Dalla Valsusa in avanti: i corpora di stampa periodica locale

Manuel Barbera & Cristina Onesti*

Università di Torino

Abstract

The present paper aims at illustrating the construction of a corpus of regional newspapers, with a particular focus on the Segusinum corpus, a tool developed at the University of Turin. The main objective of the project is to fill the gap of local newspaper corpora in Italian linguistics and, more specifically, to provide a methodological solution not only for regional sources but also for corpora of journalistic language in general.

1. Corpora giornalistici nazionali e locali

Le raccolte di testi giornalistici a tiratura nazionale mostrano una tradizione piuttosto recente che ha dato vita tuttavia a una vivace panoramica di risultati, pur eterogenei. Di quasi ogni quotidiano italiano è ormai possibile acquisire un dvd o navigare un archivio online, strumenti che mettono a disposizione materiale dal notevole valore storico, sociologico, linguistico. Ma altro sono propriamente i corpora:

> raccolte di testi (scritti, orali o multimediali) o parti di essi in numero finito in formato elettronico trattati in modo uniforme (ossia tokenizzati ed addizionati di markup adeguato) così da essere gestibili ed interrogabili informaticamente; se (come

* Per gli scopi rituali si attribuiscano a Cristina Onesti i §§ 2, 4 e 5, a Manuel Barbera il § 3, e ad entrambi il § 1.

spesso) le finalità sono linguistiche (descrizione di lingue naturali o loro varietà), i testi sono perlopiù scelti in modo da essere autentici e rappresentativi,

secondo come li definivamo in Barbera, Corino & Onesti 2007b: 70 e Barbera 2009: 126. Lo scenario degli strumenti specifici di questo tipo, dal taglio prettamente linguistico, per utenti interessati al vero e proprio versante della *corpus linguistics*, risulta infatti piuttosto circoscritto: il corpus *La Repubblica*[1] sviluppato dall'Università di Bologna (cfr. Baroni *et al.* 2004: 1771-1774) ci sembra uno dei pochi strumenti da considerare efficace per un linguista.

Ancor più limitata è la presenza di corpora giornalistici a livello locale. La stampa periodica milanese ebbe sì una propria raccolta di concordanze per l'Ottocento (Bonomi *et al.* 1983), ben prima tuttavia della possibilità di interrogazione da formato digitale attualmente possibile. Nonostante l'esistenza di strumenti quali la *Biblioteca Digitale Italiana* (BDI)[2], che raccoglie documenti scannerizzati da pubblicazioni a tiratura regionale o comunque minoritaria rispetto a quella nazionale (bollettini, annuari, ecc.), nonché la recente ed encomiabile esperienza regionale dell'archivio della stampa periodica piemontese (Periodici del Piemonte e della Valle d'Aosta[3], di carattere storico, comprendente schede, titoli bibliografici ed immagini), solo un corpus, per sua stessa definizione (cfr. *supra*), permette un'analisi linguistica – ardua con formati *.pdf* o immagini.

Il nostro gruppo di ricerca, nato intorno a bmanuel.org ed alla distribuzione di corpora.unito.it, ha voluto pertanto fare i conti con tale lacuna, a partire dagli studi resi possibili in seno al progetto FIRB *L'italiano nella varietà dei testi. L'incidenza della variazione diacronica, testuale e diafasica nell'annotazione e interrogazione di corpora generali e settoriali*, RBAU014XCF 2001, coordinatore Carla Marello, e proseguendo con la realizzazione di un corpus di testi giornalistici piemontesi all'interno del portale bmanuel.org. Con questo strumento si potrà accedere a tradizionali ricerche per lemma e calcoli di frequenze delle occorrenze; a singole parti del discorso grazie al *POS-tagging*; a indagini morfologiche, sintattiche e testuali; a ricerche specifiche su titoli, sottotitoli e occhielli; a ricerche specifiche nelle civette di prima pagina (e diversamente negli incipit delle girate); a ricerche mirate per luoghi, rubriche o testatine del giornale; a tipi di testo; a parole chiave degli articoli e di altri generi testuali talvolta negletti, quali recensioni, inserzioni, echi di cronaca, comunicati stampa, ecc. L'interesse per un corpus di dati linguistici scritti tratti da giornali piemontesi nasce inoltre dalla possibilità di archivio e interrogazione di materiali provinciali e regionali sinora mai analizzati dal punto di vista linguistico.

1 http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica.

2 La *Biblioteca Digitale Italiana* mira a promuovere e coordinare le attività di digitalizzazione del patrimonio bibliografico e archivistico italiano: http://www.bibliotecadigitaleitaliana.it/genera.jsp?s=1&l=it.

3 http://periodicipiemonte.econ.unito.it/

2. Dal masso di Rolando[4] ad una rassegna regionale

Il *Corpus Segusinum* rappresenta il primo sottocorpus di una auspicabilmente più ampia raccolta di dati scritti da varietà di italiano giornalistico, con una particolare attenzione alla realtà regionale della stampa piemontese, di cui il presente contributo tratteggia il progetto pilota.

