Note per la costituzione e trascrizione del corpus di apprendenti VALICO.

Istruzioni/Guidelines v. 74 (17.06.03 - 27.01.05).

Manuel Barbera: manuel.barbera@bmanuel.org
Elisa Corino: elisa.corino@tin.it

0. Generalità.

0 Le seguenti "norme" specificano le modalità di raccolta, archiviazione e preparazione dei testi per un corpus internazionale di apprendenti ("Learner Corpus") di italiano: VALICO (**v**arietà di **a**pprendimento della **l**ingua **i**taliana: **c**orpus **o**nline). Esse sono nate, a partire da una prima proposta di C. Marello e M. Barbera, dalle discussioni emerse nella riunione (17.06.03) di tutti i partecipanti al progetto e dai primi esperimenti di trascrizione ed osservazioni di S. Ferraris, e poi dalle prime più estese applicazioni da parte di S. Camarca, F. Minozzi e V. Saggiotto; molto hanno beneficiato anche dell'infaticabile opera di coordinazione svolta da E. Corino, che dal Novembre 2003 ha affiancato M. Barbera nel "mantenimento" di queste *Guidelines*.

0.1 I destinatari ideali del presente documento sono i **fornitori** di testi (ossia le persone che hanno assegnato e raccolto le esercitazioni degli apprendenti destinate ad entrare nel corpus) ed i **trascrittori** (ossia le persone che trascriveranno manualmente i testi degli apprendenti trasformandoli in formato elettronico con i criteri necessari alla loro successiva elaborazione automatica). Le due figure potranno, al caso, anche coincidere.

0.2 Per i fornitori di testo sono previste due diverse possibilità.

0.2.1 Nell'ipotesi di maggiore coinvolgimento dovranno fornire:

(a) una **copia meccanica** degli originali, fotocopia se gli originali sono, come spesso succede, manoscritti, o copia su dischetto se gli originali sono stati elaborati direttamente su PC;
(b) una **header** (intestazione) per ogni testo, compilata il più accuratamente possibile secondo i criteri illustrati nel capitolo 1;
(c) uno **stelloncino** con le proprie generalità, anche istituzionali e scientifiche. Nell'ipotesi di maggior coinvolgimento nel progetto essi potranno anche occuparsi (direttamente o indirettamente) della trascrizione dei testi, secondo le norme illustrate nel capitolo 2.

0.2.2 Nell'ipotesi ipotesi minima, invece, i fornitori di testi dovranno produrre:

(a) una **copia meccanica** degli originali, fotocopia se gli originali sono, come più spesso, manoscritti, o copia su dischetto se gli originali sono stati elaborati direttamente su PC;
(b) una serie di **5 questionari** (studente, docente, scuola, esercizio e test) adeguatamente compilati. Questi verranno loro forniti su files o cartacei: cfr. Appendice 1.

0.3 La trascrizione dei testi, più nel dettaglio, dovrà essere prodotta in due copie, entrambe con la medesima header (intestazione): la trascrizione diplomatica (**TD**) e la trascrizione tokenizzata e markuppata (**TTM**).

Ogni documento, sia esso in TD o TTM, è sempre costituito dalla trascrizione del testo (cfr. § 2) preceduta dalla header (cfr. § 1); il fornitore / trascrittore avrà a disposizione un file di template (il modulo `template.txt`) da usare per la compilazione delle headers e/o la trascrizione dei testi.

Per la redazione dei documenti (headers, trascrizioni e stelloncini) si raccomanda di usare un semplice editor di testo (mai Word o WordPad!!) come NotePad, Edit Pro, VEdit, WinVi ecc. I documenti devono essere in formato `.txt` di Windows (e non `.doc` o `.rtf`!!), con codifica ANSI.

Ogni trascrizione, in definitiva, dovrà pervenire agli organizzatori del corpo sotto forma di 2 files `.txt` nominati secondo il sistema *`nometrascrittore`*`###_TTM~TD.txt,` come ad esempio:

`stefania001_TTM.txt` oppure `valeria002_TD.txt`.

0.3.1 La preparazione dei due file in questione è essenzialmente operazione manuale; i TTM, tuttavia, prima di assumere la forma definitiva, passeranno attraverso un formato di transizione, generato automaticamente, che sarà sommariamente descritto nel § 3.1 ed esemplificato in Appendice 3.

0.4.1 Queste prime Guidelines riguardano solo la preparazione dei documenti non annotati ("raw"), sia in versione diplomatica sia in versione tokenizzata e markuppata. I criteri per la cernita dei testi da raccogliere non sono qui pertinenti, ed anche l'allestimento di specifiche fasce di annotazione sarà operazione successiva.

0.5.1 Il formato finale del corpus sarà XML; il formalismo qui proposto non è veramente tale, ma è un formato più "facile da scrivere" per i trascrittori, calcolato tuttavia per poter essere agevolmente (ed automaticamente) convertito in legale XML in un secondo tempo.

1. Struttura e compilazione della header (intestazione).

1 Nella intestazione o “header” va specificato, prevalentemente da parte del fornitore dei testi, un certo numero di informazioni relative al testo ed alla sua produzione, organizzate in più gerarchie: caratteristiche del documento, caratteristiche del gruppo di testi di cui il documento fa parte, dati dell’autore e caratteristiche del testo, ecc. (cfr. § 1.1). I fornitori (e trascrittori) potranno copiare ed incollare su un “bastone” vuoto (cioè uno schema vuoto predisposto nel file `template.txt` in dotazione), inserendo opzioni. È infatti fondamentale che tutte le intestazioni siano formalmente standard. Una volta compilato, il file di intestazione va copiato e incollato in testa a entrambe le trascrizioni, sia TD che TTM (cfr. § 0.3 e 2.0), di ogni documento.

1.0 Ai fornitori e trascrittori verrà inoltre richiesto (come già accennato) di compilare alcuni files di informazioni contenenti i loro dati, i dati relativi all’istituzione in cui sono stati prodotti i testi ed i dati che concernono le caratteristiche della prova somministrata (cfr. §§ 4 e 5, Appendici 1 e 2). Nella versione pubblica del corpus saranno, naturalmente, introdotte misure per la tutela della privacy del fornitore, del trascrittore e dell’autore, ma nella versione base, disponibile solo in locale, è comunque importante avere anche queste informazioni.

1.1 Il modulo vuoto per l’immissione dei dati nella base di dati collegata al corpus (quello che noi chiamiamo il “bastone vuoto”) si presenta al modo seguente:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi;unicode</charset>
       <lingua>italiano</lingua>
       <aut_NC>(nome;?,cognome;?),(nome;?,cognome;?),…</aut_NC>
       <fornitore>(nome,cognome);ente</fornitore>
       <trascr>nome,cognome</trascr>
       <data>(aaaa;0;?,mm;0;?,gg;0;?),(0;?)</data>
       <luogo>città;?,nazione,?</luogo>
       <ist>ente;scuola;azienda;privato;0;?</ist>
       <ist_nome>____;0;?</ist_nome>
   </doc-id>
   <set-id>
       <corpus>____</corpus>
       <gruppo_num>1;2;…,g1;g5;gn</gruppo_num>
       <gruppo_nome>____;0</gruppo_nome>
   </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>
       <status>1;2;3;?</status>
       <annualita>1;2;3;4;+;?</annualita>
       <lingua1>____;?,____;0;?</lingua1>
       <lingue>____;0;?</lingue>
       <scolarizzazione>an;el;md;sp;un;?</scolarizzazione>
       <permanenza>(#mesi;0;?,luogo;0;?),(#mesi;0;?,luogo;0;?)</permanenza>
       <esposizione>sc,am,fam,med;?</esposizione>
   </autore>
   <autore2>ripeti_autore_o_canc</autore2>
   <autoreN>ripeti_autore_o_canc</autoreN>
   <testo>
       <tipo_forma>c-lib_var;c-lib_descr;c-lib_narr;c-lib_reg;c-lib_arg;c-art;
                   tes;dial;ques;es-trad;dett;rias;email;lett</tipo_forma>
       <tipo_produzione>did;priv;lav;?</tipo_produzione>
       <topics>...</topics>
       <keyw>(____,____,____,____,____);?</keyw>
       <test>____;0;?</test>
       <qualita>orig;origFC;origCE;copia</qualita>
       <esecuzione>or;ms;wp;kw</esecuzione>
       <cap-min>tc;tm;0</cap-min>
   </testo>
   <ref>
       <stel>nome_F.txt;0,nome_T.txt;0,nome_G.txt;0,nome_P.txt;0</stel>
       <cons>nome_C.txt;0</cons>
       <txtext>nome1_R.txt;0,nome2_R.txt;0</txtext>
       <imgext>nome1_R.jpg;0,nome2_R.jpg;0</imgext>
       <txtint>nome1.txt;0,nome2.txt;0</txtint>
       <imgint>nome1.jpg;0,nome2.jpg;0</imgint>
   </ref>
</HEAD>
```

1.1.1 Si noti che all'interno di un `<tag>` non devono esserci spazi tra il `<tag>` e la parola adiacente, quindi si avrà, ad esempio,

