Corpus di testi giornalistici piemontesi

LINEE GUIDA per la creazione ed il markup

INDICE

1. Header

1.1. Template

```
<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.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>

<BODY>
[...]
</BODY>
```

1.2. Commento alla header

****<doc-id>****
Informazioni che serviranno ad identificare univocamente il documento una volta inserito nel corpus. Sono articolate nei seguenti attributi:

*** idN ***
Numero progressivo che sarà l'identificativo assoluto del documento: lasciare vuoto, verrà assegnato automaticamente.

*** charset ***
Il character set in cui è codificato il documento di testo. Sono possibili due soli valori alternativi: ansi, ossia il set standard in Windows, coincidente con l'ASCII ISO 8859-1 Latin 1, e unicode, da usare solo per i testi che presentino caratteri non-latini; il valore di default è ovviamente ansi (<charset>ansi</charset>).

*** Lingua ***
Di default è l'italiano.

*** Autore ***
In <aut_NC> si indica il nome dell'autore, il produttore effettivo del testo. I campi nome e cognome possono essere riempiti anche con nomi multipli o complessi usando lo spazio, per cui potremmo avere, ad es. <aut_NC>Pablo Martín Melitón,de Sarasate y Navascués</aut_NC>. È previsto il valore non definito (?) in entrambi i campi, nel caso che le generalità dell'autore fossero solo imperfettamente note.
Sono anche previsti i casi in cui gli autori siano più di uno: in questo caso si useranno le parentesi e si attiveranno le gerarchie
<autore1> ... <autoreN> per fornire i dati di ogni autore (si veda più avanti).
Se il testo è redazionale, indicheremo aut=red per redazionale (che richiama, nelle specifiche autore, l'opzione 'gruppo').

*** Fornitore ***
Il giornale che ha fornito il testo, nel Corpus Segusinum: *La Valsusa*.

*** Titolo ***
Il titolo del testo trascritto.

*** Data ***
Data di produzione del testo, espressa secondo il sistema aaaa,mm,gg (anno, mese giorno), saturabile da valori numerici o da quello non definito (?), ad es.
“14 febbraio 2001” --> 2001,02,14
“Dicembre 1999” --> 1999,12,??
estate 2003 --> 2003,06-09,??
I valori nulli o non definiti sono applicabili anche a tutto l’attributo nel suo complesso qualora tutto il campo data e non solo una sua parte risulti sconosciuto o non pertinente (<data>????,??,??</data>).
Coincide solitamente con la datazione indicata più avanti nel <f_data> (nel campo <set-id>), tranne quando nell'articolo si citi esplicitamente che l'articolo è stato redatto in data diversa.

*** Luogo ***
Quello dichiarato a volte in testa all'articolo (nome della città o *?* se ignoto).

****<set-id>****
Informazioni che serviranno ad identificare gli insiemi di testi da cui il documento proviene ed in cui confluirà.

<corpus>
Di default il valore da attribuire sarà: Corpus Segusinum.

<fonte>
Fonte è un giornale, diverso da rivista; 'giorn' sono quotidiani o settimanali (ebdomadari), 'riv' sono da intendersi con periodicità maggiore.

<doc-id_source>
Si inserisce il nome del file, costituito dalla sigla "vs" seguita da annata - numero - pagina in cui il testo è collocato - numero progressivo del testo elaborato in quella pagina (semplicemente a seconda dell'ordine in cui procedete nell'etichettarli) tutti separati da un trattino.
Separato infine da un trattino basso (_ underscore) potete facoltativamente aggiungere il numero progressivo del vostro ordine di lavorazione (utile più che altro a voi per non confondere eventuali rielaborazioni e correzioni successive dello stesso testo):
Es.:
nel caso si tratti della terza versione che fate del settimo articolo che elaborate a pagina 3 del n. 35 dell'annata 2003, avremo =>
<doc-id_source>vs_2003-35-03-07_03.txt</doc-id_source>.

<f_nome>
Si inserisce il nome per esteso della fonte, sempre: La Valsusa. Settimanale della Val Susa e Val Sangone.

*** Estremi della rivista ***
Inserire in cifre l'annata indicata in prima pagina (annata_015);
il numero progressivo del fascicolo: npf_nnn;
l'eventuale annotazione del supplemento: suppl;
la pagina (o le pagg.) in cui il testo è contenuto: pag_nnn

È possibile che un testo sia articolato su due pagine (per es. se una civetta in prima pagina rimanda ad un articolo interno o se il testo inizia semplicemente in una pagina e continua in un'altra): in questo caso segnalare entrambe le pagine con un simbolo +.
Es: l'articolo che, nel numero 31 dell'annata 107, inizia in prima pagina seguito dalla segnalazione "continua a pag. 26" risulterà:
<riv_estremi>annata_107,npf_31,0,pag_1+26</riv_estremi>.