Il corpus è attualmente costituito da due intere annate del giornale *La Valsusa*, una delle testate italiane più antiche. Fondato nel 1897 dal mons. Edoardo Giuseppe Rosaz, *La Valsusa* è un settimanale in uscita il giovedì che si occupa del territorio della Valle di Susa, Val Sangone e parte della cintura di Torino. Il formato è un tabloid di 52 pagine (in media), in parte a colori. I temi trattati vanno dalla cronaca allo sport ad approfondimenti tematici di attualità, cultura e religione. Attuale direttore responsabile, dal 1979, è Ettore De Faveri, affiancato da un gruppo redazionale in cui sono presenti oltre cinquanta giornalisti di tutte le età, il cui motto potrebbe riassumersi nelle parole di Piero Ottone: "Nessuno decide di fare il giornalista per migliorare l'umanità. Però un buon giornale la migliora".

Il *Corpus Segusinum*, che costituisce l'apripista della nostra iniziativa, sarà presto affiancato da un gemello astigiano, il *Corpus Hastense*, basato sui testi della *Gazzetta d'Asti*, di cui abbiamo acquisito i testi delle annate 2003 e 2004. Storico strumento d'informazione, la Gazzetta (in uscita ogni venerdì) rappresenta nel panorama locale una delle testate cittadine più antiche e più conosciute dell'Astigiano, di cui Masino 2007 ripercorre i primi 110 anni di attività. Settimanale cattolico fondato nel 1899 e fortemente voluto dal Vescovo di Asti di allora, Mons. Giacinto Arcangeli, dopo un decennio (1955-'65) nel quale si restrinse maggiormente alle notizie di tipo religioso, a partire dal 1967 la *Gazzetta* ritrovò il più ampio tradizionale interesse a tutti gli aspetti della vita sociale, dalla cronaca nera allo sport e al mondo del lavoro col nuovo direttore don Giulio Martinetto, assistente diocesano delle Acli, e più ancora con l'attuale direttore don Vittorio Croce, alla guida del giornale dal 1976.

Con l'arrivo di nuove collaborazioni e nuove testate il progetto punterà a offrire rinnovata visibilità a validi prodotti della realtà della carta stampata piemontese e agli argomenti che essa, e solo essa, tratta, promuovendo una tradizione locale non sempre valorizzata dalla diffusione delle sole testate a tiratura nazionale. Si tratta in alcuni casi di restituire vitalità a testi di nicchia, costretti spesso a operare con scarsi stimoli o guadagni, e di documentare realtà di grande vitalità ma talora a rischio di estinzione, poiché minacciate da altre forme di informazione (per es. gli inserti provinciali di testate nazionali).

Ovviamente il corpus, pur nato per ragioni linguistiche, potrà in futuro servire anche da archivio digitale delle singole testate giornalistiche e documentazione della loro storia.

4 Il riferimento è alla *péra ëd Roland*, "il curioso masso fesso, sito all'estremità settentrionale del comune di Villarfocchiardo (*Vilarfociàrd*), e legato alla leggenda carolingia del passaggio del paladino Orlando, che ha ispirato il logo del corpus", secondo la spiegazione offerta nel sito.

Sono stati, infatti, presi contatti con possibili nuove acquisizioni in territorio piemontese, con cui è tuttora in corso la definizione del contratto (di tipo *Creative Commons*) per le questioni legali connesse al diritto d'autore (sul quale si veda la soluzione già illustrata in Barbera, Corino & Onesti 2007)[5]. Si mira, in particolare, ad avviare a breve una collaborazione con *Il Corriere di Alba-Bra-Langhe-Roero*, *l'Eco di Biella* e *Il Mercoledì*, auspicando in seguito la possibilità di ampliare ed equilibrare i contributi in una rassegna regionale organica, che ben rappresenti le diverse anime della varie province piemontesi e della sua carta stampata.

I testi saranno sottoposti a un trattamento analogo a quello che è stato portato avanti su *La Valsusa* (cfr. oltre § 3.1), consentendo dunque una maggiore rappresentatività dei materiali del corpus. L'*homepage* del progetto di ricerca è già attualmente disponibile alla pagina web http://www.bmanuel.org/projects/vs-HOME.html, che presenta gli obiettivi del progetto, una descrizione del linguaggio di interrogazione CQP necessario per l'interrogazione linguistica con *query*; la documentazione analitica relativa al processo di *markup* dei dati; il *link* alla versione beta del corpus per l'interrogazione vera e propria dei materiali.

3. Uno standard metodologico per corpora giornalistici

L'attività di ricerca svolta sui testi del sottocorpus *Segusinum* si propone un obiettivo anche metodologico nel tracciare alcune linee guida fondanti per l'annotazione di varietà di linguaggio giornalistico. Il corpus nasce infatti con l'obiettivo di enucleare tratti peculiari del tipo testuale "articolo di giornale" così come degli altri tipi di testo ricorrenti nella stampa periodica, locale e non.

Si è posto pertanto particolare interesse alle possibili sfaccettature presenti in quotidiani e settimanali, focalizzando l'attenzione sulle tipicità della titolatura e sottotitolatura, sulla strutturazione talvolta discontinua che caratterizza girate e testi introdotti da civette e sugli elementi potenzialmente degni di attenzione per l'analisi linguistica, sia essa rivolta ad aspetti morfosintattici e lessicali, sia essa ristretta a questioni di linguistica storica e dialettologia, ma soprattutto mirando a porre al centro dell'attenzione quella dimensione testuale troppo spesso trascurata nella linguistica dei corpora, ma affatto fondamentale e fondante.