```
<permanenza>11,Livorno Ferraris,7,Saluggia</permanenza>
```

I connettori, d'altra parte, sono limitati a due, la "and" (`,`) e la "or" (`;`), più la parentesi.
Immediatamente dopo la header, contenuta nel tag `<BODY>_</BODY>`, inizia poi la trascrizione del testo (cfr. infra, capitolo 2).

1.2 Qui sotto commentiamo dettagliatamente ogni attributo e valore della header seguendo la struttura del bastone vuoto di modello.

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

1.2.1.1 **<idN>** Numero progressivo che sarà l'identificativo assoluto del documento. Va lasciato vuoto tanto dai fornitori quanto dai trascrittori: saranno poi gli allestitori del corpus a saturare il campo.

1.2.1.2 **<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, ed `unicode`, da usare solo per i testi che presentino caratteri non-latini; il valore di default è ovviamente `ansi`. Per maggiori dettagli cfr. il § 2.0.3 del capitolo sui criteri di trascrizione.

1.2.1.3 **<lingua>** Di default è l'italiano. Il valore è previsto solo per la futura interrogazione del Corpus di Apprendenti insieme ad altri corpora non sempre / solo di lingua italiana.

1.2.1.4 **<aut_NC>** Nome del 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'apprendente fossero solo imperfettamente note. (cfr. Appendice 1).
Sono anche previsti i casi in cui gli autori siano più di uno, anche se l'eventualità non è molto probabile. In questo caso si useranno le parentesi e si attiveranno le gerarchie <autore1> ... <autoreN> per fornire i dati di ogni autore (cfr § 1.2.4)

1.2.1.5 **<fornitore>** Nome della persona che ha materialmente raccolto il testo; in questo campo bastano nome e cognome (con i criteri di cui sopra), ma ogni fornitore di testi dovrà compilare uno **stelloncino a parte** con le proprie generalità, anche istituzionali e scientifiche (cfr. Appendice 2 § 5.1). Il "nome" di tale stelloncino dato dal "nomecognome" del fornitore accompagnato dalla sigla `F`, ed il suo formato sarà lo stesso `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per Tanya Roy avremo

```
tanyaroy_F.txt.
```

È anche possibile (anche se non auspicabile) che un gruppo di documenti non ci pervenga da una persona determinata, ma da un qualche ufficio o struttura amministrativa "non personale": in questo caso si userà il valore `ente`.

1.2.1.6 **<trascr>** Nome della persona che ha materialmente trascritto il testo, nel caso che questa sia distinta da chi lo ha raccolto; anche in questo campo bastano nome e cognome (con i criteri di cui sopra), ed alla stessa maniera ogni trascrittore dovrà compilare lo **stelloncino a parte** (cfr. Appendice 2 § 5.2). Analogamente, il "nome" dello stelloncino sarà dato dal "nomecognome" del trascrittore accompagnato dalla sigla `T`, ed il suo formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per Francesca Minozzi avremo

```
francescaminozzi_T.txt.
```

Nel caso che fornitore e trascrittore coincidano, l'indicazione sarà ripetuta più volte, e la sigla nel nome del file sarà `FT`, ad es.

```
silviacamarca_FT.txt.
```

1.2.1.7 **<data>** Data di produzione del testo, espressa secondo il sistema `aaaa,mm,gg` saturabile da valori numerici o da quello non definito (`?`), ad es. "14 febbraio 2001" sarà `2001,02,14`, "Dicembre 1999" sarà `1999,12,?`, estate 2003 sarà `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.

1.2.1.8 **<luogo>** Luogo di produzione del testo. Sono specificati due valori: la `città` o paese in cui il testo è prodotto e la `nazione` cui appartiene, espressa nelle convenzionali sigle internazionali scritte

in maiuscolo (quindi avremo IT per 'Italia', DE per 'Germania', IN per 'India', ecc.); sono previsti anche i valori non definiti (?). Ess.

```
<luogo>Cusano Milanino,IT</luogo>.
<luogo>Madras,IN</luogo>.
<luogo>?,HU</luogo>.
```

1.2.1.8.1 Le sigle sono quelle standard ISO usate per le estensioni TLD dei domini web internazionali; il sistema (aggiornato al 2 aprile 2002) con le sue 239 entità ricopre sostanzialmente tutti gli stati del mondo (con minime eccezioni, relative a regioni geografiche e situazioni politiche particolari). Una lista completa, se ve ne fosse bisogno, è disponibile anche sul nostro sito:

```
http://www.bmanuel.org.courses/tld.html
```

1.2.1.9 **<ist>** Tipo di istituzione nella quale è stato prodotto il testo; sono previsti anche il valore nullo (0), il valore non definito (?) ed il valore (privato) nel caso non sia coinvolta alcuna istituzione.

1.2.1.10 **<ist_nome>** Nome dell'istituzione presso o per la quale è stato prodotto il testo; sono previsti anche il valore nullo (0) e il valore non definito (?).

1.2.1.10.1 Bisognerà poi indicare su uno **stelloncino** a parte in max 720 battute (cioè c. 8 righe di 90 battute) le generalità e caratteristiche dell'istituzione (cfr. Appendice 2 § 5.3). Ogni stelloncino dovrà essere posto in un file separato, avente per nome una forma sintetica del nome dell'istituzione medesima accompagnato dalla sigla I, ed il suo formato sarà il solito .txt degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per i documenti <ist>Delhi-BA</ist> si avrà il file Delhi-BA_I.txt dal seguente contenuto:

```
Delhi University
New Delhi - India
Department of Germanic and Romance Studies. Italian studies
Grado universitario
Extra info: Bachelor of Arts (Honours) in Italian. Questo è
un corso di laurea triennale in italiano. All'anno gli stu-
denti devono superare tre corsi in lingua e due sulla cul-
tura europea in inglese.
```

1.2.2 <***set-id***> Informazioni che serviranno ad identificare gli insiemi di testi da cui il documento proviene ("gruppo") ed in cui confluirà ("corpus"):

1.2.2.1 **<corpus>** Di default il valore da attribuire sarà VALICO.

1.2.2.2 **<gruppo_num>** Esercizi con consegna uguale: numerazione. In questo campo è necessario specificare due valori separati dalla virgola: il numero assoluto dell'esercizio (1;2;…,) (dove 1 sarà tanto il primo di una serie quanto l'esercizio unico), e la consistenza del gruppo, dove sono previsti soli tre valori, (g1) per l'esemplare unico, (g5) per gruppetti inferiori a cinque e (gn) per gruppi con più di cinque esemplari. Ad esempio:

<1,g1> "esercizio unico (esemplare 1 di gruppo di 1)"
<1,g5> "primo esercizio di gruppo con meno di 5 esemplari"
<3,g5> "terzo esercizio di gruppo con meno di 5 esemplari"
<7,gn> "settimo esercizio di gruppo con più di 5 esemplari"

1.2.2.3 **<gruppo_nome>** Esercizi con consegna uguale: denominazione. In questo campo va inserito un nome che funga da identificativo per ogni gruppo di esercizi; in alternativa (per esercizi unici) è previsto il valore nullo (0). Si noti che non è necessario inserire tutta o parte della consegna: è sufficiente un nome convenzionale, possibilmente breve e originale, che permetta di riconoscere univocamente il gruppo al quale si fa riferimento, ad es.

