segmenti, di concordanze, di forme annotate semanticamente, di classi di unità lessicali selezionate ecc. In particolare è possibile costruire le matrici di dati testuali già descritte nei par. 1.2 e 1.3, selezionando le unità lessicali secondo "*start list*" predefinite dall'analista, a partire da ogni genere di annotazione operata con T2.

2.2.2. *Ricostruire un corpus*

Le funzionalità di ricostruzione del corpus (fig. 7) sono molto utili, non solo per riprodurre tutto o parte del corpus secondo i filtri a disposizione, ma soprattutto per produrre in modalità diversa dall'originale la collezione di frammenti/documenti su cui si lavora. In particolare, è possibile ricostruire un corpus che originariamente era in formato testo, in:

i) un file strutturato in campi (DB o foglio elettronico), nel quale ogni frammento trova associate nello stesso record le variabili codificate;

ii) una collezione di files in cui ogni documento viene isolato in un file a sé, avente per nome la sua label in T2; viene anche creato in automatico un unico file che raccoglie tutte le codifiche delle variabili associate a ciascun documento. Ciò è utile quando i testi della collezione sono molto numerosi e ampi (più di 32 KB ciascuno) e si prevede di aggiungere successivamente nuovi elementi alla collezione o di fondere più basi documentali.

Ulteriori funzionalità del menu File sono relative a piani di lavoro sulle unità lessicali, di cui si dirà più avanti. Passiamo ora al menu **Analisi** che raccoglie le principali funzioni del trattamento automatico con T2, sviluppate a tre livelli: il pre-trattamento, l'analisi lessicale e l'analisi testuale.

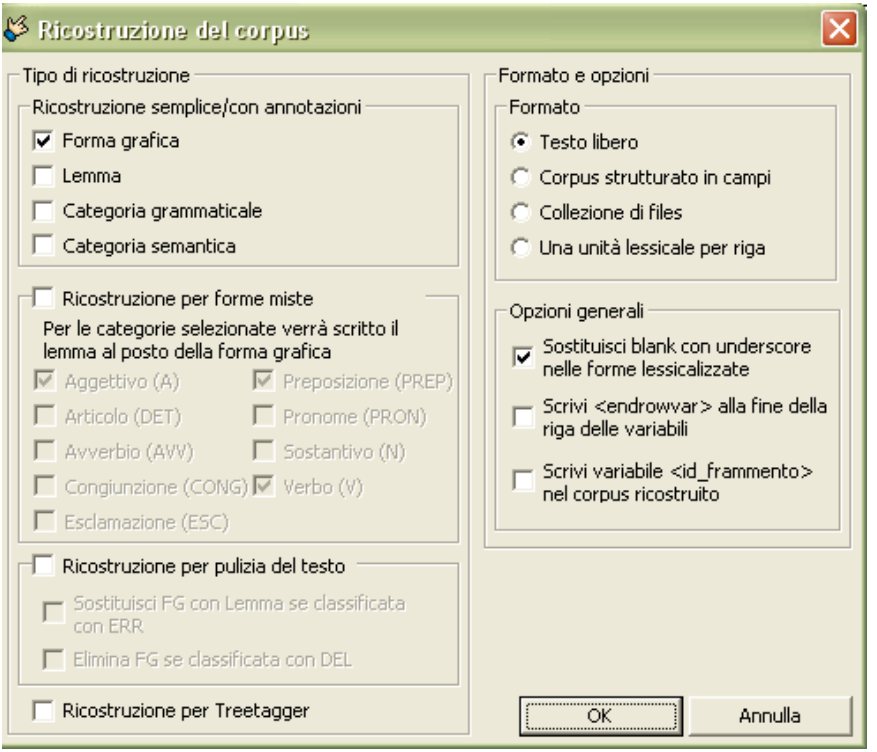

Fig. 7. Menu delle funzioni di ricostruzione del corpus

2.3. Pre-trattamento

Questa fase è particolarmente importante per la buona riuscita delle ricerche. Essa consiste nella gestione, laddove esistano, delle Sezioni con possibilità di un loro riconoscimento automatico, con possibilità di correzioni, al fine di migliorare l'estrazione di informazione al momento dell'analisi testuale. Un secondo step, che riguarda quasi ogni trattamento di un corpus, è la Normalizzazione, di cui si è già in precedenza detto e il calcolo delle Sub-occorrenze secondo le diverse variabili di raggruppamento, che possono essere associate ai dati testuali per ciascun frammento.

2.3.1. *Sezioni*

È solo il caso qui di ricordare che l'analisi per sezioni consente di studiare il corpus, selezionando una parte del testo di ciascun elemento della collezione: isolando ad esempio nelle interviste il linguaggio dell'intervistatore (la domanda) da quello dell'intervistato (la risposta), il testo del titolo di un articolo a stampa dal testo dell'occhiello o del corpo dell'articolo, oppure concentrando le ricerche solo su una parte di un documento semi-strutturato. Si pensi ad uno studio su sentenze in cui ci si concentra sulle "Parti" o sul "Dispositivo" della sentenza.

Questo step permette il riconoscimento *ex-post* delle sezioni, ovvero dopo aver effettuato il parsing. È cioè possibile acquisire un corpus che non presenti gli identificatori di sezioni nei frammenti (le righe che iniziano con ++++), ma che possieda una qualche struttura che si ripete per ogni frammento, ad esempio una suddivisione in paragrafi tutti con lo stesso titolo o quasi, e sfruttare questa suddivisione dei frammenti per creare le sezioni. Questa funzione risulta particolarmente utile nel caso in cui il corpus derivi da una collezione di file, in particolare quando tali file possiedano una struttura comune. Si pensi a documenti scientifici, che presentano sempre un abstract, un'introduzione e poi il contenuto, oppure alle rassegne stampa, in cui gli articoli sono sempre suddivisibili almeno in titolo e testo. Questa funzione evita di dover introdurre manualmente le righe di identificazione delle sezioni direttamente nei file di testo, compito ovviamente molto lungo e oneroso, ma consente di sfruttare la struttura stessa dei file per raggiungere lo scopo.

Si rimanda alla guida per il dettaglio delle funzionalità di questo step.

2.3.2. *Normalizzazione*

In questa fase, come già detto, si eliminano le varianti grafiche non significative e si riconoscono delle entità, prevalentemente "nominate" (nomi propri, personaggi, toponimi, sigle), di interesse generale e potenzialmente fonte di ambiguità (si pensi alla forma *rosa*, che se scritta in maiuscolo identifica un nome o un toponimo, e non un aggettivo). Tali riconoscimenti possono migliorare il livello di studio del contenuto di un testo, ma nel contempo possono non facilitare i confronti con testi non egualmente pre-trattati. Un esempio fra tutti è il confronto del corpus con lessici o dizionari di frequenza se non se ne conosce la

preparazione. Va anche detto che le differenze non incidono profondamente, ma costituiscono solo un fattore di precisione. Queste difficoltà possono essere "superate" o trattando i lessici alla stessa stregua dei corpus (ciò è possibile per le risorse costruite dall'analista) o scegliendo una opportuna gerarchia nelle operazioni. Ad esempio, scegliendo di stravolgere la strategia-tipo proposta nel par. 1.3 e operando un tagging grammaticale e un confronto con un lessico di riferimento "prima" di ogni trasformazione di grafie operata attraverso la normalizzazione.