L'attenzione per la titolatura, come già mostrato in numerose sedi (ricordiamo qui solo Dardano 1986; Dardano & Di Meola 1995), è giustificata dal suo fondamentale ruolo di "informazione liofilizzata" (De Benedetti 2004) che dev'essere però anche in grado di richiamare l'attenzione del lettore in modo

5 Dispendiosa in termini di tempo è stata in taluni casi la risoluzione di impedimenti burocratici nell'interazione con i direttori di giornali o legati a una istintiva diffidenza nei confronti della ricerca accademica: si è pertanto spostato, per una pur breve parte del lavoro, il focus delle energie sulla revisione del contratto da sottoporre alle testate giornalistiche e sulla stesura di una comunicazione ufficiale più efficace, che promuovesse la divulgazione degli scopi scientifici del progetto e sottolineasse l'importanza della collaborazione con le realtà locali del Piemonte. Si è in particolar modo chiarito come il progetto punti anche a offrire nuovo lustro alla stampa piemontese.

accattivante. Se la ricerca dell'effetto brillante richiamato da Dardano è tipica di gran parte dello stile giornalistico (nonché di una certa 'deriva televisiva': Loporcaro 2005: 73 sgg.), ancor di più lo è nel titolo, per ottenere un effetto originale, molto simile per certi versi allo slogan pubblicitario (Dardano 1986; si vedano anche le osservazioni di Oliva 2009 sull'uso dei forestierismi proprio in un contesto di stampa locale). Nell'analisi della neologia semantica all'interno dei titoli è pertanto comune il riscontro di numerosi nuovi derivati o composti, spesso "occasionalismi effimeri con funzione impressivo-conativa" (Dardano & Di Meola 1995: 419) o formulazioni frutto di modificazioni di unità fisse (Dardano & Di Meola 1995: 428-429).

Si tenga presente che spesso si tratta di creazioni apocrife del titolista, composte anche senza l'approvazione esplicita dell'estensore dell'articolo: viene a cadere dunque un principio importante, l'appartenenza di titolo e testo allo stesso testo. Il titolo di giornale si realizza come un atto comunicativo compiuto e indipendente, il cui contenuto informativo è consumato da molti lettori anche senza il corrispondente articolo (divenendo non soltanto il principale veicolo d'informazione ma anche canale privilegiato attraverso cui diffondere idee sociali e culturali). Ecco perché intendiamo introdurre la possibilità di *queries* autonome all'interno di titoli e sottotitoli.[6]

Il valore paradigmatico del markup elaborato nell'ambito del progetto *Segusinum* ne permette l'applicabilità a livello metodologico al di là della regione da cui siamo partiti, al di là della sola stampa periodica,[7] e persino, auspicabilmente per il futuro, a testi giornalistici in lingua diversa da quella italiana.

3.1 Markup e specifiche

Una versione beta di una prima selezione di testi del sottocorpus *Segusinum*, come si diceva, è liberamente disponibile e interrogabile alla pagina, debitamente linkata nella *home* del progetto, http://www.corpora.unito.it/valsusa/cqpmode/, grazie al linguaggio di interrogazione CQP, elaborato dall'IMS di Stuttgart (per maggiori approfondimenti si vedano Christ *et al*. 1999; Heid 2007) e adattato ai nostri scopi da Simona Colombo.

Il formato elettronico sottoposto al software segue una struttura comune, per cui ogni testo è preceduto e definito da una *header*, che contiene tutte le informazioni riepilogative del singolo testo in questione (i cosiddetti *metadata*), designando così tutti i documenti man mano inseriti nel corpus. Per informazioni dettagliate si vedano le *Guidelines*[8] del progetto.

6 Nella beta attualmente disponibile online ciò non è ancora possibile, ma è lo sviluppo previsto come prioritario.

7 Il quotidiano *La Stampa* ha altresì acconsentito a fornire al gruppo di ricerca alcuni dati della testata nazionale (annate 2002 e 2007): per motivi di *copyright*, però, non potrebbero essere resi pubblicamente disponibili, e pertanto non verranno lavorati anche se costituirebbero un prezioso repertorio per il confronto linguistico della stampa regionale con la lingua della stampa a tiratura nazionale, funzione alla quale dovrà assolvere il corpus *La Repubblica* sviluppato dall'Università di Bologna.

8 Cioè il *file* guidelines__vs.pdf, *Corpus di testi giornalistici piemontesi. Linee Guida per la creazione ed il markup*, disponibile online alla *homepage* del *Corpus Segusinum*.