```
<gruppo_nome>rane</gruppo_nome>
<gruppo_nome>miamadre</gruppo_nome>.
```

1.2.2.4 Nel caso di gruppi di esercizi, bisognerà poi indicare su uno **stelloncino** a parte in max 900 battute (cioè c.10 righe di 90 battute) le caratteristiche dell'esercizio (in alternativa il fornitore potrà compilare il "questionario esercizio", cfr. Appendice 2 § 5.4, lasciando ai trascrittori il compito di ricavarne lo stelloncino appropriato). Ogni stelloncino dovrà essere posto in un file separato, avente per nome lo stesso nome assegnato al gruppo accompagnato dalla sigla G, ed il suo formato sarà il solito .txt degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per il gruppo rane si avrà un file dal nome

```
rane_G.txt.
```

e dal contenuto seguente (che può fungere da modello anche per l'organizzazione dell'informazione al suo interno):

```
Consegna: descrivi le figure di una storia a disegni su un
bambino alla ricerca della sua rana fuggita da casa per u-
nirsi ai suoi simili nello stagno (in Berman - Slobin, Re-
lating events in narrative: a crosslinguistic developmental
study, Hillsdale, Lawrence Erlbaum Associates, 1994, 647-
654).
Scopo: verificare l'estensione del lessico dell'apprenden-
te, la sua capacità di usare i tempi verbali e i connettivi
testuali, coerenza e coesione del testo.
Contesto: esercitazione in classe.
Extra info: agli studenti non sono stati suggeriti vocaboli
o strutture morfosintattiche.
```

1.2.2.4.1 Si noti che la "descrizione del gruppo" fornita negli stelloncini `*_G.txt` non è coincidente con la "riproduzione integrale della consegna" fornita negli stelloncini `*_C.txt` (cfr. § 1.2.6.2): per un gruppo di documenti potremmo infatti non avere a disposizione la consegna originaria, così come potremmo invece disporre della consegna originaria per un documento singolo. Le informazioni contenute nei due stelloncini non sono inoltre coincidenti: in uno si fornisce una descrizione "dall'esterno" dell'esercizio, nell'altro si riproduce quanto effettivamente consegnato agli apprendenti.

1.2.3 <***autore***> Informazioni sul produttore del testo.

1.2.3.1 <**specifiche**> Informazioni (`specifiche`) sul sesso del produttore del testo, maschile o femminile o non definito (`?`), se si tratta di individuo, altrimenti si specifica 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`).

1.2.3.2 <**eta**> Sono previste sette fasce di età (`1-7,8-13,14-18,19-25,26-30,30-40,40-50,oltre`) oltre al valore non definito (`?`).

1.2.3.3 <**status**> Status sociale, in base al reddito: modesto (`1`), medio (`2`), alto (`3`), non definito (`?`).

1.2.3.4 <**annualita**> L'anno di scolarità in italiano; sono previsti quattro valori (`1,2,3,4,+`) oltre al non definito (`?`); il valore (`+`) è per qualsiasi scolarità superiore a quattro.

1.2.3.5 <**lingua1**> Informazioni sulla lingua di partenza dell'apprendente. Sono previsti due campi cui attribuire un valore. Nel primo si fornisce il nome della lingua madre vera e propria; nel caso non sia nota è previsto anche il valore non definito (`?`). Nel secondo si indica la L1 veicolare se diversa dalla lingua madre (come ad es. avviene spesso in India, o come nel caso dell'arabo letterario rispetto agli arabi "volgari", ecc.), altrimenti si pone il valore nullo (`0`: lingua madre e L1 coincidono) od alla peggio non definito (`?`).

1.2.3.6 <**lingue**> Informazioni sulle altre lingue note all'apprendente; sono previsti anche il valore nullo (`0`) e il valore non definito (`?`). Se particolare cura va rivolta a che i valori dell'attributo precedente `<lingua1>` siano accurati, spesso non si può purtroppo entrare molto nel merito sul grado di conoscenza delle altre lingue: la proposta sarebbe di inserire tutte le lingue che l'apprendente dichiara di conoscere (quindi anche le lingue con conoscenza scolastica e non solo le L2 effettive) in supposto ordine decrescente di conoscenza. Si noti, tra l'altro, che l'italiano, dato che è la lingua che tutti gli autori stanno studiando, non viene mai indicata: è già presupposta di default.

1.2.3.7 <**scolarizzazione**> La scolarizzazione di partenza: analfabeta (`an`), elementare (`el`), media (`md`), superiore (`sp`), universitaria (`un`), non definita (`?`).

1.2.3.8 <**permanenza**> Quantificazione e localizzazione dei soggiorni degli apprendenti in territorio italofono: sono previsti due campi associati per il numero dei mesi e la località (`#mesi,luogo`; ad es. "`7,Rho`"), ripetibili liberamente (ad es. "`(3,Perugia),(2,Torino)`"), oppure usabili genericamente (ad es. "`1,Italia`", come può essere il caso per certe vacanze turistiche). Il valore nullo (`0`) in entrambi i campi sarà, naturalmente, da assegnare in documentata assenza di qualsiasi soggiorno in territorio italofono. In completa mancanza di informazioni in proposito sono, invece, sempre previsti i valori non definiti (`?`).

1.2.3.9 <**esposizione**> Il tipo di esposizione alla lingua italiana che il produttore del testo ha avuto. Si possono esprimere anche più valori contemporaneamente, in alternativa al sempre previsto valore non definito (`?`) se si è privi di informazioni in proposito. I valori finora previsti sono: scuola (`sc`), amici (`am`), famiglia (`fam`), media (`med`).

1.2.4 <***autore1***> ... <***autoreN***> Nel caso (prevedibilmente poco frequente) in cui siano stati posti più autori come valore del campo <aut_NC>, i loro dati andranno forniti in tante gerarchie quanti, appunto, gli autori, e la struttura interna di ogni gerarchia riprodurrà quella di <autore>, avremo quindi (per usare un esempio di fantasia):