Quanto alle normalizzazioni personalizzate, occorre dire che il tagging semantico può fare questa funzione per alcune sue caratteristiche intrinseche (vedi par. 2.4.2). Si preferisce questa strada in quanto le condizioni nella sintassi del riconoscimento delle grafie da cambiare non è di facile gestione da parte dell'utente che incorrerebbe in frequenti errori, vanificando il lavoro di uniformazione grafica auspicato da quella funzione.

2.3.3. *Sub-occorrenze*

Il calcolo delle sub-occorrenze è funzionale a conoscere elementi di dettaglio sulla situazione del corpus. Interessa infatti tenere sotto controllo come si distribuiscono le occorrenze totali di ogni unità lessicale contenuta nel vocabolario del corpus. La loro partizione secondo le modalità delle variabili codificate permette di calcolare il linguaggio caratteristico di una parte rispetto all'altra, mediante l'analisi delle specificità, e ciò anche a livello di classi di parole (lemmi, tratti semantici e concetti, classi grammaticali) o dell'intero imprinting del corpus che è la riunione di tutte queste statistiche del testo. Ma le sub-occorrenze sono essenziali a ricostruire le matrici di frequenza per le analisi multidimensionali che ci forniscono la visione d'insieme delle principali relazioni fra le parole, fra i documenti e fra loro raggruppamenti secondo le modalità anzidette.

Vale segnalare alcune particolarità nell'utilizzo di questa funzione. Nel caso di un corpus composto da poche interviste ma in profondità – se ci sono le quantità minime di occorrenze (ordine delle migliaia per ogni intervista) –, si può utilizzare una variabile strumentale per raggruppare i frammenti di una stessa intervista e quindi avere, tramite questa variabile, la possibilità di confrontare le singole interviste (nello schema riportato in fig. 4 sarebbe la variabile associata al nome dell'intervistato). Oppure un secondo caso: quando il corpus è diviso in sezioni,

l'applicazione di questa funzione consente di calcolare le occorrenze delle parole anche per ciascuna sezione, rilevando facilmente le diversità di linguaggio fra sezioni.

2.4. Analisi lessicale

Il dominio dell'analisi lessicale in T2 è la tabella vocabolario. Ogni ricerca e attività di questo tipo insiste su di essa: ad esempio, la selezione di parti del linguaggio, di cui si è già parlato nei paragrafi precedenti. In questo paragrafo si aggiungono informazioni utili non ancora toccate nei par. 1.2 e 1.3, in particolare alcuni dettagli sulle operazioni di tagging.

2.4.1. *Tagging grammaticale*

È appena il caso di ricordare che la procedura di tagging grammaticale in T2 si svolge "fuori contesto", basandosi solo sulla lista delle forme e non sul contesto di ogni frase. Non si effettua quindi una lemmatizzazione completa del corpus, ma ci si limita ad attribuire alle forme non ambigue la categoria grammaticale, il lemma e le caratteristiche morfologiche (imprinting morfematico). Per le forme ambigue TaLTaC^2 fornisce comunque l'elenco di tutte le possibili categorie grammaticali di appartenenza, i lemmi a cui possono essere riferite e le relative caratteristiche dell'imprinting. A partire da un'analisi delle concordanze, il ricercatore può decidere di assegnare una determinata categoria grammaticale alle occorrenze della forma in questione, disambiguandola, quando l'insieme delle occorrenze attualizzate nel corpus sono di una stessa categoria. Questo tipo di etichettatura automatica è quindi esatta ma incompleta e migliora notevolmente (riducendo l'incompletezza) quando si è pre-categorizzato il testo con l'individuazione di entità nominate o *multiword* negli step precedenti. Nel caso dei verbi le occorrenze non ambigue arrivano a coprire facilmente oltre l'80% del totale delle forme verbali, quindi la loro lemmatizzazione senza errori supera questa percentuale e, dal punto di vista statistico, le proporzioni trovate per i lemmi sono stabili, ovvero non cambierebbero di molto quando si arrivasse ad identificare il 100% delle flessioni dei verbi.

TaLTaC^2 consente, in alternativa, di seguire una diversa procedura di tagging grammaticale, "ereditando" una lemmatizzazione completa effettuata all'esterno con altro software. Ciò da un lato risolve i problemi di disambiguazione, ma introduce inevitabilmente un certo tasso di er-

rore nell'attribuzione della categoria grammaticale, di cui occorrerà tener conto. Questo approccio, praticato fra gli altri attraverso l'uso di TreeTagger, implica che la chiave primaria di ogni entrata del vocabolario sia la coppia (FormaGrafica, CAT) e non più soltanto la forma grafica. Ciò comporta varie conseguenze nell'uso degli strumenti di Taltac, di cui tener conto nelle ricerche e nell'estrazione d'informazione.

2.4.2. *TAGGING SEMANTICO*

Attraverso il tagging semantico è possibile associare un'etichetta a tutte quelle forme del Vocabolario legate ad un determinato tema oggetto di studio. In questo modo è possibile categorizzare semanticamente le forme, distinguendole dalle altre. Il tagging semantico può essere eseguito in tre modalità – **Lista da file esterno**, **Liste generate da DB** e **Metalista –**, per ognuna delle quali è necessario avere a disposizione un particolare tipo di risorsa. La prima è un semplice dizionario monotematico, in cui tutte le sue entrate rispondono alla stessa etichetta. La seconda è una lista plurilabels, ovvero un dizionario i cui elementi appartengono a differenti sottoinsiemi o temi, ciascuno con una propria etichetta (ad esempio un dizionario di cibi distinti per tipo di piatti: primi piatti, secondi, contorni, dessert). Questo dizionario deve essere previamente importato come tabella in T2 o nel DB di sessione o nelle Risorse di sistema. Un esempio di quest'ultimo tipo è la risorsa degli aggettivi positivo/negativo per l'analisi del sentiment di un testo, la cui selezione è riportata in fig. 8.

La Metalista invece è una modalità più complessa, poiché lancia un piano di lavoro composito che racchiude in sé vari tipi di risorse, incluse query predefinite, quindi "modelli" di ricerca costruiti sulla base di diversi tipi di interrogazioni, non solo di parole (vedi par. 2.6.2).