*** <f_data> ***
Data di pubblicazione del giornale o rivista, espressa nella forma aaaa,mm,gg (=anno, mese, giorno): è possibile il valore "??" per mm e gg, ma l'anno è obbligatorio.

*** <gruppo_Sez> <gruppo_Rub> <gruppo_Ins> ***
<gruppo_Sez> sezione generale, di solito su titolo corrente della pagina = "testatina" (Nome che viene dato alle singole pagine, in genere posto in alto a sinistra)

<gruppo_Rub> rubrica particolare = "rubrica" (È un appuntamento fisso tenuto da un esperto o da un commentatore che può riguardare settori del mondo politico, sociale, scientifico o dello spettacolo).

<gruppo_Ins> inserto speciale = "inserto" (Gruppo di pagine che, pur essendo parte integrante di un giornale, svolgono un ruolo del tutto autonomo e possono essere staccate e conservate).

*** Autore ***
<specifiche>: Informazioni sul produttore del testo:
se si tratta di individuo, si daranno informazioni (specifiche) sul sesso del produttore del testo (quando accertabile), maschile (m) o femminile (f) o non definito (?); altrimenti si specifica solamente se l'erogatore del testo è un ente od istituzione di qualche natura (ente), o se invece il testo è il risultato del lavoro collettivo di un gruppo di persone (gruppo). Badasi bene che gli articoli redazionali vanno certo ascritti al "gruppo", ma che semplici articoli anepigrafici non devono automaticamente assumersi come redazionali, ma, semplicemente, appunto, come adespoti, e pertanto "?".

In <eta> selezionare la fascia d'età, quando nota e se pertinente.

In <qualifica> segnalare eventuali qualifiche, titoli, ecc. (per es.: Dott.; Cav.; Sindaco di Rosta)

<autore2> <autoreN>
Da riempire, replicando le entrate di <autore>, solo nel caso di più autori di uno stesso testo. In caso contrario cancellare la riga.

****<testo>****
Caratterizzazione testuale del documento.

<tipo_forma>
*** art;comred;petiz;bio;mosc;ins;lett;rec;nov;poem;c-lib ***
Sono i tipi_forma ereditati dagli altri corpora a cui si aggiungono: "comred", "mosc" ed "ins" (qui sotto definiti).

Il tipo "art" è l'ipercategoria di cui i due tipi_art che seguono nella header, <tipo_artP> e <tipo_artS>, sono sottodistinzioni, il primo per specificare le caratteristiche posizionali ed il secondo quelle strutturali (mentre quelle propriamente contenutistiche - cronaca nera, rosa, interni, ecc. - sono desumibili dalla prima keyword, v. più avanti).
"Rec" (=recensione), che potrebbe essere plausibilmente posta in <tipo_artS>, ne è invece esclusa a favore di <tipo_forma> perché già assegnata al <tipo_forma> in altri corpora precedentemente prodotti (Valico, Vinca, Athenaeum).

I vari tipi di forma del testo si distinguono secondo i seguenti criteri:

comred
comunicato redazionale (brevi comunicazioni ai lettori da parte della redazione - o di autori di essa facenti parte - per scuse, annunci, ecc.)

mosc
moscone (Breve notizia a pagamento che annuncia una morte, una nascita, un matrimonio, ecc...) = annuncio non commerciale; la differenza degli echi e dei necr non sono redazionali.

ins
inserzione (Messaggio pubblicitario che viene pubblicato a pagamento) = annuncio commerciale e/o pubblicitario.

lett
comprende Lettera aperta (Articolo sotto forma di lettera o lettera inviata al direttore di una testata per la pubblicazione) e Lettere al direttore (Rubrica in cui il direttore risponde ai suoi lettori).

petiz
petizioni

bio
biografie di personaggi (esclusivamente tali: quindi non articoli su Pippo, ma di solito pezzi biografici in appoggio ad un articolo su Pippo sul Rocciamelone, ecc.)

rec
è usato per le applicazioni tradizionali di recensioni di spettacoli, film, libri e simili; le eventuali cronache di convegni e simili, in Athenaeum perlopiù poste sotto "rec", sono nei giornali più propriamente da considerare "res" (resoconti): cfr. oltre sotto <tipo_artS>.

nov
una novella o altra breve composizione narrativa per varie ragioni pubblicata o riportata nel giornale.

poem
un testo poetico per varie ragioni pubblicato o riportato nel giornale.

c-lib
composizione libera, come ad es. un tema scolastico riportato in un servizio. In Valico si distingueva ulteriormente: di tipo misto o imprecisabile (c-lib_var), di tipo descrittivo (c-lib_descr), narrativo (c-lib_narr), regolativo (c-lib_reg), argomentativo (c-lib_arg); cosa che qui (finora) non è parso il caso di fare.