```
<doc-id>
    […]
    <aut_NC>(Gwynfor,Dwyryd),(Siân,Llewellyn)</aut_NC>
    […]
<autore>
    <specifiche>m</specifiche>
    <eta>19-25</eta>
    <status>1</status>
    <annualita>3</annualita>
    <lingua1>gallese,inglese</lingua1>
    <lingue>?</lingue>
    <scolarizzazione>un</scolarizzazione>
    <permanenza>3,Locate Triulzi</permanenza>
    <esposizione>sc,am,fam</esposizione>
</autore>
<autore2>
    <specifiche>f</specifiche>
    <eta>19-25</eta>
    <status>2</status>
    <annualita>3</annualita>
    <lingua1>gallese,inglese</lingua1>
    <lingue>?</lingue>
    <scolarizzazione>un</scolarizzazione>
    <permanenza>2,Cava Manara</permanenza>
    <esposizione>sc,am,fam</esposizione>
</autore2>
    […]
```

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

1.2.5.1 <**tipo_forma**> Tipo "formale" di testo: libera composizione rispettivamente 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), composizione in forma di articolo di giornale (c-art), tesina (tes), testo dialogico scritto da una persona singola (dial), questionari liberi e "comprehension" (ques), esercizio di traduzione (es-trad), dettato (dett), riassunto (rias), lettera elettronica (email) o tradizionale (lett). Sono escluse le traduzioni dall'italiano (perché il corpus è di italiano), i questionari con risposte obbligate o troppo brevi per essere di alcuna rilevanza linguistica, e, per analoghe ragioni, i cloze; non sono, almeno in questa prima fase, previsti i dettati.

1.2.5.2 <**tipo_produzione**> Tipo di condizioni nel quale il testo è stato prodotto: nell'attività didattica (did), privatamente (priv) o nel quadro dell'attività lavorativa (lav). È stato previsto anche il valore indefinito (?) ma non quello nullo.

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

1.2.5.4 <**keyw**> Per le medesime ragioni si possono indicare alcune keywords che aiutino ad individuare l'argomento del documento; il numero di queste è fissato a 5, ma è stato previsto anche il valore indefinito (?), nel caso il documento non abbia un singolo e/o preciso argomento. In questa prima fase anche questo campo viene semplicemente ignorato e verrà completato dai curatori del corpus in un momento successivo alla trascrizione dei documenti.

1.2.5.5 <**test**> Qui va inoltre specificato se l'elaborato, quale che ne sia il tipo, è una prova di esame di fine anno o una prova in itinere. In tal caso il raccoglitore userà una formulazione riconoscibile nel sistema scolastico del paese, ad es.

`<test>3d Degree</test>`.

1.2.5.5.1 In uno **stelloncino a parte** illustrerà poi tale dicitura, chiarendo anche le condizioni di svolgimento della prova (tempo dato, possibilità di consultare dizionari monolingui o bilingui di italiano o altri testi di riferimento; cfr. Appendice 2 § 5.5). Il "nome" dello stelloncino sarà dato da una forma convenzionalmente abbreviata del nome della prova accompagnato dalla sigla P, ed il suo formato sarà il solito .txt degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3), ad es.

3dDegree_P.txt.

È previsto anche il valore nullo (`0`), se il documento prodotto non è una prova, ed il non definito (`?`), se semplicemente l'informazione non è nota al raccoglitore.

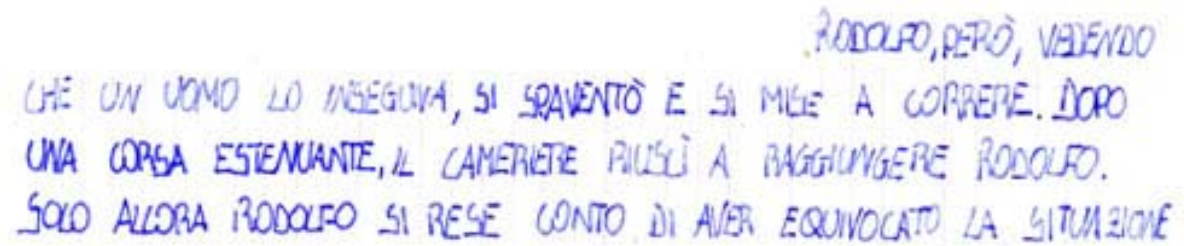
1.2.5.6 <**qualita**> La natura dell'**antigrafo** del testo trascritto: si tratta dell'originale prodotto dall'apprendente, materialmente (`orig`) od in fotocopia (`origFC`) od in copia elettronica (`origCE`), o piuttosto di una sua copia indiretta, già digitata dal raccoglitore o da chi per esso (`copia`)

1.2.5.6.1 Nel caso di e-mails si usa il valore (`orig`) quando l'antigrafo è la mail originaria, direttamente estratta dal mail reader, si usa invece (`origCE`) quando l'antigrafo è già una conversione dal formato originario del mailer, con eventuale perdita di informazioni (headers, fini riga, ecc.) – cfr. anche § 2.1.1.3.

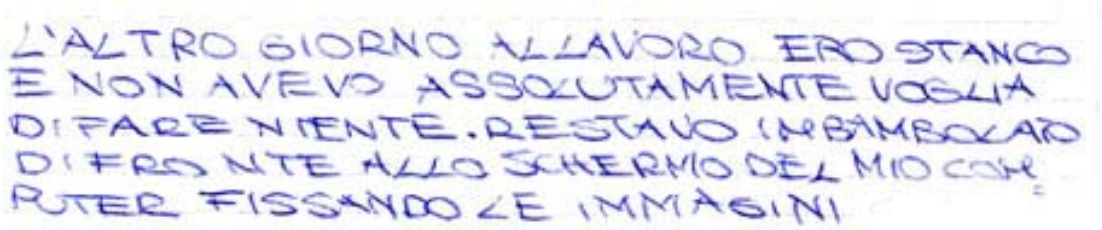
1.2.5.7 <**esecuzione**> Il modo di produzione materiale del testo: se orale (`or`), manoscritto (`ms`), scritto al computer con un programma di videoscrittura (`wp`), o dattiloscritto (`kw`). I materiali che prevediamo di avere sono tutti scritti (prevalentemente manoscritti), ma si è voluto lasciare una finestra aperta per eventuali materiali orali che fosse dato di raccogliere.

1.2.5.8 <**cap-min**> Il sistema ortografico normale delle lingue scritte in latinica quali l'italiano prevede la normale alternanza di due set di grafi: **capitali** ("ABCABC") e **minuscoli** ("abc*abc*"). Può capitare che singoli scriventi uniformino la propria ortografia ad uno solo dei set, scrivendo tutto in grafi esclusivamente attinti al canone capitale (meno frequente il contrario). Si tratta di una caratteristica da distinguere dall'uso specifico per singole porzioni di testo (singole parole o frasi) del maiuscolo (trascritto come tale) o del maiuscoletto (cfr. le marche di evidenziazione, §2.4.7.1). Per evitare di trascrivere testi intieri in capitali (appesantendo inutilmente il formario del POS-tagger) si è scelto di marcare tale caratteristica nella header e poi trascrivere il testo in normale minuscola (con eventuale ricorso a maiuscole per marcare cambi di corpo, anche se non di canone, del carattere). I valori previsti sono pertanto: il valore nullo (`0`) per l'uso normale, il valore (`tc`) per i testi tutti in capitali, ed il valore (`tm`) per quelli tutti in minuscole. Si vedano

a



b



gli ess. *a* e *b*, cui va attribuito `<cap-min>tc</xap-min>`, e la loro trascrizione TD:

```
a   Rodolfo, però, vedendo
    che un uomo lo inseguiva, si spaventò e si mise a correre. Dopo
    una corsa estenuante, il cameriere riuscì a raggiungere Rodolfo.
    Solo allora Rodolfo si rese conto di avere equivocato la situazione
b   l'altro giorno al lavoro ero stanco
    e non avevo assolutamente voglia
    di fare niente. restavo imbambolato
    di fronte allo schermo del mio com|puter
    fissando le immagini
```

1.2.6 <***ref***> I links, o riferimenti ipertestuali (href), istituiti dal e nel documento, intendendo con ciò tanto i riferimenti esterni chiesti dalla header (stelloncini, ecc.), tanto i riferimenti interni ad immagini od allegati testuali contenuti nel testo.

1.2.6.1 <**stel**> Devono essere indicati, nella corretta sequenza, i nomi degli stelloncini richiesti dal documento in questione (nell'ordine: fornitore, trascrittore, gruppo, prova) con i nomi che sono stati descritti nei §§ 1.2.1.5, 1.2.1.6, 1.2.2.3 e 1.2.5.4.

```
<stel>tanyaroy_R.txt,francescaminozzi_T.txt,sogno_G.txt,3dDegree_P.txt.</stel>
```

Oltre alla specifica dei nomi è previsto naturalmente anche il valore nullo (`0`).

1.2.6.2 <**cons**> In condizioni ideali, oltre agli elaborati degli apprendenti, si dovrebbe acquisire anche la consegna materialmente assegnata dal docente. In tal caso questa va trascritta integralmente (secondo i criteri della TD) su file separato, il cui "nome" sarà dato da una forma convenzionalmente abbreviata del titolo della consegna (perlopiù la stessa del nome del gruppo, quando presente: cfr. § 1.2.2.3) accompagnato dalla sigla `C`, ed il suo formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3), ad es. la consegna dei documenti del gruppo `storia`, indicata al modo sg.