In tutti questi tre casi di tagging, le risorse tematiche possono contenere sia parole, sia *multiword*. Nel corpus GRS ad esempio un dizionario di piatti di carne contiene termini come: "*vitello*, *carré d'agnello*, *abbacchio scottadito*, *spezzatino*, *stinco di maiale*, *scaloppine*, …". Prima di questo tagging, il Vocabolario può non contenere le multiword, se queste non sono state già riconosciute nelle fasi di Normalizzazione o Lessicalizzazione, ma ciò non toglie che queste esistano come occorrenze nel corpus. Per questo il tagging semantico comporta una rilettura del testo e di fatto opera un *nuovo parsing del corpus* con la creazione di alcune nuove occorrenze, laddove esistano.

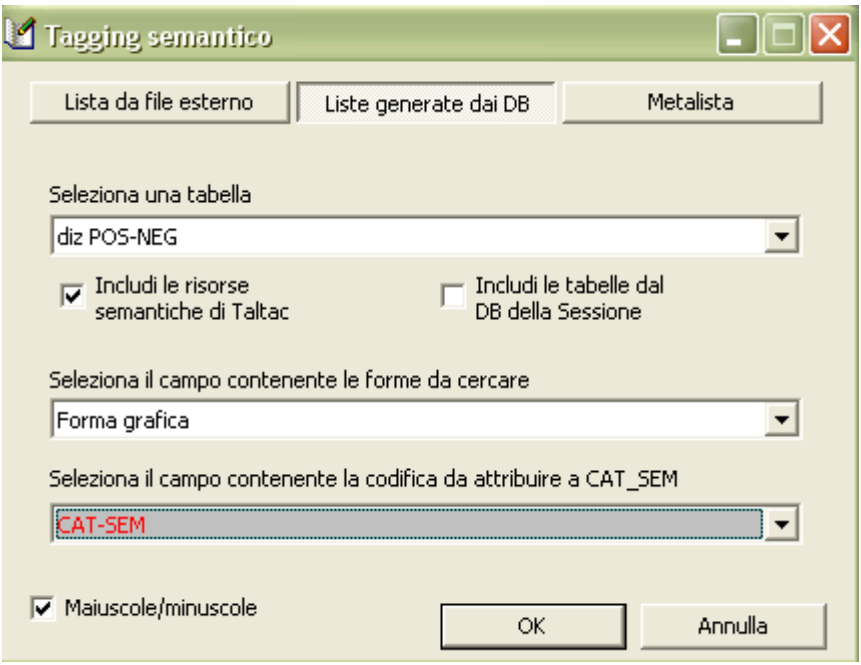

Fig. 8. Maschera del Tagging Semantico da tabella (dizionario positivo/negativo)

Questa operazione in tal modo può fungere da riconoscimento di entità e di normalizzazione delle grafie (se il riconoscimento è sensibile o meno alle Maiuscole). Così facendo, di fatto assolve all'opportunità di sottoporre una normalizzazione personalizzata, ad integrazione di quella standard effettuata nella fase di pre-trattamento.

2.4.3. Query elementari e complesse

Il livello di annotazione lessicale, al di là di quanto finora esposto, può essere personalizzato per gli scopi più diversi. Le funzionalità per il *text/data mining* di base in T2 sono ampiamente descritte nella Guida online. Una volta selezionato un campo di una qualsiasi tabella, con una query elementare si ricerca una parola, o più elementi accomunati da una stessa radice, una sequenza o quant'altro facendo uso di elementi di interrogazione classici, come i caratteri jolly (* ? [ ] !) delle ricerche effettuate con il criterio LIKE (vedi esempio della ricerca *zucc* in fig. 9).

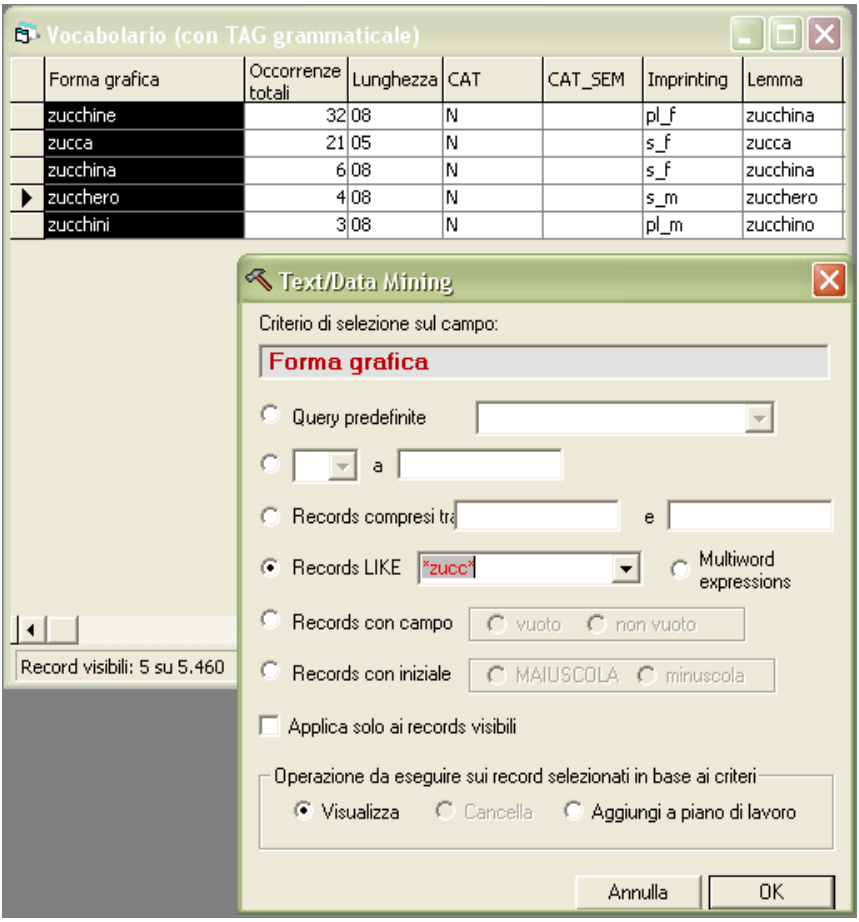

Fig. 9. Funzionalità di text/data mining a livello dell'analisi lessicale: menu di ricerca e sullo sfondo risultato della query impostata nel campo Records LIKE, con visualizzazione dei records estratti dalla tabella Vocabolario.

Fra le funzioni di text/data mining più complesse figurano le cosiddette "query predefinite", ossia set di istruzioni da sottoporre con un solo comando, che riconoscono differenti insiemi di entrate del vocabolario a partire da più query elementari. Una volta estratte le unità ricercate, ciascuna di esse viene annotata con una etichetta nel campo "Informazioni aggiuntive" del vocabolario. Vedi più avanti il par. 2.6.1.

Fin da questo livello di ricerca si possono impostare dei piani di lavoro che permettono di rilanciare una serie di singole query elementari che possono produrre annotazioni in campi predefiniti della tabella vocabolario. Vedere sulla guida il piano di correzione degli errori ortografici.