Per chiarire meglio le informazioni fornite dalla *header*, ne riportiamo qui di seguito lo scheletro portante (racchiuso invero tra le indicazioni <HEAD> e </HEAD>):

```
<HEAD>
  <doc-id>
<idN>XXXXnnnnnnnnnnnnnnn</idN>
<charset>ansi</charset>
<lingua>italiano</lingua>
<aut_NC>(nome;?,cognome;?),(nome;?,cognome;?),...;red</aut_NC>
<fornitore>La Valsusa</fornitore>
<titolo>____;0;?</titolo>
<data>(aaaa,mm;0;?,gg;0;?);(0;?)</data>
<luogo>città;?</luogo>
  </doc-id>
  <set-id>
<corpus>Corpus Segusinum</corpus>
<fonte>giorn</fonte>
<doc-id_source>nomefile[vs_2003-35-01-02.txt];0</doc-id_source>
<f_nome>La Valsusa. Settimanale della Val Susa e Val Sangone</f_nome>
<riv_estremi>annata_nnn;0;?,npf_nnn;0;?,suppl;0,pag_nnn-
nnn;nnn+nnn;0;?</riv_estremi>
<f_data>(aaaa,mm;0;?,gg;0;?);(0;?)</f_data>
<gruppo_Sez>____;0</gruppo_Sez> =[Testatina]
<gruppo_Rub>____;0</gruppo_Rub> =[Rubrica]
<gruppo_Ins>____;0</gruppo_Ins> =[Inserto]
  </set-id>
  <autore>
<specifiche>(m;f;?);ente;gruppo</specifiche>
<eta>1-7;8-13;14-18;19-25;26-30;30-40;40-50;oltre;?</eta>
<qualifica>___;?</qualifica>
  </autore>
  <autore2>ripeti_autore_o_canc</autore2>
  <autoreN>ripeti_autore_o_canc</autoreN>
  <testo>
                <tipo_forma>art;comred;petiz;bio;mosc;ins;lett;rec;nov;poem;c-
lib</tipo_forma>
<tipo_artP>aper;box;traf;fond;spal;sspal;?</tipo_artP>
<tipo_artS>cors;edit;elz;serv;interv;comm;pubred;res;comst;appg;dist;necr;spig
;agen;echi;?</tipo_artS>
<tipo_taglio>a;m;b;am;mb;amb</tipo_taglio>
<tipo_stile>giorn;inserz;usl</tipo_stile>
<tipo_fine>divulg;spec;artist;intratt;inform;celeb;emot;d-o</tipo_fine>
<topics>...</topics>
<keyw>____,____,____,____,____</keyw>
<qualita>derEdE</qualita>
  </testo>
  <ref>
    <imgint>nome1.txt;0,nome2.txt;0</imgint>
  </ref>
</HEAD>Tavola 1 Template della header del Corpus Segusinum.
```

Tavola 1 *Template* della *header* del *Corpus Segusinum*.

La prima sezione di tale *header* (<doc-id>) fornisce informazioni di identificazione univoca del documento, articolate nei seguenti attributi: idN (numero progressivo assegnato automaticamente che costituirà l'identificativo assoluto del documento); charset (ovvero *character set* in cui è codificato il documento di test, di default ANSI – il set standard in Windows, praticamente coincidente con l'ASCII ISO 8859-1 Latin 1); la lingua del testo; l'autore e il fornitore del testo; il titolo e la data di produzione; infine il luogo (dichiarato a volte in testa a un articolo).

Il campo <set-id> definisce le informazioni che serviranno a identificare gli insiemi di testi da cui il documento proviene e in cui confluirà. Si raccolgono qui: l'indicazione del sottocorpus cui il testo afferisce (<corpus>: per ora *Segusinum* o *Hastense*); la fonte (distinguendo tra ebdomadari, quotidiani o settimanali, e riviste, da intendersi con periodicità maggiore); la <doc-id _ source> che istituisce la denominazione del file (precisata con una specifica sintassi, costituita dalla sigla "vs" per La Valsusa o "ga" per la Gazzetta d'Asti, seguita da annata - numero - pagina in cui il testo è collocato - numero progressivo del testo elaborato nella pagina); il nome per esteso della fonte; gli estremi della rivista; alcune specifiche relative alla sezione generale in cui il testo è pubblicato (di solito corrispondente alla 'testatina') e ad eventuali rubriche o inserti; dettagli sul produttore del testo (è talvolta possibile conoscerne il sesso, la fascia d'età e l'eventuale qualifica, specificando viceversa se si tratta di un testo redazionale o se l'erogatore del testo è un ente o istituzione).

Infine il terzo settore della *header*, <testo>, fornisce una caratterizzazione testuale del documento: non mette qui conto riferire della interminata bibliografia e di tutte le discussioni innescate dal problema della classificazione dei cosiddetti *generi testuali*; una soluzione assoluta, crediamo, probabilmente non esiste, ma quello che qui si propone è solo una griglia empirica che, sovrapponendo diversi tipi di parametro, consenta una caratterizzazione efficace delle realtà testuali che il corpus comprende, così consentendo all'utilizzatore di orientarcisi e delimitare le sue ricerche in modo soddisfacente. Viene qui *in primis* precisato il tipo testuale, attingendo alla categoria <tipo _ forma> ereditata dagli altri corpora del gruppo, a cui si aggiungono tipi peculiari della varietà giornalistica – con un'ipercategoria "art", articolo, accanto alla quale si contemplano comunicati redazionali, mosconi, inserzioni, lettere aperte e lettere al direttore, petizioni, biografie di personaggi, recensioni, brevi composizioni narrative, eventuali testi poetici per varie ragioni pubblicati nel giornale, composizioni 'libere', come ad es. un tema scolastico riportato in un servizio. Le righe <tipo _ artP> e <tipo _ artS> rendono conto delle caratteristiche posizionali e strutturali del testo: nella prima è possibile selezionare un valore tra apertura, box, fondo, spalla, sottospalla; per quanto concerne la struttura sono molte le possibilità previste dal mondo giornalistico che può interessare interrogare: corsivo, editoriale, elzeviro, servizio, intervista, commento, pubbliredazionale, resoconto, pezzo d'appoggio, comunicato stampa, distico, necrologio (comprendente anche i pezzi talvolta designati come "coccodrilli"), spigolature, agenda, echi di cronaca. La classificazione in base alle tipologie strutturali degli articoli a volte è sovrapponibile a quella posizionale di <tipo _ artP>, altre volte sono tut-

tavia esclusive e non incrociabili, nel qual caso si assegnerà il valore nullo al campo non pertinente.