```
<cons>storia_C.txt</cons>
```

punterà al file `storia_C.txt` che riproduce nella sua interezza la consegna originale, cioè:

```
Continua la storia:
Era una notte buia e tempestosa, il vento soffiava tra le cime degli alberi e la pioggia battente scrosciava tra le fronde. Geppino e Mariolina erano perduti, non avrebbero mai più trovato la strada di casa, ma ecco che ad un tratto …
```

Oltre alla specifica del nome è previsto naturalmente anche il valore nullo (`0`).

1.2.6.2.1 Per la differenza tra "consegna" e "gruppo", così come tra stelloncini `*_G.txt` e stelloncini `*_C.txt` cfr. § 1.2.2.3-4.

1.2.6.3 <**txtext**>Nella consegna si può fare riferimento a testi esterni che siano stati letti in classe (dettati, traduzioni, esercizi di *comprehension*, sono necessariamente basati su un testo esterno). Qualora ne fossimo in possesso, di questi va fornita o la trascrizione (per brani di pubblico dominio) od il rinvio bibliografico (per testi estesi, facilmente reperibili, o coperti da copyright). Valgono le solite avvertenze sul nome e formato del file, vale a dire che il "nome" sarà dato da una forma convenzionalmente abbreviata del titolo del brano di riferimento (eventualmente il medesimo del nome del gruppo e/o della consegna: cfr. §§ 1.2.2.3 e 1.2.6.2) accompagnato dalla sigla `R`, ed il suo formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Possono essere indicati anche più files (separati tra loro dalla virgola) o nessuno (`0`). Ad esempio, il testo di riferimento per i documenti del gruppo `tartari`, indicato al modo sg.

```
<txtext>tartari_R.txt</txtext>
```

punterà al file `tartari.txt` che conterrà il testo:

```
Dino Buzzati, Il deserto dei Tartari, Milano, A. Mondadori, 1979, 6a ed. - cap.4
```

1.2.6.4 <**imgext**> Analogamente nella consegna si può fare riferimento ad immagini esterne che siano state usate come base per l'esercitazione. Si tratterà in questo caso di files di immagine, scannate di solito in `.jpg`, il cui "nome" sarà dato da un titolo convenzionale (eventualmente il medesimo del nome del gruppo e/o della consegna: cfr. §§ 1.2.2.3 e 1.2.6.2) accompagnato dalla sigla `R`.
Ad es., l'immagine di riferimento per i documenti del gruppo `pescatore`, indicata al modo sg.

```
<imgext>br-g=pogopesca_R.jpg</imgext>
```

punterà al file `br-g=pogopesca_R.jpg` che conterrà l'immagine :

Possono essere indicati anche più files (separati tra loro dalla virgola) o nessuno (`0`).

1.2.6.5 **<txtint>** Nel testo possono essere compresi allegati di natura testuale (i.e. ritagli di giornale, ecc.). In tal caso questi saranno trascritti integralmente in files separati secondo i criteri della TD, i cui "nomi" saranno dati da un titolo convenzionalmente abbreviato (a volte lo stesso del nome del gruppo: cfr. § 1.2.2.3), ed il cui formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3).

Ad es., il testo allegato nei documenti del gruppo `annuncioVecchia`, indicato al modo sg.

```
<txtint>annuncioVecchia.txt</txtint>
```

punterà al file `annuncioVecchia.txt` che conterrà il testo:

```
AAA vecchia multimiliardaria residente ad Acapulco cerca gio-
vane aitante bella presenza e fisico prestante per assaporare
ultimi istanti di vita. Si promette una morte a breve termine
coronata da cospicua eredità. Chiedere di tota Bina.
```

Potranno naturalmente essere indicati più files (separati tra loro dalla virgola) o nessuno (`0`).

1.2.6.5.1 Si noti che i nomi dei files di riferimento interni non presentano sigle, a differenza degli esterni che ne erano sempre contrassegnati.

1.2.6.6 **<imgint>** Nel testo possono essere compresi materiali di tipo grafico, come disegni o schizzi dell'autore (cfr. il commento al tag `img` infra § 2.4.9 e sgg.). Si tratterà in questi casi di files di immagine, scannate di solito in `.txt`, il cui "nome" sarà dato da un titolo convenzionale.

Ad es., l'immagine allegata in un documento del gruppo `mipresento`, indicata al modo sg.

```
<imgint>br-g=omino4.jpg</imgint>
```

punterà al file `br-g=omino4.jpg` che conterrà l'immagine:

Potranno naturalmente essere indicati più files (separati tra loro dalla virgola) o nessuno (`0`).

1.2.6.6.1 Si noti sempre che i nomi dei files di riferimento interni non presentano sigle, a differenza degli esterni che ne erano sempre contrassegnati.

2. Criteri di trascrizione.

Il trascrittore produrrà pertanto una prima versione, praticamente diplomatica [TD], e poi, a partire da quella, una seconda versione [TTM] in cui la trascrizione sarà sottoposta ad una appropriata tokenizzazione e corredata da un markup testuale comprendente anche alcune categorie di pre-tagging. Entrambe saranno inserite, immediatamente dopo la `<HEAD>`, nel tag `<BODY>_</BODY>.`

I generali **requisiti dei files** così prodotti sono già stati indicati, ma li ricapitoliamo brevemente (§ 0) prima di scendere più nel dettaglio delle **norme per la trascrizione** vere e proprie (§ 1-6):

2.0. ***Files.***

Uno per ogni trascrizione (e quindi due per testo).

2.0.1 Il **formato** deve essere esclusivamente `.txt` e deve essere prodotto da un editor di testo semplice come NotePad / BloccoNote di Windows , EditPro, VEdit ecc,. - mai comunque, con Word, Write / Wordpad, o qualsiasi altro programma che rischi di sporcare il puro testo con codici di formattazione.

2.0.2 I **nomi** dei files `.txt` dovranno essere costruiti secondo il sistema

`nometrascrittore``###_TTM~TD.txt`

quindi si avrà, ad esempio, `stefania001_TTM.txt` o `valeria002_TD.txt`.

Si badi che la numerazione procede in un'unica serie continua per ogni trascrittore, indipendentemente dagli eventuali gruppi cui il documento appartenga.

2.0.2.1 I files di Header prodotti dai fornitori dovranno essere analogamente strutturati:

`nomefornitore``###_HD.txt`

quindi si avrà, ad esempio, `tanya001_HD.txt`. Il fornitore dovrà altresì apporre il medesimo contrassegno (proprio nome + numero sequenziale di documento) sulle copie cartacee o sui dischetti (cfr. sopra la discussione del tag della Header `<qualita>`, § 1.2.5.5) che produrrà.

2.0.2.2 Files passati attraverso una trafila completa (in cui, ossia, fornitore e trascrittore non coincidono) avranno pertanto la struttura:

`nomefornitore_nometrascrittore-``###_TTM~TD.txt`

2.0.3 Il **character set** di base sarà quello ANSI base di Windows (praticamente coincidente con l' ASCII ISO 8859-1 Latin 1 universalmente corrente anche in Unix, e diverso dal vecchio set ASCII di DOS, che era l' ISO 646-RV a 7 bit), i cui codici fuori tastiera ("over-122") si ottengono digitando `alt+0+#cod` sul tastierino numerico (una comoda e opportunamente stampabile lista dei codici carattere è facilmente accessibile in rete alla pagina `http://www.netstrider.com/tutorials/HTMLRef/ASCII/`).

2.0.3.1 C'è possibilità di ricorso all'Unicode per caratteri non latini eventualmente presenti nei testi.

2.0.3.2 Non devono mai essere usati i caratteri doppi, tipo (*`e'`*) per i semplici (*`è`*), in quanto creerebbero incoerenze e problemi di riconoscimento da parte del software.

2.1. ***Layout.***

È rispettato il più possibile l'originario.

2.1.1 Le **righe** dell'originale sono mantenute con le stesse andate a capo; ossia si andrà a capo, senza ulteriori contrassegni, quando e solo quando anche l'originale vada a capo. [TD+TTM].

2.1.1.1 Le parole divise nell'**accapo** sono riportate alla riga iniziale (compresi gli eventuali segni di interpunzione attaccativi); il punto di divisione è segnato con il diacritico | [ANSI 0124]. [TD+TTM]. Esempi.