2.4.4. Sull'analisi delle specificità

In Taltac 2.10 è possibile operare l'analisi delle specificità anche su tabelle diverse da quella Vocabolario purché dotate di sub-occorrenze: ad esempio, sulla tabella dei lemmi dei verbi, così come sull'imprinting (statistica di tutte le categorie). Questa lacuna è stata colmata. Infatti per i verbi le specificità sulle forme flesse sono assai poco significative e per ottenere le specificità sui lemmi occorreva esportare un testo lemmatizzato e risottoporlo in quella forma a T2 prima di effettuare l'estrazione del linguaggio caratteristico. La procedura da seguire, ora permette di calcolare le specificità su qualsiasi tabella con sub-occorrenze; anche su una tabella importata nella quale siano state dichiarate le sub-occorrenze di una variabile (vedi fig. 6 nel par. 2.2). Per procedere al momento dell'analisi delle specificità, si apre, nella versione 2.10, una finestra per scegliere tabella e partizioni (anche più di una nello stesso processo), come illustra la fig. 10.

Fig. 10. Finestra di scelta della tabella e delle partizioni per l'analisi delle specificità

2.5. Analisi testuale

Le principali funzioni di analisi testuale in T2 consistono nel recupero delle concordanze di semplici parole, singole e per categorie (linguistiche o semantiche), e di segmenti; l'estrazione dell'informazione dal testo tramite categorizzazioni da dizionari o da espressioni regolari e tramite parole chiave (calcolo del TFIDF) ai fini della categorizzazione automatica dei documenti. Alcune di queste procedure producono effetti, principalmente, sulla tabella Frammenti nel DB di Sessione. Un doppio click su un record della tabella frammenti rimanda in modo ipertestuale alla visualizzazione del contenuto del testo ad esso relativo, nel quale osservare direttamente le entità trovate, evidenziate in giallo.

Sulle concordanze si è già detto nel par. 1.2. Sul calcolo del TFIDF per i dettagli tecnici si rimanda alla guida online. Qui vale dire, oltre quanto già esposto nel par. 1.3, che l'indice è particolarmente utile per la categorizzazione automatica dei documenti, a partire da query che identifichino dei modelli di classificazione tematica (Bolasco, Pavone 2008).

2.5.1. *Il text mining in TaLTaC²: la ricerca di entità*

La ricerca di entità è la funzione del text mining in T2 a livello testuale. Essa consente, come già detto, di rintracciare nel corpus singole occorrenze di parole, di classi di unità lessicali, di entità più complesse, ma soprattutto di *relazioni fra* tali *occorrenze*. Ciò avviene facendo uso di query scritte attraverso *espressioni regolari*, che mettono a frutto le annotazioni operate al precedente livello di analisi sul Vocabolario del corpus. Le entità di interesse possono essere le più varie: da una semplice categoria grammaticale [i verbi: CATGR(V)] o un singolo lemma [LEMMA(parlare)] ad un gruppo nominale del tipo N_AGG o AGG_N ["CATGR(N) CATGR(A)" OR "CATGR(A) CATGR(N)"]. Oppure ad una "quasi sequenza", in cui un sintagma nominale può recuperarsi anche con un inserto, come nel caso di un <*calamaretti saltati al Marsala*> nell'occorrenza di un <*calamaretti al Marsala*> ritrovabile con l'espressione "CATGR(N) LAG2 CATGR(NM)" (vedi fig. 11), in cui la funzione LAG# implica un ritardo del sostantivo da 0 a # parole dall'aggettivo. Le classi di types compatibili con tali espressioni regolari possono anche provenire da categorie semantiche [CATSEM] o da lessemi [agnell*]. Le espressioni regolari possono essere anche molto complesse e collezionare in una sola espressione decine e decine di relazioni fra gruppi

in OR. È il caso di modelli complessi utili da applicare in blocco su un testo (Bolasco et al. 2007). Questi insiemi possono anche essere pianificati nella cosiddetta metaquery (vedi il par. 2.6.3).

Quando le entità di interesse sono formate da sequenze o da relazioni fra queste è possibile ottenere la lista delle entità ritrovate, la loro localizzazione nei frammenti, la quantità di occorrenze ritrovate per frammento ed è possibile creare una nuova variabile nella tabella frammenti che contenga l'informazione (RifTerrit in fig. 11).

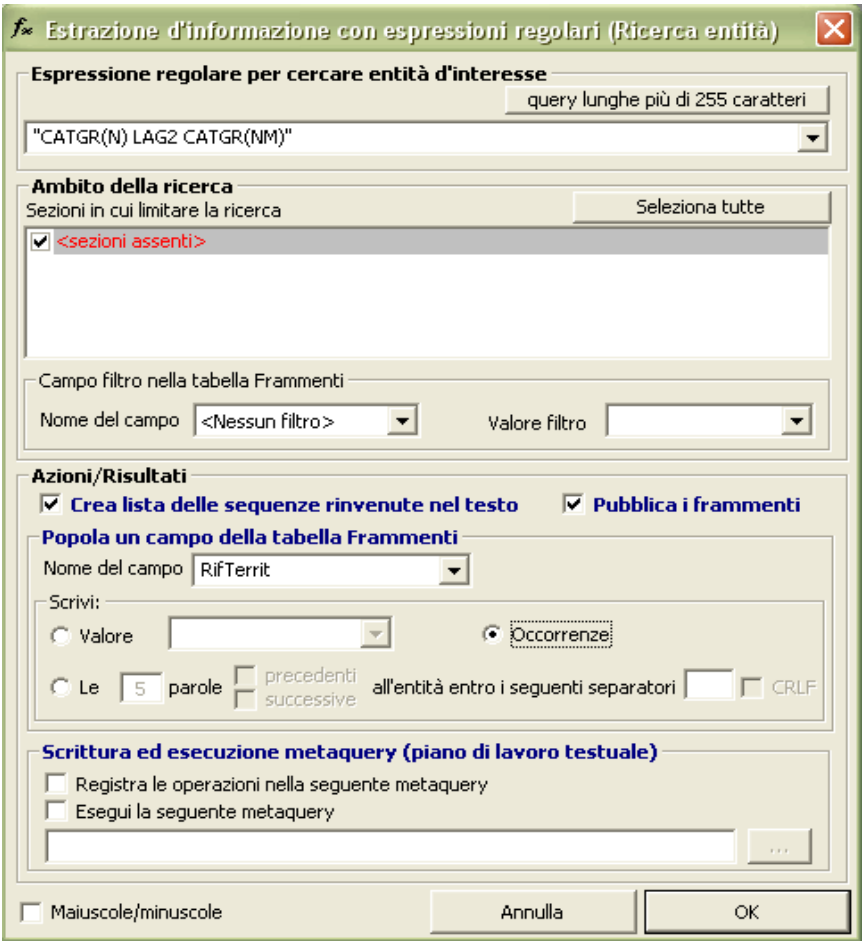

Fig. 11. Funzionalità di text mining per la ricerca di entità mediante espressioni regolari

Quest'ultima è una forma di categorizzazione automatica, utile come processo di tipo ETL per catturare informazione non strutturata e sparsa nel testo e trasformarla in dato strutturato in una tabella.