<tipo_artP>
*** aper;box;traf;fond;spal;sspal;? ***

Il <tipo_artP>, che specifica le caratteristiche posizionali in genere dell'articolo, richiede, come ovvio, obbligatoriamente il valore "art" espresso nel campo <tipo_forma>. La classificazione in base alle tipologie posizionali degli articoli a volte è sovrapponibile a quella strutturale di <tipo_artS>, ed in tal caso saranno

selezionati valori non nulli in tutti e due i campi, ma altre volte è esclusiva e non incrociabile con quella di <tipo_artS>, nel qual caso si assegnerà il valore nullo al campo non pertinente.

Spiegazione dei valori:

aper
Apertura: Articolo pubblicato in prima pagina, dedicato alla notizia più importante del giorno.

box
Piccolo spazio evidenziato nella pagina dedicato ad un approfondimento o ad un inciso. Comprende anche i pezzi che a volte sono chiamati "Incorniciato" (tra due linee orizzontali o circondata da un filo tipografico), "Palchetto" (su una o due colonne, circondato da una cornice) e "Riquadro" (all'interno di una cornice)

traf
Trafiletto (Notizia molto breve posta generalmente in fondo alla pagina). Comprende anche i pezzi che a volte sono chiamati "Pallino" (Breve notizia senza titolo) e "Breve" (Notizia di poche righe senza titolo)

fond
Fondo (Commento autorevole ad un fatto di notevole importanza collocato quasi sempre in prima pagina). Comprende anche i pezzi che a volte sono chiamati "Fondino" (Articolo di fondo che compare in una prima pagina di settore. Rispetto al fondo è più breve), comunque distinguibili incrociando con <gruppo_sez>.

spal
Spalla: (l'articolo collocato in prima pagina in alto a destra che in genere ospita un articolo di rilievo).

sspal
Sottospalla (Articolo che occupa le prime due colonne in alto a destra del foglio).

Nota:
Alcuni "tipiP" di articolo non qui espressi sono comunque ricavabili altrimenti. Così:
Postilla (Nota o citazione a pié di articolo) è solo in markup => nota.
Centro (Notizia, di solito nella prima pagina, che occupa una posizione centrale) => taglio.
Centrotesta (Spazio tra apertura e spalla, in alto, al centro della pagina) => taglio.
Pié di pagina (Posizione dell'articolo pubblicato a più colonne sul fondo della pagina) => taglio.
Piedino (Annuncio pubblicitario o breve articolo pubblicato in fondo alla pagina) => taglio.

Se non si riuscisse ad attribuire uno di questi valori all'articolo, inserire *?* (di cui in ogni caso è preferibile non abusare).

<tipo_artS>
*** cors;edit;elz;serv;interv;comm;pubred;res;comst;appg;dist;necr;spig;agen;echi;? ***

Il <tipo_artS>, che specifica le caratteristiche strutturali in genere dell'articolo, richiede, come ovvio, obbligatoriamente il valore "art" espresso nel campo <tipo_forma>. La classificazione in base alle tipologie strutturali degli articoli a volte è sovrapponibile a quella posizionale di <tipo_artP>, ed in tal caso saranno selezionati valori non nulli in tutti e due i campi, ma altre volte è esclusiva e non incrociabile con quella di <tipo_artP>, nel qual caso si assegnerà il valore nullo al campo non pertinente.

Spiegazione dei valori:

cors
Corsivo: Commento breve ma incisivo e polemico scritto, generalmente, in carattere corsivo.
edit
Editoriale: Articolo principale, in genere non firmato, pubblicato sulla prima pagina. Esprime il parere della testata sul fatto politico, sociale, economico più rilevante del giorno.

elz
Elzeviro: Articolo in bella scrittura, destinato alle pagine culturali.

serv
Servizio: È un articolo lungo che prevede un approfondimento dei fatti, con corredo di dati e testimonianze.

interv
Intervista: Riproduzione scritta, (televisiva o radiofonica) rivista e corretta, di un dialogo avvenuto tra il giornalista e l'intervistato. La distinzione, a volte istituita, tra "togata" (quando il colloquio avviene con una persona nella sua veste ufficiale) e "volante" (quando riguarda un personaggio alla ribalta per un fatto del giorno) non è qui perseguita.

comm
Commento: Articolo che non descrive un fatto, ma esprime un'opinione o un'interpretazione. In genere affianca un articolo in cui vengono riportate le notizie del momento.

pubred
Pubbliredazionale: Articolo pubblicitario redatto in stile giornalistico.

res
Resoconto: Riassume dibattiti, lavori congressuali, convegni, sedute parlamentari o le fasi salienti di un processo.

appg
Pezzo d'appoggio: Articolo complementare con note informative o esplicative.

comst
Comunicato stampa: Notizia sintetica priva di commenti su iniziative o manifestazioni diramate da enti, istituzioni, partiti politici, aziende o associazioni.

dist
Distico: Breve articolo che introduce un nuovo collaboratore, una nuova rubrica o una nuova serie di articoli.