Il <tipo__taglio> è introdotto per marcare la posizione orizzontale occupata nella pagina dai testi, quali che siano le colonne (distinguendo dunque tra un taglio alto, medio, basso o che copre più zone di una pagina).

Il <tipo__stile> marca lo stile, in genere, in cui un testo è stato scritto, con l'attuale differenziazione in tre valori: stile giornalistico, proprio di articoli di vario genere; inserzionistico; infine quello che è stato definito usuale/privato, tipico di molti mosconi o lettere, particolarmente significativi peraltro nella stampa locale, per un maggiore senso di appartenenza dei lettori alle testate regionalmente circoscritte o, a maggior ragione, relative solo a una determinata provincia, zona o vallata.

Il <tipo__fine> marca la finalità con cui un testo è stato prodotto: si distingue tra i valori divulgativo, informativo, specialistico (per es. per testi scientifici), artistico (cfr. novelle, testi poetici), di intrattenimento, celebrativo, con funzione emotiva, domanda/offerta.

In prospettiva dell'armonizzazione del corpus con altri corpora in allestimento, sarà successivamente introdotta una classificazione tematica adeguata di ogni documento (in questa prima fase il campo viene semplicemente ignorato, lasciando <topics>...</topics>). Sono state inoltre indicate con annotazione manuale cinque parole chiave (*keyw* sta per "keywords") che contribuiscano ad individuare l'argomento del documento. Se si tratta di articolo, come prima parola chiave va sempre assegnato uno dei settori di competenza: interni, esteri, cronaca (bianca, nera, rosa), sport, economia, cultura, costume, spettacoli.

Infine la *header* mostra le indicazioni di <qualita>, cui si attribuisce di *default* per i corpora *Segusinum* e *Hastense* il valore "derEdE", cioè derivato di copia elettronica (di entrambe le testate possediamo infatti versioni in *.pdf*); e <ref>, sezione dedicata ai riferimenti interni ad immagini o allegati testuali contenuti nel testo: di norma, non essendo le immagini (<imgint>) da considerarsi nella composizione del corpus, la riga è cancellata. In considerazione delle potenziali esigenze del ricercatore e del linguista, una siffatta *header* permette dunque di risalire immediatamente al numero di pagina in cui il testo è collocato; alla posizione del testo all'interno della pagina; ai tipi di testo; alle parole chiave degli articoli; al testo contenuto in rubriche o testatine del giornale, e così via.

3.1.2 Il markup interno del corpus

Anche il *markup* interno dei testi elaborato per il corpus segue alcuni elementi 'ordinari' e coerentemente presenti in tutti gli altri corpora del gruppo torinese, cui si va ad aggiungere un *markup* 'speciale', creato ad hoc per le peculiarità del linguaggio giornalistico e delle sue declinazioni. Si illustreranno di seguito alcuni aspetti ritenuti rilevanti per la classificazione e lo studio linguistico della stampa periodica, per i cui dettagli rimandiamo nuovamente alle *Guidelines* del progetto.

Grazie al *markup* cosiddetto ordinario si potranno, in primo luogo, interrogare i dati presenti nel corpus per forestierismi, toponimi, antroponimi, datazio-

ni, evidenziazioni grafiche della testata (corsivo, neretto, maiuscoletto, caratteri sottolineati o puntinati), turni di dialogo e discorso diretto, citazioni, eventuali indirizzi web. Speciali etichette, poi, evidenziano all'interno del documento gli occhielli (<occhiello>), i titoli (<tit>) e i sottotitoli (<catenaccio>, come da gergo giornalistico), per le motivazioni poc'anzi avanzate (cfr. *supra* § 3): tali *tags* consentiranno infatti *queries* di specifiche occorrenze o parti del discorso limitando la ricerca all'interno delle titolature.

Trovano altresì una loro eco tutte le occorrenze di girate e civette (<girata>; <civetta>) – differenziando dunque rispettivamente la parte di articolo che continua in una pagina successiva da quella in cui l'articolo è iniziato rispetto alla mera segnalazione in prima pagina di un articolo posizionato nelle pagine interne del giornale. Sono escluse dalla trascrizione le immagini e didascalie, eventuali testi pubblicitari e righe del tipo: "continua a pag. 27", che si evincono dalle indicazioni di pagina nella *header* così come di girata nel corpo del testo.