Le ricerche sul testo possono essere indirizzate solo su alcune delle sezioni di ciascun documento (ad esempio gli abstract in una raccolta di papers scientifici) oppure su una parte dei frammenti, sulla base delle variabili di partizione disponibili (ad es. i papers relativi ad un anno, o ad un autore).

2.6. Strumenti avanzati di ricerca ed estrazione di informazione

2.6.1. *Creazione/modifica di una query predefinita*

Una query predefinita permette di individuare e categorizzare un insieme di forme grafiche (o lemmi), semplici e/o complesse, presenti in una qualsiasi tabella di T2. È costituita semplicemente da una lista di parole/poliformi o loro riduzioni a stem, infissi, suffissi e cosi via (una per ogni riga). Eventuali poliformi presenti nella query devono già esistere nella tabella per essere individuati. Le query predefinite non danno luogo al processo di Lessicalizzazione. Una query predefinita può essere eseguita dalla funzione Text/Data Mining o attraverso l'esecuzione di una Metalista che ne contenga un riferimento.

Le query predefinite sono registrate nel file <Query predefinite.txt> presente nella cartella di installazione di TaLTaC[2]. Tale file può essere personalizzato con l'aggiunta di ulteriori query, seguendo le istruzioni che vi sono riportate all'inizio. Le query predefinite presentano la seguente sintassi:

```
----– INIZIO QUERY –----
Parentela
[CAT] ='N' OR [CAT] ='J' OR [CAT] ='A'
babb?
bisnonn?
cognat?
…
genitor?
[mp] adr?
mamm?
nipot*
…
sorell*
*suocer?
zi?
----– FINE QUERY –-----
```

All'interno dei due marcatori di INIZIO e FINE QUERY, la query inizia con il proprio nome (*Parentela*), segue una riga di vincoli (opzionale) che permette di limitare l'effetto della ricerca a quei record che soddisfano la condizione descritta (nel caso in esame i soli record che abbiano la categoria grammaticale uguale ad N, J o A), ed infine gli elementi della query vera e propria. Applicata ad una tabella, la query dell'esempio selezionerà tutte le parole come *babbo*, *babbi*, *bisnonno*, *bisnonna*, *bisnonni* ecc. che abbiano, come categoria grammaticale, uno dei tre valori espressi nella riga dei vincoli (N o J o A). Gli stessi record verranno contrassegnati, nel campo Informazioni aggiuntive, tramite il nome della query (*Parentela*), al fine di permettere una rapida identificazione o recupero nel seguito del lavoro.

In fig. 12, la maschera per inserire elementi in una query predefinita, per modificarla e salvarla

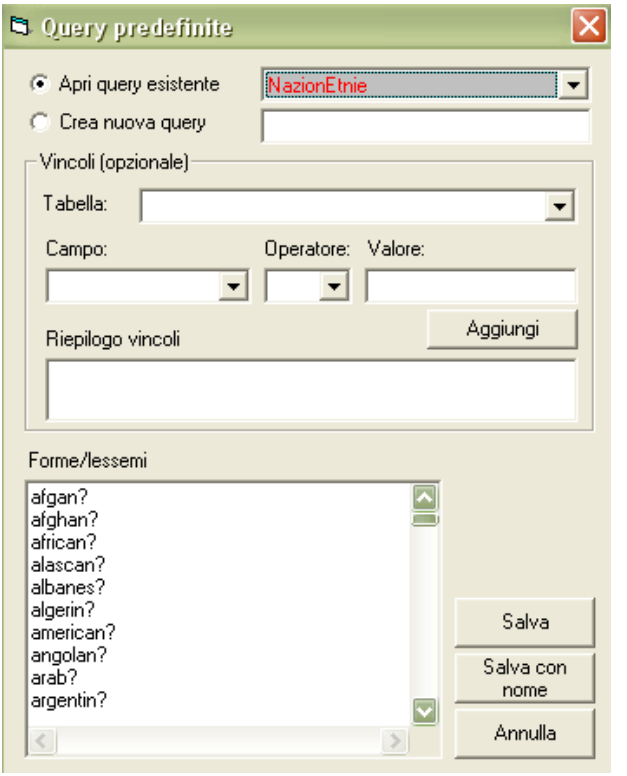

Fig. 12. Maschera di inserimento di una query predefinita

2.6.2. *META-LISTA*

Talvolta si ha interesse a sottoporre una serie di query anche complesse, secondo un "piano di lavoro", che ricostruisca le annotazioni necessarie alla costruzione di un modello di risorsa, precedentemente "verificato". Ciò è possibile attraverso il livello più sofisticato di ricerca lessicale: la metalista (fig. 13). Essa consente di unificare in un unico processo le varie modalità di categorizzazione semantica (Tagging semantico e Query predefinite) disponibili in TaLTaC[2]. Una metalista è formata da più elementi quali:

- liste esterne (in formato .txt);
- tabelle del DB di sessione o delle Risorse Statistico Linguistiche;
- query predefinite (vedi par. 2.6.1);
- annotazioni già attribuite (ad es. categorie grammaticali o semantiche);
- singole forme/lessemi.