necr
Necrologio: Articolo commemorativo per ricordare un defunto. Comprende anche i pezzi che a volte sono chiamati "Coccodrillo"
(Articolo commemorativo, già confezionato, su un personaggio pubblico che viene pubblicato in occasione della sua scomparsa).
A differenza dei "mosc" (cfr. <tipo_forma>) sono o redazionali o firmati da giornalista.

spig
Spigolature: Raccolta di brevi notizie eterogenee, divagazioni o aneddoti, raggruppati sotto un unico titolo.

agen
Agenda: Rubrica in cui si segnalano scadenze, santi del giorno, viabilità, previsioni meteo ed altre informazioni utili. Comprende anche i pezzi che a volte sono chiamati "effemeride" (diario, cronaca giornaliera, raccolta di dati astronomici, rassegna di scienza o letteratura).

echi
Echi di cronaca: Breve notiziario della cronaca cittadina che informa di nascite, morti, matrimoni, lauree, diplomi, inaugurazioni, ecc. A differenza dei "mosc" (cfr. <tipo_forma>) sono redazionali e a differenza dei "necr" (qui sopra) sono brevi.

<tipo_taglio>
*** a;m;b;am;mb;amb ***
Il <tipo_taglio> è introdotto per marcare la posizione orizzontale occupata nella pagina (qualichessiano le colonne) dai testi. I valori possibili sono:

a = taglio alto
m = taglio medio
b = taglio basso
am = taglio alto + medio
mb = taglio medio + basso
amb = tutte le 3 zone della pagina

<tipo_stile>
*** giorn;inserz;usl **

Marca lo stile, in genere, in cui un testo è stato scritto. Qui si distingue (per ora) essenzialmente tra tre valori:

giorn	stile giornalistico, proprio di articoli di vario genere
inserz	stile inserzionistico di molti "ins", "pubred", ecc.
usl	stile usuale, privato, tipico di molti "mosc" o "lett"

<tipo_fine>
*** divulg;spec;artist;intratt;inform;celeb;emot;d-o ***

Marca la finalità con cui un testo è stato prodotto. Si distingue tra i seguenti valori:

divulg	fine divulgativo
spec	fine specialistico (per testi scientifici per es.)
artist	fine artistico (v. novelle, testi poetici)
intratt	fine di intrattenimento
inform	fine informativo
celeb	fine celebrativo
emot	funzione emotiva (per molti mosc)
d-o	domanda/offerta (per molti ins)

<topics>
In prospettiva dell'armonizzazione del corpus con altri corpora in allestimento, sarà successivamnete introdotta una classificazione tematica adeguata di ogni documento. In questa prima fase il campo viene semplicemente ignorato, lasciando <topics>...</topics>.

*** Keywords ***
Si devono indicare alcune parole chiave (keyw sta per "keywords") che aiutino ad individuare l'argomento del documento; il numero di queste è fissato a **5**.
Se si tratta di articolo, come prima keyword va sempre assegnato uno dei seguenti valori:

interni
esteri
cronaca (+ bianca, nera, rosa) (anche per molte delle cosiddette "attualità")
sport
economia
cultura
costume
spettacoli

Nelle keywords successive deve sempre esserci anche il LUOGO cui l'articolo è riferito.

<qualita>
Per il Corpus Segusinum si attribuisce di default derEdE, cioè derivato di copia elettronica.

**** <ref> ****

Parte della header per i riferimenti interni ad immagini od allegati testuali contenuti nel testo.
Di norma, non essendo le immagini da considerarsi nella composizione del corpus, la riga può essere cancellata (cfr. <imgint> per le eccezioni a tale norma).

*** <imgint>nome1.txt;0,nome2.txt;0</imgint> ***

Le immagini che accompagnano l'articolo non sono di norma introdotte nel corpus, né le didascalie che le accompagnano.
Si riprodurranno in un file di testo separato solo le eventuali (e complessivamente rare) immagini con contenuto testuale rilevante, vuoi in sè (quantità), vuoi in relazione all'articolo che accompagna.

2. MARKUP

2.1. MARKUP SPECIALE

<girata>

Il testo continuo (non l'annuncio di un articolo + l'articolo: cfr. "civetta") che si snodi su 2 pagg. differenti verrà riportato in un unico file ma segnalando la seconda parte, detta girata (=la parte di articolo che continua in una pagina successiva da quella in cui l'articolo è iniziato) con un tag contenente eventuali titoletti e l'indicazione della pagina, secondo lo schema:

<girata="PAPA">026
testotestotestotestotestotestotesto</girata>

Un esempio potrebbe essere il seguente:

```
<BODY>
$011$
%001% <occhiello><emph_ng>Devozione sul <topn>Rocciamelone</topn></emph_ng></occhiello>
<tit><emph_b>La consacrazione dei meloni</emph_b></tit>
<catenaccio><emph_i>Gremita la grande e mistica Sagra del Melone</emph_i></catenaccio>
#001# Eccezionale afflusso di pellegrini alla consueta
Sagra del Melone che si tiene tradizionalmente ogni
<date_????-08-15>Ferragosto</date> sul<topn>Rocciamelone</topn>. Ben diciassettemilatrecentosei
i meloni consacrati dall'instancabile Don <anth>Tabone</anth>
sotto gli approvanti sguardi della Madonnina del
<topn>Rocciamelone</topn>.
<girata="Sagra dei meloni">
$017$
#002# I pellegrini che, incuranti della
canicola e della completa assenza di ombra
sulla china del monte assassino, hanno comunque
raggiunto la cima senza gravi inconvenienti,
portando nello zaino il prezioso carico di
meloni da consacrare, non hanno certo avuto
paura di nulla. E ne sono stati ben rimeritati!
Agli altri, trascurabili vittime di vari
simpatici inconvenienti lungo l'amenissima
risalita, è stato comunque offerto un ricovero
premio all'ospedale di <topn>Susa</topn>.
#003# Sulla aprica vetta i mistici meloni,
ormai caldi da fare schifo ...
</girata>
</BODY>
```

<occhiello>

Occhiello: Frase posta sopra il titolo che introduce l'articolo. Talvolta viene stampato in negativo, ossia in lettere bianche su fondo scuro. Comprende anche quello che a volte è chiamato "sopratitolo" (riga che è posta sopra il titolo e riassume il fatto descritto nel testo).

<tit>

<tit> marca il titolo del testo. Segnalare comunque anche il tipo di carattere (grassetto, corsivo, etc.) col markup apposito (cfr. sotto in "Evidenziazioni").

<catenaccio>
Catenaccio: Ulteriore titolo posto sotto a quello principale come una sorta di sommario dell'articolo (sottotitolo).

<civetta>
<civetta>: Segnalazione in prima pagina di un articolo posizionato nelle pagine interne del giornale. Comprende anche i pezzi che a volte sono chiamati "Contornato" (richiamo e riassunto a notizie o servizi contenuti nelle pagine interne) e "Manchette" (riquadro della prima pagina che preannuncia un articolo nelle pagine interne). Si badi che la civetta è solo l'annuncio di un articolo, non la prima parte di un articolo proseguito poi nella "girata" (cfr.). Quindi avremo qualcosa del tipo:

<body>
<civetta> 001
%001% <tit><emph_b>Clamoroso!</emph_b></tit>
#001# Prodigioso nostro servizio sull'evento dell'anno:
la benedizione delle trote da parte del parroco di
Borgone! Correte a pagina 17 a vedere i nostri
servizi sullo stupefacente evento. </civetta>
017
%002% <occhiello><emph_ng>straordinario a <topn>Borgone</topn></emph_ng></occhiello>
<tit><emph_b>La benedizione delle trote</emph_b></tit>
<catenaccio><emph_i>Mai così in forma il parroco di Borgone!</emph_i></catenaccio>
#001# <topn>Borgone</topn>. Gran folla stamane alla curva della statale,
con lievi incomodi per il traffico,
bloccato per sole sette misere orette.
Numerosissimi i pellegrini accorsi da
ogni parte della Valle per la attesissima
e tradizionale <emph_b>benedizione delle trote</emph_b>,
fornite come sempre dal locale allevamento
<ent>"Pharmaton - Da <anth>Paco</anth>"</ent>.
Grande prestazione del parroco di <topn>Borgone</topn,
che, in splendida forma, si è strafogato undici
dei viscidi animaletti, incurante della successiva
lavanda gastrica, che si è presto resa imperiosamente
necessaria. Applausi della folla in estasi.
</body>

<finestra>
Finestra: Testo incorniciato posto all'interno di un articolo.

*** Lettere e simili ***
In caso di lettere o allocuzioni, si mantenga la marcatura già utilizzata negli altri corpora, e cioè:

** Protocollo <pcoll> **
per le formule iniziali (per es.:<pcoll>Gentile signore e signori!</pcoll>)

** Escatocollo <ecoll> **
per le formule di congedo (per es.: <pcoll>porgo distinti saluti.</pcoll>)

** post scripta <PS> **

il testo aggiunto in lettere ed emails dopo l'escatocollo. Sovente è esplicitamente segnato già nel teso con PS, PPS ecc.

*** Indice ***
Nei casi in cui sia presente un indice, e questo sia sufficientemente esteso, può essere mantenuto se compreso nel tag <indice></indice> al cui interno la paragrafatura è indipendente dal testo successivo.