```
ac|capo [TD+TTM]
tranqui|lla). [TD]
tranqui|lla ) . [TTM]
```

2.1.1.2 Le eventuali **righe bianche** vengono mantenute come tali; bisogna, ossia, porre tante righe bianche nella trascrizione quante ve ne erano nell'originale. [TD+TTM].

2.1.1.3 Nei testi degli **e-mail**, qualora le originali fine-linea non fossero state conservate nel trasferimento dal formato iniziale (.eml, di solito) a quello di archiviazione (.txt), si impone arbitrariamente l'accapo entro la 60a battuta (che è il valore medio più diffuso nei mail reader). Cfr. anche § 1.2.5.6.1. [TD+TTM].

2.1.2 Eventuali spazi bianchi a sinistra (**indentature**) od al centro della riga vanno riprodotti in TD con altrettanti spazi bianchi, mentre in TTM vanno indicati con il tag `<blank>` e per valore il numero approssimativo di parole che lo spazio occupa. Ad esempio l'inizio di una lettera sarà risolto così:

```
Cara Amalasunta,
                grazie del pacco di brigidini
che mi hai                mandato.  [TD].
Cara Amalasunta ,
<blank_2> grazie del pacco di brigidini .</blank>
che mi hai <blank_2></blank>mandato .  [TTM].
```

2.1.2.1 Si noti che il tag se aperto ad inizio riga (e chiuso non all'interno della riga medesima, ma alla fine della stessa o di più righe) si riferisce a tutte le righe che vi sono incluse, per cui, nel caso di indentature continuate (come talvolta in dialoghi e questionari) basta aprirlo e chiuderlo all'inizio di ogni blocco indentato), come nell'esempio seguente (dove si prescinde dalla marca per turno, che sarebbe necessaria, per chiarezza, in quanto viene spiegata in séguito, cfr. 5.4):

-
```
Commesso : Buongiorno Signor , cosa potrei fare
           per lei?
Io : 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. [TD]
```
-
```
Commesso : Buongiorno Signor , cosa potrei fare
<blank_2> per lei ? </blank>
Io : Oggi è il compleanno di mia amica .
<blank_1> Ho preparato una torta buona ma il
mio cane la ha mangiata e devo
procurarla da qualche mezzi . </blank> [TTM]
```

2.1.2.2 Nel caso in cui il margine sinistro sia irregolare, ma non presenti evidenti indentature o spazi bianchi intenzionali, non è necessario riprodurre il layout originale. [TD+TTM]

2.1.3 Le **pagine** vengono marcate con un `$001$` ecc. all'inizio di ogni pagina. [TD+TTM]

2.1.4 Eventuali **capitoli** o **paragrafi** vanno numerati al loro inizio con `%001%` ecc. per i capitoli, e `#001#` ecc. per i paragrafi. [TD+TTM].

2.1.4.1 La definizione di "capitolo" e "paragrafo" è delicata in quanto il "livello" deve potere essere confrontabile con quello di testi "normali" (pena la inconfrontabilità a livello testuale con altri corpora), nonostante la diversa specificità della maggior parte dei testi del Learner Corpus. Se la maggior parte degli elaborati che saranno raccolti di norma non conterrà veri e propri "capitoli" (che potrebbero essere presenti solo in tesine), l'identificazione del livello inferiore (paragrafo) sarà invece spesso necessaria anche se delicata; il punto a capo, comunque, nella nostra prospettiva, rappresenta un livello di organizzazione del testo ancora più basso di quello del paragrafo, e non deve normalmente esser fatto con esso coincidere. Per identificare un paragrafo proponiamo due regole pratiche (senza alcuna pretesa teorica) che si applicano a catena:

(1) *formalmente* saranno paragrafi distinti solo blocchi di testo di una certa estensione chiaramente individuati oltre che da un punto a capo anche da linee bianche od altri espliciti segnali grafici demarcativi (tratti di penna, indentature, ecc.).

(2) *testualmente* possono essere considerati paragrafi distinti blocchi testuali formalmente delimitati da un mero punto a capo se e solo se sono di considerevole estensione e semanticamente presentano un notevole e indiscutibile e consistente cambio di argomento.

2.1.5 Gli **elenchi puntati** saranno contrassegnati con il tag `<el>`, il cui uso è indipendente da (ed a sua volta combinabile con) i tags di capitolo (cfr. § 2.1.4), paragrafo (cfr. § 2.1.4) e titolo (cfr. § 2.5.1.1). [TTM]

Avremo quindi