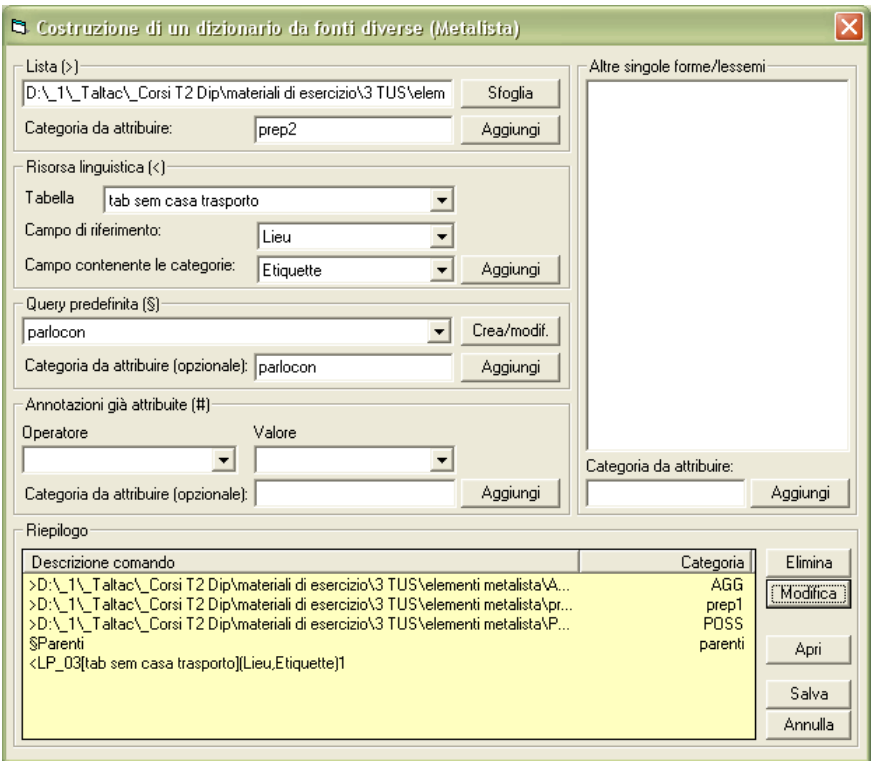

Fig. 13. Menu di compilazione di una metalista per la costruzione della risorsa TUS di Locuzioni di luogo

Tramite la funzione del menu **File/Crea/Modifica metalista** si apre la finestra illustrata in fig. 13 che consente di comporre *ex-novo* una metalista o di modificarne una esistente.

Non è necessario inserire nella metalista un elemento di ogni tipo; al contrario è possibile inserire più elementi di un medesimo tipo (più liste esterne, più query predefinite ecc.). Una volta terminata la sua composizione, e salvata in un file .txt, la metalista potrà essere riutilizzata in qualsiasi momento attraverso la funzione Tagging semantico.

2.6.3. *META-QUERY*

Analogamente a quanto visto per le procedure batch del livello lessicale (piani di lavoro, query predefinite, metaliste), anche nell'analisi testuale è possibile – nel corso delle attività di training di un modello – accumulare via via query ben verificate dagli esiti assai positivi (ad alta precisione), per poi rilanciarle all'interno di un set di operazioni mediante un solo comando, oppure riapplicarle in corpus diversi ma omologhi al primo. Questi piani di lavoro testuali sono chiamati in T2 *metaquery*, in quanto collezionano una serie di query anche complesse, applicate al corpus attraverso il menu della ricerca di entità. Tali piani memorizzano tutti i parametri di questo menu: dominio di applicazione della ricerca (ovvero in quali sezioni insiste la ricerca?), eventuali filtri sui documenti, lista delle entità trovate, creazione della variabile ex-post nella tabella frammenti, conteggio di occorrenze ecc. La metaquery viene memorizzata in un file .txt apribile in Excel che può essere direttamente alimentato con nuove istruzioni in quella sede, pronto all'esecuzione (vedi fig. 14).

Una metaquery si giustifica tanto più quanto più ampia è l'attività da ripetere e tanto più lunga è stata la fase di training delle query. Solitamente ciò avviene nella costruzione di modeli di grammatiche locali (vedi Bolasco et al. 2007; Bolasco, Pavone 2010) che fanno uso di sistemi ibridi. Questi ultimi integrano funzionalità dell'analisi lessicale – basate sia su regole (query) che su dizionari applicati al Vocabolario – con funzionalità di ricerca nel corpus; ossia ricerca di relazioni fra classi, concetti o gruppi di parole, query testuali per la ricerca di strutture linguistiche e non.

Una metaquery in genere "eredita" almeno una metalista, in quanto è un piano di lavoro testuale che "lancia" un set di espressioni regolari

fondate su tutta una serie di annotazioni del Vocabolario, effettuate quasi sempre tramite una metalista. Tutto ciò rappresenta una cosiddetta "soluzione di text mining", che struttura molte nuove informazioni e che è in grado di essere riapplicata senza ulteriori costi, ad esempio per una categorizzazione automatica di documenti o per la costruzione di una nuova risorsa.

Espressione regolare	Crea lista entità	Pubblica i frammenti	Campo da creare nella tabella Frammenti	Scrivi occorrenze	Scrivi valore	Scrivi le n parole successive all'incipit	Valore	Numero parole rispetto all'incipit	Precedenti	Successive	Caratteri di stop	CRLF	Campo filtro	Valore filtro	Maiuscole/minuscole	Sezioni
CATSEM(PREP1) CATSEM(casa) OR "CATSEM(PREP1) CATSEM(POSS) CATSEM(casa)"	0	1		0	-1	0		5	0	0		0	<nessun filtro>		0	0
CATSEM(PREP1) CATSEM(AGG) CATSEM(casa) OR "CATSEM(PREP1) CATSEM(POSS) CATSEM(AGG) CATSEM(casa)" OR "CATSEM(PREP1) CATSEM(casa) CATSEM(POSS)"	0	1		0	-1	0		5	0	0		0	<nessun filtro>		0	0

Fig. 14. Tabella excel della metaquery

2.7. PER CONCLUDERE

In chiusura, vale la pena indicare che da tempo in T2 esiste nel DB di Sessione una tabella che descrive il giornale di bordo dei lavori svolti (fig. 15). Il giornale elenca analiticamente le operazioni fatte con i valori dei parametri utilizzati e pertanto garantisce il principio della riproducibilità della prova, molto importante in un ambito come questo che è comunque prevalentemente di ricerca qualitativa.