2.2. MARKUP ORDINARIO

*** Livelli di marcatura del testo ***

Tipo	Obbligatorio	Simbolo
Capitolo	SI	%___%
SottoCap1	NO	&___&
SottoCap2	NO	‰___‰
Sottocap3	NO	£___£
Sottocap4	NO	¢___¢
Paragrafo	SI	#___#

In questo modo sono stati definiti 2 livelli assoluti (Capitolo, Paragrafo), di cui l'ultimo per definizione il più basso (più o meno "capoverso") nella gerarchia, che devono obbligatoriamente essere marcati per ciascun testo.
Gli altri sono opzionali e vanno applicati secondo l'ordine della gerarchia (prima il "subcap1", poi se necessario il "subcap2" ecc.); non se ne immagina per ora l'applicazione anche nella Valsusa, ma sono eventualmente disponibili.
Si noti che alcuni livelli superiori sono stati previsti per il corpus legale:

Libro	¥___¥
Titolo	ð___ð
Capo	Þ___Þ
Sezione	µ___µ
"Paragrafo" (legale)	Ÿ___Ÿ

ma non se ne immagina per ora l'applicazione anche nella Valsusa.

I diversi livelli di marcatura del testo vanno numerati al loro inizio con %001%, %002%, ecc. per i capitoli, e #001#, #002#, ecc. per i paragrafi. Ogni sottolivello riprende la propria numerazione da 001 ogniqualvolta inseriamo un nuovo elemento di categoria superiore a tale sottolivello: ciò significa che se inizia un nuovo capitolo, i paragrafi che ne fanno parte non continueranno la numerazione precedente ma riinizieranno da capo.
Solo per i paragrafi è sufficiente porre #___# senza numerare manualmente: lo farà per voi un apposito script in seconda battuta. Gli altri livelli, invece, vanno numerati manualmente in ordine progressivo.

Ricordiamo, infine, che formalmente saranno considerati paragrafi distinti solo blocchi di testo di una certa estensione, di una relativa unità tematica, chiaramente individuati da un punto a capo.

*** Citazione <citaz>...</citaz> ***

Si usa per frasi o porzioni di testo (non per testi completi o loro sezioni ampie, significative e compiute), riconducibili di solito a personaggi del passato, o se del presente citati come auctoritas e non come persona dialogante, o di cui si fa nel testo menzione estemporanea - diversamente da quanto succede per estratti di discorso diretto (cfr. <ddir>) di politici, sportivi, etc. durante interviste, dichiarazioni, e così via.

*** Citazione annidata <citaz2> ***

Si usa quando si debba marcare una citazione DENTRO un'altra citazione.
Es: Come scrive il Testacalda : <citaz> " Secondo Pomponazzi, che scriveva <citaz2> 'ciao ! '</citaz2>, Aristotele pescava le trote con la lenza " </citaz>

*** Discorso diretto <ddir> ***

Marca il discorso diretto.
Es.: E Giorgio disse: <ddir> " Oh basta là " </ddir>.
Si veda anche, diversamente, <citaz> (cfr. supra).

*** Date <date_yyyy-mm-dd> ***

Marca le espressioni di datazione presenti nel testo (siano esse numeriche o frasali).
Esempi:
<date_1231-??-??>1231</date>
<date_1931-03-13>13 marzo 1931</date>
<date_????-12-25>Natale</date> (se considerato in quanto festività/ricorrenza religiosa più che in quanto data; altrimenti si fa capo all'annata del giornale: <date_2004-12-25>Natale</date> se si è certi che il riferimento sia al Natale di quell'anno).

*** turni <turno_xyz>***

Si marcano i turni del dialogo, con indicazione convenzionale (A, B, C, ...), come nell'esempio seguente (tratto dal corpus Valico):

<turno_A>Io : <ddir>Buongiorno , potrebbe aiutarmi ? </ddir></turno>
<turno_B>Commesso : <ddir>Buongiorno Signor , cosa potrei fare per lei ? </ddir></turno>
<turno_A>Io : <ddir>Oggi è il compleanno di mia amica .
Ho preparato una torta buona ma il mio cane la ha mangiata e devo
procurarla da qualche mezzi . </ddir></turno>

*** lingue straniere <lng_nomelingua>...</lng> ***

Saranno markuppate non singole parole considerabili come prestiti non adattati, ma solo le parole, i sintagmi, le frasi od i paragrafi effettivamente non in italiano.
Per es.:
<lng_inglese>last but not least</lng>
<lng_francese>chapeau</lng>

*** indirizzi mail <adress>adress@mah.com</adress> ***
Marca gli indirizzi mail.

*** indirizzi web <url></url> ***
Marca gli indirizzi web.

*** indirizzi telefono <tel> ***
Marca i numeri di telefono e fax.

*** Parti in versi <versi> ***
Per eventuali testi non in prosa (si vedano, per es., le poesie proposte dai lettori che saranno segnalate tra i tags <versi> e </versi>)