```
La lista della spesa:
<el>1.</el> pane
<el>2.</el> latte
<el>3.</el> giornale
```

2.2. ***Ortografia e processi correttori.***

Si conserva sempre l'ortografia dell'autore (ma con le precisazioni di cui oltre).

2.2.1 L'uso delle **maiuscole** e minuscole va mantenuto come è. [TD+TTM].

2.2.1.1 Si badi che altra cosa sono le *maiuscole* ("capitals": MAIUSCOLO) dal *maiuscoletto* ("small caps": MAIUSCOLETTO), che viene qui trattato con uno dei tags di `<emph>` (cfr. § 2.4.7.1)

2.2.2 L'**accento**, di solito non distinguibile nella scrittura manuale, è riportato di default all'uso standard. Eventuali casi di autori che distinguono sistematicamente tra acuto e grave saranno risolti se e quando compariranno. [TD+TTM].

2.2.3 Quanto alla stratigrafia delle **correzioni**, così come alle **inserzioni** correttive, intendendosi con queste gli interventi dello scrivente sul proprio testo, il testo trascritto rispecchia sempre l'ultima correzione introdotta. [TD+TTM].

2.2.3.1 Le **correzioni**, ossia le lezioni scartate, possono tuttavia essere interessanti linguisticamente, e vanno pertanto riportate con due diversi sistemi, uno basato sulle parentesi graffe `{` (ANSI 0123) e (ANSI 0125) `}` in TD, ed uno basato sul tag `<CORR>` in TTM. In entrambi i casi le lezioni scartate devono sempre seguire l'ultima versione introdotta dallo scrivente.

2.2.3.1.1 Nella trascrizione diplomatica per un "allora le ho detto" con *gli* cassato sul rigo, una lezione non recuperabile cancellata sopra il rigo a sinistra (resa con `{x}`) ed un *li* cassato sopra il rigo a destra, avremo:

`allora le {gli,x,li} ho detto` [TD].

2.2.3.1.2 Nella trascrizione tokenizzata e markuppata, avremo invece:

`allora le <CORR>gli,x,li</CORR> ho detto` [TTM].

2.2.3.2 Le **inserzioni** saranno rese in modo analogo in entrambe le trascrizioni, in TD ricorrendo allo zero, ed in TTM al tag `<INS>`. Immaginiamo "allora le ho detto" e "allora a lei ho detto" con "le" ed "a lei" inseriti nell'interlinea; avremo risp.

2.2.3.2.1 `allora le {0} ho detto` [TD].

`allora a lei {00} ho detto` [TD].

2.2.3.2.2 `allora le <INS>le</INS> ho detto` [TTM].

`allora a lei <INS>a lei</INS> ho detto` [TTM].

Si badi alla diversa strategia (numero di zeri in TD) per indicare l'estensione dell'inserzione in TD e TTM.

La posizione dell'elemento inserito, sopra, sotto o a lato della riga, non è rilevante.

2.2.3.3 Si noti inoltre che i tags di markup testuale per interlinee e marginalia (§ 2.5.1) non sono invocabili per specificare la distribuzione materiale delle varianti nella pagina. Entrambe le notazioni prescindono, infatti, dalla corretta resa della specificità paleografica delle correzioni (cfr. anche quanto specificato sul trattamento delle indentature, § 2.1.2.2): esse sono intese alla creazione di un mero Learner Corpus (corpus di apprendenti), non all'allestimento di una vera edizione critica XML di un testo, in cui la accuratezza nella rappresentazione della natura anche materiale del testo manoscritto è invece fondamentale.

2.2.3.4 Per un esempio di inserzione e correzione combinate (correzione innestata in inserzione) cfr. § 2.7.2.

2.2.4 **Nota bene**: gli interventi correttivi del docente **non** devono essere considerati.

2.2.5 La **varianti**, ovvero più di una proposta per uno stesso termine (si noti che non si tratta di una correzione, né di un'inserzione, le due lezioni coesistono sullo stesso piano), in TD saranno semplicemente separate dal diacritico ¦ (ANSI 0166) e in TTM saranno racchiuse nel tag `<VAR>`. Avremo quindi:

C'era una volta un bambino di nome Gigi. Un giorno il bambino ¦ ragazzo, mentre portava a spasso il suo cane si imbattè in un orco. [TD].

C'era una volta un bambino di nome Gigi . Un giorno il <VAR>bambino ¦ ragazzo</VAR> , mentre portava a spasso il suo cane si imbattè in un orco . [TTM].

2.2.6 Le **lacune**, cioè le zone del testo non leggibili per difetto (e.g. fotocopia malfatta) o guasto meccanico della copia (e.g. bruciatura di tabacco, macchia di caffè, incrostazione di brioscina, ecc.), devono essere adeguatamente segnalate. [TD+TTM].

2.2.6.1 In TD vanno notate con le parentesi quadre al cui interno si pongono tante *x* quanti grossomodo sono i caratteri che potrebbero starvi; se si riesce a leggere o indovinare con ragionevole sicurezza qualcuno dei segni contenuti nella lacuna, queste congetture vanno ugualmente inserite nelle quadre, con o senza altre *x*. L'esempio seguente è tratto da una lista numerata di temini, di cui costituiva l'ultimo; la fotocopia risultava leggermente tagliata sul margine sinistro e fortemente annerita al fondo della pagina: [TD]

Ti ricordi di quando avevi due anni ? Scrivi della cosa che ti piaceva di più e quella che odiavi di più.

- Quando ero bambino mi piaceva mangiare i cioccolati. Tutto il tempo io mangiavo i cioccolati perché era molto buoni . Il sapore era molto buono . Potevo fare tutto per i cioccolati . Ogni giorno pensavo che abbia mangiato 4 O cinque (5) cioccolati.
Ma non mi piaceva gioc [xxxxxxxx] qualcuno perché volevo giocare solo {0} con mia mad[xxxxxxxx] mia madre . Tutto il tempo volevo [xxxtxxxxxxxxxx] madre [xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx] [xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx] [TD]

2.2.6.2 In TTM il trattamento è identico, solo che le *x* e le letture congetturali invece che essere racchiuse tra quadre sono inserite nel tag <LAC>xxx</LAC>. La riga 8 dell'esempio precedente in TTM sarebbe: [TTM]

<blank_3></blank> Ma non mi piaceva gioc <LAC>xxxxxxxx</LAC>

2.2.7 Dal guasto meccanico va in linea di principio distinta la **difficoltà paleografica**: può infatti capitare che, nonostante ogni ragionevole sforzo, non si riesca a decifrare alcune parole (o parti di parole) semplicemente perché illeggibili di per sé. Nella maggior parte dei casi ciò avviene nel corso di un processo correttorio o variantistico, ed in questi casi si è già suggerito come comportarsi: basta inserire tante *x* quante sono le (ragionevolmente supponibili) lettere illeggibili ed inserire poi queste nei tags volta per volta appropriati.

Può raramente capitare che l'impossibilità di lettura si verifichi in una zona neutra del testo: in questo caso, onde non generare "parole fittizie", le *x* vanno inserite tra doppie parentesi quadre [[xxx]] (in TD; le parentesi sono doppie per distinguere la lezione semplicemente illeggibile dalla danneggiata, notata dalle parentesi semplici, cfr. supra) o nel tag <illegg>xxx</illegg> (in TTM).

Superfluo mi sembra naturalmente avvertire che la rinuncia interpretativa vada contenuta al minimo possibile.

2.3. ***Divisione delle parole.***

Deve sempre essere ricostruibile l'originaria.

2.3.1 Nella trascrizione diplomatica iniziale la divisione delle parole originaria viene mantenuta come è, giusta o sbagliata che sia. Avremo pertanto

odeto al lamico dell'ele fante [TD].

2.3.2 Nella versione tokenizzata ogni token deve essere separato da spazio, e bisognerà pertanto ricomporre i token corretti senza perdere l'informazione sulla divisione dell'originale, grazie all'introduzione del diacritico ¬ (*logicalnot*, ANSI 0172: spazio inserito) e + (*plus,* ANSI 043: spazio eliminato). L'esempio precedente diventerà pertanto:

o¬ deto al+l¬ amico dell'¬ ele+fante [TTM].

2.3.2.1 Si badi in particolare che le parole con apostrofo nella tokenizzazione andranno sempre separate con uno spazio (più eventualmente il logicalnot) dalla parola seguente, cfr. § 2.4.5 [TTM].

2.3.2.2 Nei testi originariamente elettronici (emails ecc.) può capitare che le parole siano divise da più di uno spazio (doppi spazi, tripli, ecc.):

nella trascrizione diplomatica tale caratteristica deve essere essere conservata [TD]

nella trascrizione tokenizzata (in cui le "parole" sono ridotte a "tokens") diventa invece indispensabile eliminare l'anomalia (è sufficiente un "search" di doppio blank, di solito) in quanto ogni token deve essere preceduto/seguito da uno ed un solo spazio (o fine/inizio riga). [TTM]

2.3.3 Al momento si rinuncia ad introdurre una divisione di token per le preposizioni articolate (che richiederanno così una POS supplementare) e le catene clitiche. Il problema sarà tuttavia riesaminato durante il POS-tagging. [TD+TTM].

2.4. ***Interpuntemi, diacritici, caratteri grafici.***

Si mantiene di norma il sistema dell'originale.

2.4.1 Tutti i segni di **punteggiatura ordinaria** (punto, due punti, ecc. ecc.), come che siano posizionati nell'originale e nella trascrizione diplomatica [TD], nella tokenizzazione vanno separati da spazio [TTM]:

`virgola, punto.` [originale]

`virgola , punto .` [TTM]

2.4.1.1 Si badi, però, che le serie di interpuntemi, tipo *!!! ?!?* ... ecc., sono trattate come interpuntemi compatti e quindi non sono spaziate al loro interno. [TTM].

2.4.1.2 In TD si riproduce entro ragionevoli limiti la situazione degli originali. La formula cautelativa è dovuta al fatto che non sempre nella grafia manuale degli apprendenti è facile distinguere quando un interpuntema sia attaccato alla parola che precede o sia separato da essa con uno spazio. La raccomandazione è pertanto di trascrivere come separati da spazio in TD solo quei casi dove la spaziatura sia con buona certezza voluta dallo scrivente, in base alla evidenza paleografica (l'interpuntema è sensibilmente staccato) e/o all'uso relativamente sistematico (l'apprendente si comporta regolarmente così). Tutti gli altri casi dove non vi sia una sufficiente certezza vanno ricondotti all'uso normale (interpuntema attaccato alla parola che precede). [TD]

2.4.2 Accanto alla punteggiatura ordinaria è introdotto anche un carattere speciale usato da solo (#) od in combinazione con altri, specie il punto (`.#`) per l'**andata a capo** [TTM]. L'uso del marcatore # non deve intendersi come una semplice marca di fine riga (l'andare a capo meccanico nella trascrizione è già sufficiente allo scopo) ma come specificazione introdotta per un segno di interpunzione (solitamente punto fermo, esclamativo, interrogativo, puntini di sospensione, lineetta) usato come finale.

2.4.3 Particolare attenzione va posta al **punto**. Il punto come segno di interpunzione (sia esso a capo o di seguito) va infatti regolarmente tokenizzato in TTM (es. `punto .`), ma il punto come segno abbreviativo no (es. `i.e.`). [TTM]

Quindi avremo, ad es.