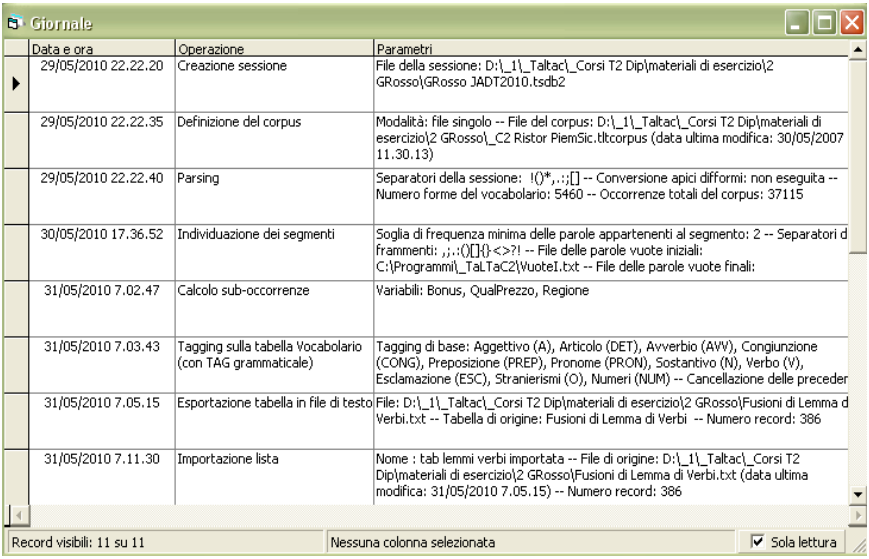

Fig. 15. Stralcio di giornale di bordo di una sessione

Come detto nell'introduzione, con questo contributo si apre una serie di Quaderni intesi ad ospitare esperienze applicative o sviluppi metodologici, che la piattaforma TaLTaC consente. Le potenzialità di lavoro grazie a TaLTaC sono vastissime. In particolare a livello di "costruzione di modelli" per l'analisi del testo, laddove il ricercatore sappia integrare *risorse linguistiche* importanti con *regole* "specifiche" o *grammatiche locali*. Una prova recente è l'esercizio fatto sul corpus di Repubblica (*Rep90*) partendo dalla risorsa delle locuzioni verbali del Gradit (De Mauro 2007). La grammatica locale, messa a punto in quel caso (Bolasco 2010), ha dato luogo ad un algoritmo, ora presente in TaLTaC, per riconoscere in qualsiasi testo i *verbi idiomatici* che, nel loro insieme in un corpus, forniscono una vista significativa sull'imprinting del discorso. Ad esempio nel corpus "Obama" (OBM), si estraggono, dai discorsi del suo primo anno in carica, verbi quali: *porre fine, dare vita, andare avanti, portare avanti, dare il via, prestare servizio, farsi carico, porre rimedio, avere nel cuore, fare parte*, diversi da quelli di altri uomini politici.

3.
RIFERIMENTI BIBLIOGRAFICI

3.1. BIBLIOGRAFIA INTORNO A TALTAC E JADT

Aureli E., Bolasco S. (a cura di) (2004). *Applicazioni di analisi statistica di dati testuali*, Casa Editrice Università "La Sapienza", Roma, pp. 181.

Baiocchi F., Bolasco S., Canzonetti A., Capo F.M. (2005). Estrazione di informazione da testi per la classificazione automatica di una base documentale: la soluzione di Text Mining per l'Authority della Concorrenza, in S. Bolasco, et al. *Text mining: uno strumento strategico per imprese e istituzioni*, CISU, Roma, 2005, pp. 45-54.

Balbi S., Bolasco S., Verde R. (2002b). Text Mining on Elementary Forms in Complex Lexical Structures in A. Morin, P. Sébillot (eds.) *JADT 2002*, St Malo 13-15 marzo, IRISA-INRIA, pp. 89-100.

Bolasco, S. (1990). "Sur différentes stratégies dans une analyse des formes textuelles: une experimentation à partir de données d'enquête", in M. Bécue, L. Lebart, N. Rajadell (eds.), *JADT 1990*, Barcellona, pp. 69-88.

Bolasco, S. (1992). "Criteri di lemmatizzazione per l'individuazione di coordinate semantiche", relazione al convegno "Ricerca Qualitativa e Computer nelle Scienze Sociali", Dipartimento di Sociologia, Università di Roma "La Sapienza" – 1-2 dicembre 1992, in Cipriani R., Bolasco S. (a cura di), (1995). *Ricerca qualitativa e computer*, Franco Angeli, Milano, pp. 87-111.

Bolasco, S. (1993). Choix de lemmatisation en vue de reconstructions syntagmatiques du texte par l'analyse des correspondances, in *Proceedings of JADT 1993*, pp. 399-410, ENST-Telecom, Paris.

Bolasco, S., Lebart, L., Salem, A. (eds.) (1995). *JADT 1995 – Analisi statistica dei dati testuali*, CISU, Roma, tome1, pp. 409; tome 2, pp. 410.

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

Bolasco, S. (1998). Meta-data and Strategies of Textual Data Analysis: Problems and Instruments, in Hayashi *et al.* (eds.) *Data Science*, *Classification and Related Methods*, Springer Verlag, Tokio, pp. 468-479.

Bolasco, S., Morrone, A. (1998). La construction d'un lexique fondamental de polyformes selon leur usage, in S. Mellet (ed.), *JADT 1998*, Univ. Sophie Antipolis, Nice, pp. 155-166.

Bolasco S., Morrone A., Baiocchi F. (1999). A Paradigmatic Path for Statistical Content Analysis Using an Integrated Package of Textual Data Treatment, in M. Vichi, O. Opitz (eds.), *Classification and Data Analysis. Theory and Application*, Springer-Verlag, Heidelberg, pp. 237-246.

Bolasco, S. (1999). *L'analisi multidimensionale dei dati.* Carocci, Roma, pp. 358. (4^ ristampa 2010).

Bolasco S. (2000a). TALTAC: un environnement pour l'exploitation de ressources statistiques et linguistiques dans l'analyse textuelle. Un exemple d'application au discours politique, in *JADT2000*, EPFL, Lausanne, tome 2, pp. 342-353.

Bolasco S. (2000b). Déclarations et répliques gouvernementales dans le discours parlementaire italien, deux genres discoursifs. *Mots*, 64, pp. 97-112.

Bolasco S., Baiocchi F., Morrone A. (2000). *TALTAC. Versione 1.0 – Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto*, Cisu, Roma, pp. 80.

Bolasco S. (2001a). Analisi di interviste aperte mediante metodi testuali. Un'indagine sulla Sardegna. in A. Tuzzi (a cura di) *Dall'intervista alla notizia*, Edizioni Sapere, Padova, pp. 89-110.

Bolasco S. (2001b). Statistiche sulla partecipazione nel sito FO e analisi testuale dei messaggi, in M. Radiciotti (a cura di) *La formazione on-line dei docenti Funzioni Obiettivo. Indagine qualitativa sugli esiti dei forum attivati dalla Biblioteca di Documentazione Pedagogica*. Franco Angeli, Milano, pp. 19-78.

Bolasco S., Verde R., Balbi S. (2002a). Outils de Text Mining pour l'analyse de structures lexicales à éléments variables, in A. Morin, P. Sébillot (eds.) *JADT 2002*, St Malo 13-15 marzo, IRISA-INRIA, pp. 197-208.

Bolasco S. (2002). Integrazione statistico-linguistica nell'analisi del contenuto, in B. Mazzara (a cura di) *Metodi qualitativi in Psicologia Sociale. Prospettive teoriche e strumenti operativi*, Carocci, Roma, pp. 329-342.

Bolasco S., Giovannini D. (2002). Il Trattamento Automatico Lessico-Testuale per l'Analisi del Contenuto (TALTAC): un'applicazione allo studio delle immagini della formazione professionale trentina, in B. Mazzara (ed.), *Metodi qualitativi in Psicologia Sociale. Prospettive teoriche e strumenti operativi*, Carocci, Roma, pp. 343-361.

Bolasco S. (2004a). L'analisi statistica dei dati testuali: intrecci problematici e prospettive in E. Aureli, S. Bolasco (eds.), *Applicazioni di analisi statistica di dati testuali*,

Casa Editrice Università "La Sapienza", Roma, pp. 9-26.