*** Elenco <el> ***
La sintassi generale del tag è la seguente:
<el>Elemento marcante</el> Testo

Il testo dell'elenco deve a sua volta essere contrassegnato adeguatamente (ad es. se è un paragrafo sarà marcato con #___# ecc.).
Esempi:

Testo originale:

1. testo prima linea
2. testo seconda linea

Testo marcato:

<el>1.</el> testo prima linea
<el>2.</el> testo seconda linea

Testo originale:

1. testo prima linea lungo e significativo
2. testo seconda linea lungo e significativo

Testo marcato:

#___# <el>1.</el> testo prima linea lungo e significativo
#___# <el>2.</el> testo seconda linea lungo e significativo

Testo originale:

1. testo a volte articolato in paragrafi
1.1 testo a volte articolato in paragrafi

Testo marcato:

%001% <el>1.</el> #001# testo
‰002‰ <el>2.</el> #001# testo

Testo originale:

1. Titolo capitolo 1 dell'articolo
1.1 Titolo sottocapitolo 1 dell'articolo

Testo marcato:

%001% <el>1.</el> <tit> Titolo capitolo 1 dell'articolo </tit>
‰001‰ <el>1.1</el> <tit> Titolo sottocapitolo 1 dell'articolo </tit>

Semplici elenchi a lista senza punti elenco o simili non riceveranno <el> ma saranno trattati come semplici righe (ed il nuovo paragrafo sarà riservato solo per elenchi le cui voci siano blocchi di frasi e non semplici liste). Ad es.
#___# Le opere migliori, oggetto della mostra organizzata nei locali
del Dipartimento di Biologia animale e dell'uomo a corollario del
convegno di presentazione del Calendario, sono state realizzate da:
Elisabetta Berra
Rosanna Gigantiello
Mauro Mantovani
Giovanna Minoggio
Maria Nazario
Luca Zanvercelli
#___#

*** Pagine ***
Le pagine vengono marcate con un 001 ecc. all'inizio di ogni pagina.
Le pagine in cifre romane vengono marcate con un $R001$ ecc. all'inizio di ogni pagina.

*** Note <nota> ***
Per le note a piè di pagina, si proceda come nel seguente esempio che chiarifica l'uso del tag <A href="x">:

%001% [...]
#002# Ai locali in uso all'Università e all'Accademia di Medicina si
accede da un piccolo portone aperto nell'800 in forme neoclassiche
sotto i portici di via Po al n° 18, che certo non spicca tra le aggressive vetrine dei
negozi che lo assediano. Il cortile che così si raggiunge fu il
chiostro con giardino del Convento, oggi appena
percepibile a causa delle profonde trasformazioni che ha subito nel
tempo, soprattutto a partire dall'800: le arcate
di tre dei quattro lati sono state via via chiuse in vario modo, quelle
centrali del lato sud sono in parte scomparse.
Si percorre a sinistra il portico del chiostro nel tratto delle tre
arcate ancora aperte e si raggiunge uno scalone
sormontato da cupola con tamburo ottagonale e lanterna in cui domina
una Crocefissione, affresco attribuito al
Guidobono <A href="#(2)">(2)</A> (foto B).

%003%<tit>NOTE</tit>
[...]
#002# <nota><A name="#(2)">(2)</A>Nel 2000 l'affresco viene
ricollocato, dopo essere stato restaurato, sulla parete da dove era stato staccato
nel 1970 perché gravemente danneggiato. Il restauro dell'affresco si
deve all'Accademia di Medicina con il contributo della Fondazione CRT.
All'Università degli Studi di Torino si deve il restauro della cupola e

dello scalone. </nota>

Si badi che per esigenze delle scripts di preparazione del corpus la
stringa <A href="#_">_</A> non deve mai essere spezzata su due righe.

*** Evidenziazioni ***
corsivo <emph_i;bi>__</emph_i;bi>
I valori previsti sono corsivo normale “i” e grassetto corsivo “bi”.
[Si noti che, come per tutte le indicazioni con punto e virgola, i valori "i" e "bi" sono ovviamente alternativi: o corsivo normale o grassetto corsivo!]

grassetto <emph_b;bb>__</emph_b;bb>
I valori previsti sono corsivo normale “b” ed extra-bold “bb”

sottolineato <emph_u1;u2;u3>__</emph_u1;u2;u3>
I valori previsti sono singolo “u1”, doppio “u2” e triplo “u3”

tratteggiato <emph_h1;h2;h3>__</emph_h1;h2;h3>
I valori previsti sono singolo “h1”, doppio “h2” e triplo “h3”

puntinato <emph_d1;d2;d3>__</emph_d1;d2;d3>
I valori previsti sono singolo “d1”, doppio “d2” e triplo “d3”

maiuscoletto <emph_sc>__</emph_sc>
Il valore previsto è solo “sc” (*small capitals*). Le porzioni di testo in maiuscoletto vanno trascritte in normale tondo minuscolo e marcate <emph_sc>, a differenza delle parole in maiuscole (all caps) che restano semplicemente in stampatello.

espanso <emph_xp>__</emph_xp>
Il valore previsto è solo “xp” (expanded)

apice/pedice <emph_ap;pd>__</emph_ap;pd>
I valore previsti sono solo “ap” (apice) e “pd” (pedice)

negativo <emph_ng>__</emph_ng>
Il valore previsto è solo “ng” (negativo); usato per i bianchi su fondo grigio, ecc.