Tale *markup* richiede notevoli risorse in termini di annotazione manuale; ad oggi sono circa 2200 i testi già elaborati, a fronte di una quantità decisamente maggiore di testi che è attualmente in fase di etichettatura.

```
<HEAD>
  <doc-id>
    <idN>XXXXnnnnnnnnnnnnnnn</idN>
    <charset>ansi</charset>
    <lingua>italiano</lingua>
    <aut__NC>Serena,Oggero</aut__NC>
    <fornitore>La Valsusa</fornitore>
    <titolo>Come si stampava nel Settecento?</titolo>
    <data>2003,09,04</data>
    <luogo>Rivoli</luogo>
  </doc-id>
  <set-id>
    <corpus>Corpus Segusinum</corpus>
    <fonte>giorn</fonte>
    <doc-id__source>vs__2003-33-43-07.txt</doc-id__source>
    <f__nome>La Valsusa. Settimanale della Val Susa e Val Sangone</f__nome>
    <riv__estremi>annata__103,npf__33,0,pag__043</riv__estremi>
    <f__data>2003,09,04</f__data>
    <gruppo__Sez>Spettacoli e recensioni</gruppo__Sez>
    <gruppo__Rub>0</gruppo__Rub>
    <gruppo__Ins>0</gruppo__Ins>
  </set-id>
  <autore>
    <specifiche>f</specifiche>
    <eta>?</eta>
    <qualifica>?</qualifica>
  </autore>
  <testo>
    <tipo__forma>art</tipo__forma>
    <tipo__artP>0</tipo__artP>
    <tipo__artS>serv</tipo__artS>
    <tipo__taglio>b</tipo__taglio>
    <tipo__stile>giorn</tipo__stile>
    <tipo__fine>inform</tipo__fine>
    <topics>...</topics>
    <keyw>cultura,Rivoli,mostra,stampa,Settecento</keyw>
    <qualita>derEdE</qualita>
  </testo>
</HEAD>
```

<BODY>
043
%001%<occhiello><emph__ng>RIVOLI</emph__ng>
<blank__D>Sabato <date__2003-09-06>6 settembre</date></blank></occhiello>
<tit><emph__bb>Come si stampava nel Settecento?</emph__bb></tit>
#001#Si inaugura sabato <date__2003-09-06>6 set|tembre</date> ,
alla <topn>Casa del
<anth>Conte Verde</anth></topn> di <topn>Rivoli</topn> la
mostra “La Stamperia
Reale di <topn>Torino</topn> e le tecni|che
di stampa del Sette|cento” .
Nocciolo prezioso
della manifestazione la
presentazione di pregiate
edizioni settecentesche
della Stamperia Reale di
<topn>Torino</topn> , affiancate da una
sezione dedicata alla
stampa e alle tecniche
dell’ incisione e da una
mostra didattica sulle legature
curata dal profe|ssor
<anth>Malaguzzi</anth> . “<ddir>Lo scopo è
migliorare la conoscenza
della produzione editoria|le
del Settecento in <topn>Pie|monte</topn>
presentando un
centinaio di opere già pre|senti
nei cataloghi fino ad
ora pubblicati e altre cin|quanta
opere sconosciute
o mal note , per ricordare
attraverso i libri , perso|naggi ,
interessi e conqui|ste
dell’ ingegno umano
nel campo scientifico , ar|tistico
e letterario</ddir> ” spiega
<anth>Alessandro Bima</anth> , curato|re
dell’ esposizione delle
opere librarie .
#002#Saranno in mostra veri e
propri capolavori artistici ,
ma anche libri legati alla
vita quotidiana , scolastici ,
religiosi . Alla Stamperia
Reale , nata nel <date__1740-??-??>1740</date> , si
devono infatti numerose
edizioni di buon livello ti|pografico
e di contenuto
valido , che spaziano dal|l’ antiquaria
alle scienze ,
dalla storia ecclesiastica
al diritto , dall’architettu|ra ,
all’arte militare e alla
medicina . Tra i titoli pre|senti
un <emph__i>“<oper>Corpus Juris Ci|vilis</oper>”</emph__i>
del <date__1782-??-??>1782</date> , l’ <emph__i>“<oper>Introdu|zione
allo studio della reli|gione</oper>”</emph__i>
di <anth>Giacinto Sigi|smondo
Gerdil</anth> (<date__1755-??-??>1755</date>) , te|sti
scientifici come gli
<emph__i>“<oper>Elementi d’ algebra</oper>”</emph__i> di

```
<anth>Pietro Paoli</anth> e <anth>G. T. Miche|lotti</anth> .
Anche <topn>Susa</topn> diventa
protagonista della mostra
con l' opera <emph__i><emph__b>"</oper>L' arco anti|co
di <topn>Susa</topn> descritto e
disegnato dall'architetto
<anth>Paol'Antonio Masaz|za</anth></oper>"
di</emph__b></emph__i> <anth>Paol'Antonio Mas|sazza</anth>
(<date__1750-??-??>1750</date>) , "<citaz>bellissima
opera che unisce il rigore
scientifico del contenuto
alla pregevole veste tipo|grafica</citaz>" .
#003#I volumi creano un per|corso
cronologico e tema|tico
insieme , mentre una
sezione illustra l' attività
della Stamperia Reale co|me
tipografia ufficiale del|lo
Stato sabaudo . L' inizia|tiva
offre un piccolo as|saggio
delle numerosi edi|zioni
della Stamperia , per
la prima volta una mostra
dedicata ad una singola
tipografia torinese del
Settecento . L' appunta|mento
per l' inaugurazione
è sabato <date__2003-09-06>6 settembre</date> , alle
ore 18 in <topn>via Fratelli Piol ,
8</topn> , <topn>Rivoli</topn> .
004#La mostra , promossa dal|l' Assessorato
alle Politiche
Educative e Culturali del|la
città di <topn>Rivoli</topn> , sarà
aperta dal <date__2003-09-07>7 settembre</date> al
<date__2003-11-16>16 novembre</date> con i se|guenti
orari : martedì-ve|nerdì
ore 15-19 , sabato e
domenica ore 10-12,30/
15-19 . Per maggiori infor|mazioni :
#005#<tel>011.9563020</tel>/<tel>011.95368
09</tel> oppure in rete al sito
<url>www.comune.rivoli.to.it</url> .
</BODY>
```

Tavola 2 Esempio di articolo markuppato per il corpus *Segusinum*.