Bolasco S. (2004b). Il linguaggio dei protagonisti: una analisi lessicale e testuale in R. M. Morani, M. C. Salustri, E. Tais (eds.) *Sapere i Sapori – Comunicazione ed Educazione alimentare*, Anicia Editore, Roma, pp. 109-134.

** Bolasco S. (2004c). "Il Text Mining in banca: una nuova sfida per semplificare il flusso di contatti con il cliente", Workshop su Data Mining e Text Mining, in atti del 3^ Convegno ABI "CRM 2004 – Fidelizzare la Clientela Privata e lo Small Business", Roma, Bancaria Editrice.

Bolasco S., Baiocchi F., Canzonetti A., Della Ratta F., Feldman A. (2004a). Applications, sectors and strategies of Text Mining, a first overall picture in S. Sirmakessis (ed.), *Text Mining and Its applications*, Springer Verlag, Heidelberg, pp. 37-52.

Bolasco S., Bisceglia B., Baiocchi F. (2004b). Estrazione di informazione dai testi in *Mondo Digitale*, III, **1**, 2004, pp. 27-43.

** Bolasco S., Bolasco M. (2004). Il gusto delle parole: il lessico della critica enogastronomica. relazione al Convegno "*Comunicare il Gusto*", Di p. di Sociologia e della Comunicazione, Università di Roma "La Sapienza", 19 aprile 2004.

Bolasco S., DellaRatta Rinaldi F. (2004). Experiments on semantic categorisation of texts: analysis of positive and negative dimension. in G. Purnelle, C. Fairon, A. Dister (eds.), *JADT2004 Le Poids des mots*, UCL Presses Universitaires de Louvain, vol. 1, pp. 202-210.

Bolasco S. (2005a). La reperibilità statistica di tendenze diacroniche nell'uso delle parole, in I. Chiari e T. DeMauro (eds.) *Parole e Numeri. Analisi quantitativa dei fatti di lingua*, Aracne, Roma, pp. 335-354.

Bolasco S. (2005b). Statistica testuale e text mining: alcuni paradigmi applicativi, *Quaderni di Statistica*, Liguori Ed., 7, pp. 17-53.

Bolasco S., Canzonetti A. (2005). Some insights into the evolution of 1990s' standard Italian using Text Mining techniques and automatic categorisation, in M. Vichi, P. Monari, S. Mignani e A. Montanari (eds.) *New developments in classification and data analysis* (Serie *Studies in Classification*, *Data Analysis*, *and Knowledge Organization*), Springer-Verlag, Berlin, pp. 293-302.

Bolasco S., Canzonetti A., Capo F.M. (2005b). *Text mining: uno strumento strategico per imprese e istituzioni*, CISU, Roma, pp. 202.

Bolasco S., Canzonetti A., Capo F.M., della Ratta-Rinaldi F., Singh B. K. (2005c). Understanding Text Mining: a Pragmatic Approach, in S. Sirmakessis (ed.), *Knowledge Mining*, (Series: Studies in Fuzziness and Soft Computing, Springer Verlag), Springer-Verlag, Heidelberg, pp. 31-51.

Bolasco S., Galli de' Paratesi N., Giuliano L. (2006). *Parole in libertà. Analisi statistica e linguistica dei discorsi di Berlusconi*, ManifestoLibri, Roma, pp. 142.

Bolasco S., D'Avino E., Pavone P. (2007). Analisi dei diari giornalieri con strumenti di

statistica testuale e text mining, in M.C. Romano (a cura di), *I tempi della vita quotidiana. Un approccio multidisciplinare all'analisi dell'uso del tempo,* Roma, ISTAT, pp. 309-340. [1]

Bolasco S. (2008a). Corpora e liste di frequenza d'uso: criteri e tecniche per l'analisi automatica dei testi, in M. Barni, D. Troncarelli e C. Bagna (eds.) *Lessico e apprendimenti. La dimensione lessicale nell'educazione linguistica*, F. Angeli, Milano, pp. 113-142.

** Bolasco S. (2008b). "Dal lessico delle guide, i tipi della ristorazione in Italia", relazione al convegno sul tema: "Gli attori della cucina italiana" – ALMA Graduate School, Università di Bologna, 6 maggio 2008.

Bolasco S., Pavone P. (2008). Multi-class categorization based on cluster analysis and TFIDF, in S. Heiden & B. Pincemin (eds.) *} JADT2008*, Presses Universitaires de Lyon, vol. 1, pp. 209-218.

Bolasco S. (2010). Il riconoscimento automatico di locuzioni verbali con l'ausilio del software Taltac[2], *Rassegna Italiana di Linguistica Applicata*, **1**: 39-56.

Bolasco S., Chiari I., Giuliano L. (eds.) (2010). *JADT2010. Statistical Analysis of Textual Data*, Proceedings of 10th International Conference JADT, LED, Milano, 2 voll., pp. 1330.

Bolasco S., Pavone P. (2010). Automatic Dictionary and Rule-Based Systems for Extracting Information from Text, in F. Palumbo, C. o N. Lauro, M. Greenacre (eds.) *Data Analysis and Classification*, Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società Italiana di Statistica, Springer, Berlin-Heidelberg. pp. 189-198.

Canzonetti A. (2001). "Il lessico di frequenza del linguaggio economico finanziario: differenza tra old e new economy", Tesi di laurea, Università degli studi di Roma "La Sapienza", Facoltà di Economia.

Cipriani R., Bolasco S. (a cura di), (1995). *Ricerca qualitativa e computer*, Franco Angeli, Milano, pp. 443.

Giuliano L., La Rocca G. (2008). *L'analisi automatica e semi-automatica dei dati testuali. Software e istruzioni per l'uso*, LED, Milano, pp. 247.

Morrone A. (1993). Alcuni criteri di valutazione della significatività dei segmenti ripetuti, in Aa.Vv., *Secondes journées internationales d'analyse statistique de données textuelles*, Telecom-Enst, Paris, pp. 299-309.

[1] Pubblicato sul sito Istat (http://www.istat.it/dati/catalogo/20070807_00/) e tradotto anche in inglese (http://www.istat.it/dati/catalogo/2008061201/arg0835time_use_in_daily_life.pdf).

3.2. Altri riferimenti bibliografici

De Mauro T. (1999, 2003, 2007). *Grande dizionario di italiano dell'uso*, Utet, Torino.

Giuliano L. (2004). Il lessico della guerra nei newsgroups della categoria it. politica durante la guerra in Iraq, in G. Purnelle, C. Fairon, A. Dister (eds.), *JADT2004 Le Poids des mots*, UCL Presses Universitaires de Louvain, vol. 1, pp. 504-514.

Lafon P. (1980). Sur la variabilité de la fréquence des formes dans un corpus. Mots, 1, pp. 127-165.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*, Klincksieck, Paris.

Salton G. (1989). *Automatic Text Processing: The Transformation*, *Analysis and Retrieval of Information by Computer*, Addison-Wesley.