I codici sono combinabili solo con tags distinti, per cui un maiuscoletto grassetto corsivo spaziato sarebbe:
<emph_sc><emph_b><emph_i><emph_xp>_____</emph_xp></emph_i></emph_b></emph_sc>
(si noti il tipo di embricatura delle etichette, che non prevede "incroci").

indentature, centrature, ...
Eventuali indentature, in particolare il centrato (C) e l'allineato a destra (D), si marchino con le etichette:
<blank_C>___</blank>
<blank_D>___</blank>
L'allineato a sinistra, presente di default, non dovrà essere segnalato con nessun <blank>.

righe bianche
Le eventuali righe bianche vengono mantenute come tali; bisogna, ossia, porre tante righe bianche nella trascrizione quante ve ne erano nell'originale.

antroponimi,toponimi,enti/tà <anth><topn><oper><ent>

Come già in Valico, dalle cui Guidelines citiamo, «distinguiamo antroponimi (anth), toponimi (topn), tutti i nomi di creazioni artistiche, manufatti ed opere culturali in genere (oper), siano essi i Promessi sposi, Santa Maria Novella o la Gioconda, e tutti i nomi propri che non riguardano persone o animali (ent), siano essi marche di scarpe, di detersivi o nomi di alberghi:

<anth>__</anth>
<topn>__</topn>
<oper>__</oper>
<ent>__</ent> »

Anche gli indirizzi stradali in genere sono trattati come <topn>.
Particolare attenzione va posta alle etichette embricate, per es.: il <ent>Cotonificio <topn>Valle Susa</topn></ent>.

Un altro es. (tratto da JUS: lReg\lr-abr_00-001.htm):
&001& #___# <el>1.</el> Alla costituzione dell' intero patrimonio iniziale della <ent>Fondazione</ent> provvederà la <ent>Regione <topn>Abruzzo</topn></ent> . Esso sarà formato dall' immobile sito in <topn>Chieti Scalo</topn> – <topn>Viale Abruzzo 322</topn> - di proprietà della <ent>Regione <topn>Piemonte</topn></ent> con tutti i suoi arredi , strutture e pertinenze .

Si tenga inoltre conto dei seguenti casi specifici, molto frequenti all'interno de "La Valsusa":

- Suor(e), don, canonici, vescovi, monsignori e simili cariche ecclesiastiche si indichino sempre fuori dai tags, es.: consorella Suor <anth>Franca</anth>; don <anth>Luciano</anth>; Can. <anth>Barberis</anth>

- <ent>Padre Eterno</ent>, <ent>Dio</ent> (non <anth>, negli altri corpora del gruppo di ricerca era già così) nota invece: <anth>Redentore</anth>; <anth>Cristo</anth>

- <anth>San Francesco</anth>; <anth>Beato Rosaz</anth>

- <anth>Santo Curato d'Ars</anth> senza embricare il <topn> Ars

- "chiesa Santa Maria Maggiore" --> chiesa <oper>Santa Maria Maggiore</oper>

- Valle di Susa --> <topn>Valle di Susa</topn> (senza embricate anche "Susa" a parte)

- cattedrale di Susa --> <oper>cattedrale di <topn>Susa</topn></oper> (nota che qui invece si deve embricare)

- <ent>Casa di Riposo San Giacomo</ent> (senza embricare l'anth San Giacomo)

*** Colori diversi ***
L'uso intenzionale di colori diversi nel testo può essere rappresentato con il tag <col_red;green,...>__</col_red;green,...>.

formule matematiche <mat>
Le espressioni numerico-matematiche o comunque in cifre, ad esclusione dei semplici numerali "linguistici" espressi in cifre anziché in lettere e dei punti-elenco, sono contrassegnate con <mat>. Quindi avremo "voglio 15 giorni di vacanza" e "<el>1.</el>" senza marche ma "<mat>15 + 3 / 2 = 8</mat> è sbagliato" con marca.

**** ESCLUSI ****

Sono da escludersi dalla trascrizione:

- immagini;
- didascalie;
- testi pubblicitari;
- righe del tipo: "continua a pag. 27", che si evincono dalle indicazioni di pagina nella header e di girata nel BODY;
- numero del giornale, data e titolo quando si trascriva la testatina: se ne riporta solamente il titolo;
- nome e cognome dell'autore (o abbreviazione) a fine articolo, poichè il dato è già presente nella header sotto <aut_NC>.