4. Il *Corpus Segusinum* come chiave di accesso alla linguistica dei corpora

I *files* corredati di markup, ordinario e specialistico, sono stati elaborati nel corso dei tirocini curricolari attivati da Manuel Barbera e Carla Marello[9] dell'Università di Torino: scopo del tirocinio è stato ed è tuttora quello di favorire l'acquisizione di familiarità con strumenti informatici di trattamento del lin-

9 Tirocini "Annoluce.com" presso la Facoltà di Lingue e Letterature Straniere, Università di Torino.

guaggio naturale e gestione di banche di dati non numerici, nonché l'avvicinamento alla terminologia e allo stile del linguaggio giornalistico. Il lavoro ha permesso soprattutto di fornire una conoscenza dei principi basilari della linguistica dei corpora e della costruzione concreta di un corpus. A partire dal 2007 il coordinamento e la gestione di tali percorsi di tirocinio hanno visto la formazione di numerosi studenti impegnati nell'annotazione manuale dei testi, seguiti da un tutor. Questo tipo di incarico ha avvicinato decine di persone a scoprire 'dall'interno' la *corpus linguistics*, ponendosi come uno degli sporadici casi di formazione sul campo in questo àmbito ancora parzialmente negletto nel panorama linguistico italiano.

L'inserimento dei dati nella *header* e del *markup* nei testi è stato recentemente reso più semplice da una maschera di trascrizione, il *TransPaper*, ideata e realizzata da Mauro Costantino e perfezionata da Luca Procopio: si tratta di un compilatore (*open source*, e presto disponibile sul sito) che, per aggiustamenti successivi, è diventato il più possibile funzionale e *user-friendly*, mirando a rivolgersi senza ambiguità anche a un etichettatore poco esperto. Potrà essere in tal modo sfruttato in futuro anche da trascrittori di testi che, dopo una breve formazione iniziale, potranno lavorare a distanza, risparmiando notevoli risorse, *in primis* umane, per l'ampliamento del corpus – il tipo di risorse che più ci è stato fondamentale rispetto alle decisioni di etichettatura manuale.

5. Applicazioni *in fieri*: una conclusione provvisoria

L'elaborazione di dati dalla realtà piemontese è legata non solo a un interesse diatopico, ma ancor più a un aspetto diafasico, considerata la situazione enunciativa della stampa locale che presenta frequentemente caratteristiche dissimili dalla stampa a tiratura nazionale. Essa è infatti svincolata da un pubblico nazionale, nella scelta degli argomenti per esempio, che evidenzia episodi che possiamo considerare di valore solo per una specifica comunità.

Il linguaggio della stampa regionale ha inoltre maggiori probabilità di essere soggetto al contatto con forme dialettali. Rispetto a tale aspetto potrebbero per es. intervenire costruzioni sintattiche che seguono il modello di noti piemontesismi: si pensi a una struttura quale "solo più", registrata nel GRADIT come regionalismo piemontese e da più parti analizzata come vero e proprio calco del gergale *mac pi*[10] (es.: "Ho solo più tre giorni per studiare"). "Sicurissima spia dialettale" per Cortelazzo (1982: 120 sgg.), il rafforzamento degli avverbi *solo*, *soltanto*, *solamente* è approfondito da Telmon (1993, 2001) e da Regis (2006), il quale ne evidenzia specialmente le ragioni che possono giustificare la sua persistenza nel parlato piemontese a fronte di altre forme equivalenti: "L'impressione è che *solo più* costituisca, agli occhi del parlante piemontese, la scelta più economica per

10 Una consulenza linguistica sorta dal quesito di una navigatrice della Crusca online dà adito alla curiosa testimonianza di Maria Corti che, in un'intervista di Maria Grazia D'Oria, dichiarò: " 'solo più' […] salì all'onore della citazione sulla "Stampa" allorché un assassino ricercato vi inviò una lettera; tutti sospettavano di un veneto, ma il linguista Benvenuto Terracini intervenne sulla *Stampa*, chi scriveva 'solo più' non poteva essere che piemontese" (http://www.accademiadellacrusca.it/faq/faq__risp.php?id=8326&ctg__id=44).

esprimere sinteticamente un significato complesso" (Regis 2006: 280); funzionalità su cui insiste ancor più recentemente De Benedetti (2009) da una prospettiva semantica. Un'altra motivazione enucleata da Regis di tale stabilità (ma anche dell'estensione all'uso di parlanti piemontesi senza competenza dialettale, che potrebbe a maggior ragione interessare il nostro caso), è il rapporto di simmetria con la locuzione *neanche/nemmeno più*, uno "speculare negativo" di *solo più* (corrispondente peraltro alla coppia speculare del dialetto *gnanca/manch pì* da un lato, e *mac pì* dall'altro; Regis 2006: 281 sgg.). L'analisi del corpus sosterrà, grazie a una mole di dati che vuole essere rappresentativa di diverse zone del Piemonte, analisi linguistiche di questo tipo accanto a costruzioni quali *non…mica* o *fare che* + VERBO.

Può rivelarsi interessante anche la diversa penetrazione di alcuni forestierismi. Il prototipo in locale ci ha già permesso di analizzare la presenza di termini stranieri in articoli della zona astigiana: i primi carotaggi sulla presenza di forestierismi nella *Gazzetta d'Asti* da parte di Oliva (2009) mostrano un diverso utilizzo del repertorio dei francesismi rispetto ai contesti d'uso degli anglismi, e non solo per motivazioni puramente semantiche, bensì per cause – in numerosissimi casi – stilistiche. È inoltre emerso come, a livello di stampa periodica locale, larga influenza eserciti anche l'impiego dei dialettismi, i quali da un lato forniscono maggiore colore locale, dall'altro suggeriscono una linea comune con le scelte editoriali del giornale.

Lo studio ha voluto allargare, pur parzialmente, le prospettive degli studi linguistici sulla stampa a tiratura regionale, puntando anche a stimolare l'interesse rivolto alla dimensione locale. Questo resta uno dei nostri obiettivi precipui, che dischiuda tuttavia anche gli orizzonti dell'applicazione metodologica a corpora di varietà giornalistiche in senso lato.

Riferimenti bibliografici

Barbera M., Corino E. & Onesti C. (2007) (a cura di) *Corpora e linguistica in rete*, Perugia, Guerra Edizioni.

Barbera M., Corino E. & Onesti C. (2007) "Che cos'è un corpus? Per una definizione più rigorosa di corpus, token, markup", in Barbera M., Corino E. & Onesti C. (a cura di), *Corpora e linguistica in rete*, Perugia, Guerra Edizioni, pp. 25-88.

Bonomi I., De Stefanis Ciccone S. & Masini A. (1983) (a cura di) *La stampa periodica milanese della prima metà dell'Ottocento: testi e concordanze*, Pisa, Giardini.

Baroni M. *et al.* (2004) "Introducing the "La Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian", in *Proceedings of LREC 2004*, Lisbon, ELDA.

Christ O. *et al.* (1999) *The IMS Corpus Workbench: Corpus Query Processor (CQP). User's Manual*, Institut für maschinelle Sprachverarbeitung, Stuttgart, August 16, 1999 (CQP V2.2), http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/CQPUserManual/PDF/cqpman.pdf (consultato il 10/01/2011).

Cortelazzo M. (1982) *Memoria di parole. Dialetto tra vita e letteratura*, Ravenna, Edizioni del Girasole.

Dardano M. (1986) *Il linguaggio dei giornali italiani*, Roma-Bari, Laterza.

Dardano M. & Di Meola C. (1995) "Note sulla semantica dei titoli della stampa italiana e austriaca", in Dardano, M., Dressler, W.U. &

Di Meola, C. (Hgg.), *1995. Parallela 5. Akten der 6. österreichisch-italienischen Linguistentagung (Rom, 20-22.9.1993)*, Roma, Bulzoni, pp. 415-453.

De Benedetti A. (2004) *L'informazione liofilizzata. Uno studio sui titoli di giornale (1992-2003)*, Firenze, Cesati.

De Benedetti A. (2009) *Val più la pratica - Piccola grammatica immorale della lingua italiana*, Bari, Laterza.

De Mauro T. (2000), *Grande dizionario italiano dell'uso*, Torino, UTET (ed. in CD-ROM).

Gualdo R. (2007) *L'italiano dei giornali*, Roma, Carocci.

Heid U. (2007) "Il Corpus WorkBench come strumento per la linguistica dei corpora. Principi ed applicazioni", in Barbera M., Corino E. & Onesti C. (a cura di), *Corpora e linguistica in rete*, Guerra Edizioni, Perugia, pp. 89-108.

Loporcaro M. (2005) *Cattive notizie*, Milano, Feltrinelli.

Masino S. (2007) *110 anni con la Gazzetta d'Asti (1899-2009), la Storia in prima pagina*, Asti, Edizioni Gazzetta d'Asti.

Oliva G. (2008-09) *Per un corpus di giornali regionali piemontesi. Alla ricerca dei francesismi perduranti*, tesi di laurea non pubblicata, Università di Torino.

Regis R. (2006) "Breve fenomenologia di una locuzione avverbiale: il 'solo più' dell'italiano regionale piemontese", in *Studi di lessicografia italiana*, XXIII, pp. 273-289.

Telmon T. (1993) "Varietà regionali", in Sobrero A.A. (a cura di), *Introduzione all'italiano contemporaneo. La variazione e gli usi*, Bari-Roma, Laterza, pp. 93-149.

Telmon T. (2001) *Piemonte e Valle d'Aosta*, Roma-Bari, Laterza.

Sitografia

BDI
http://www.bibliotecadigitaleitaliana.it/genera.jsp

Corpus Segusinum
http://www.bmanuel.org/projects/vs-HOME.html

Periodici Piemonte e Val d'Aosta
http://periodicipiemonte.econ.unito.it/

Corpus *LaRepubblica*
http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica

(consultati il 10/01/2011)