```
Sono stufo . Vado a dormire .#
Dammi gli attrezzi : martello , pinza , ecc. , e viti normali e
    parker , i.e. autofilettanti .#
```

2.4.4 Le **virgolette**, semplici o doppie, non sono di solito ulteriormente specificate nella scrittura manuale, ed anche nella videoscrittura la scelta tra le "diritte" ("_", uguali in Times ed in Courier) e le curve od "inglesi" (**"_"** in Courier e "_" in Times) è attuata automaticamente da Word e simili programmi. Le coppie curve od inglesi **'_'** (**'** ANSI 0145 e **'** ANSI 0146; in Times '_') o **"_"** (**"** ANSI 0147 e **"** ANSI 0148; in Times "_") sono usate come i rappresentanti convenzionali di queste virgolette generiche. Ogni qual volta l'autore distinguerà esplicitamente differenti tipi di virgolette, quali perlopiù i caporali (" ANSI 0171 e " ANSI 0187), queste verranno riprodotte e mantenute come tali. [TD+TTM].

2.4.5 L'**apostrofo** sarà sempre rappresentato con la forma diritta (' ANSI 039), per l'utilità di averlo rappresentato da un codice diverso da quelli della virgoletta semplice (resa con l'apice inglese, **'_'** ANSI 0145-6; in Times '_') già fin dalla trascrizione raw. [TD+TTM].

Si badi che in TTM i gruppi con apostrofo vanno sempre tokenizzati introducendo uno spazio di norma a destra (e non a sinistra), quindi si avrà "`l' amaca`", "`un po' di birra`", "`ciao , 'notte !`".

2.4.5.1 Per quanto riguarda la separazione o meno dell'apostrofo, in TD si riproduce solo entro ragionevoli limiti la situazione degli originali: ogni caso dubbio va infatti ricondotto automaticamente all'uso standard secondo quanto detto per le interpunzioni al § 2.4.2 [TD]

2.4.6 **Simboli** quali asterischi, trattini e freccette vengono trattati come normali caratteri del testo, nella fattispecie per gli asterischi (e segni di richiamo a stella in genere) si userà il carattere "`*`" (ANSI alt-042), per lineette e trattini in genere si userà il carattere "`-`" (ANSI alt-045), per le freccette si useranno le combinazioni, rispettivamente con verso a destra ed a sinistra, risp. "`-›`" (ANSI alt-045 + alt-0155) e "`‹-`" (ANSI alt-0139 + alt-045), e per i segni ondulati di 'circa' 'alternanza' e simili si userà la semplice tilde "~" (ANSI alt-0126). [TD+TTM]

`mia madre è casa+linga * .` [TTM]

`- In India l'istituzione del matrimonio è molto forte,` [TD]

`-› Bologna è una città antica.` [TD]

2.4.6.1 Un particolare insieme di simboli è quello degli **emoticons**. Questi vanno tokenizzati e riprodotti come sono, quindi con spazi all'esterno ma non all'interno; ad esempio:

`:-(  ^__^  ;-)` [TD+TTM]

2.4.7 Per le **marche di evidenziazione** o enfasi, quali le sottolineature (più frequenti nella scrittura manuale), i corsivi, grassetti maiuscoletti ed espansi (più frequenti nella videoscrittura) si ricorre all'attributo `<emph_`*`valore`*`>__</emph>`, nella maniera seguente (la notazione – "`label`" – dei valori dei tag assume qui la base inglese – underlined, dotted, bold, italics, ecc. – per via della pressoché universale conoscenza e diffusione di tale terminologia grazie ai software di videoscrittura):

2.4.7.1 Il **sottolineato** è rappresentato con `<emph_u1;u2;u3>__</emph>`. I valori previsti sono singolo "`u1`", doppio "`u2`" e triplo "`u3`". [TD+TTM].

Il **tratteggiato** è rappresentato con `<emph_h1;h2;h3>__</emph>`. I valori previsti sono singolo "`h1`", doppio "`h2`" e triplo "`h3`". [TD+TTM].

Il **puntinato** è rappresentato con `<emph_d1;d2;d3>__</emph>`. I valori previsti sono singolo "`d1`", doppio "`d2`" e triplo "`d3`". [TD+TTM].

Il **corsivo** è rappresentato con `<emph_i;bi>__</emph>`. I valori previsti sono corsivo normale "`i`" e grassetto corsivo "`bi`". [TD+TTM].

Il **grassetto** è rappresentato con `<emph_b;bb>__</emph>`. I valori previsti sono corsivo normale "`b`" ed extra-bold "`bb`". [TD+TTM].

Il **maiuscoletto** è rappresentato con `<emph_sc>__</emph>`. Il valore previsto è solo "`sc`" (small capitals). [TD+TTM].

L' **espanso** è rappresentato con `<emph_xp>__</emph>`. Il valore previsto è solo "`xp`" (expanded). [TD+TTM].

Il **cerchiato** è rappresentato con `<cerc>__</cerc>`. [TD+TTM].

2.4.7.2 Per **evidenziazioni complesse** si possono liberamente combinare i valori semplici, così ad esempio un maiuscoletto grassetto con doppia sottolineatura sarà marcato: `<emph_sc,b,u2>__</emph>.` [TD+TTM].

2.4.8 L'uso intenzionale di **colori** diversi nel testo può essere rappresentato con il tag `<col_red;green,...>__</col>`. [TD+TTM].

2.4.9 La presenza di **disegni** può essere resa da un set limitato di sigle convenzionali; quelle per ora proposte (ma altre potrebbero venire aggiunte in base ad esigenze specifiche) sono: `SG` "segno grafico" generico (per ogni altro disegno-carattere, tipo faccine, fulmini, ecc.), `DN` (per disegni naturalistici estesi anche su più righe), `DT` (per disegni tecnici, come un pezzo di circuito elettrico), `DS` (per diagrammi schematici, tipo schema a blocchi, ecc.). Se i disegni non sono rilevanti per la comprensione del testo è sufficiente sostituirli con le rispettive sigle, senza acquisirli otticamente (cfr. § 2.4.9.3).

2.4.9.1 In TD le sigle di disegni vanno poste in parentesi uncinate, per cui ad esempio avremo:

`Oggi <SG> è una bella giornata` [TD].

in cui `<SG>` sta, ad esempio, per un sole che ride (non è necessario esplicitare ulteriormente la natura del disegno perché, come detto sopra, le sigle sono usate appunto solo per disegni non rilevanti per la comprensione del testo).

2.4.9.2 In TTM le sigle vanno invece chiuse nel tag `<imgint>`, per cui lo stesso esempio precedente sarà reso al modo seguente:

`Oggi <imgint>SG</imgint> è una bella giornata` [TTM].

2.4.9.3 Se il disegno è rilevante per la comprensione del testo o per il profilo psicologico dell'autore (cosa frequente soprattutto negli elaborati infantili), questo può essere acquisito otticamente (scannato e digitalizzato) (cfr. § 1.2.6 dove si trova l'elenco di tutti i riferimenti ipertestua-

li istituiti dal documento e § 1.2.6.4, che si riferisce specificamente al trattamento delle immagini). In questo caso nella trascrizione TTM si aggiungerà un link HTML-like al file esterno, `<imgint src="nomefile.jpg">`, che precisa dove l'immagine vada effettivamente inserita; nel caso dell'es. precedente avremo:

`Oggi <imgint src="solecheride.jpg">SG</imgint> è una bella giornata` [TTM].

2.4.10 La presenza di **allegati di natura testuale**, quali ritagli di giornale, ecc., sarà invece rappresentata con la sigla `TX` che analogamente agli elementi grafici saranno poste in parentesi uncinate in TD (cfr. § 2.4.9.1) ed invece chiuse nel tag `<txtint>`nella trascrizione TTM (cfr. § 2.4.9.2). Nel caso in cui il testo in questione venga trascritto in file a parte, secondo le modalità specificate nel § 1.2.6.5 (cui cfr.), nella trascrizione TTM si aggiungerà un link HTML-like al file esterno, `<txtint src="nomefile.jpg">`, che precisa dove il testo vada effettivamente inserito; ad es. per il caso riferito nel § 1.2.6.5 avremo:

`Ho visto l' anuncio <txtint src="annuncioVecchia.txt">TX</txtint> sulla Notte` [TTM].

2.5. ***Markup testuale.***

Contrassegna nella TTM le strutture principali del testo.

2.5.1 Si marcano, in primo luogo, **zone speciali** del testo, come

2.5.1.1 <**titolo**>il titolo del brano, del paragrafo o del capitolo</**titolo**> [TTM]

2.5.1.2 <**pcoll**>le formule iniziali (protocollo) nelle lettere, es. *Dear, Ciao!* </**pcoll**> [TTM]

2.5.1.3 <**ecoll**>le formule di congedo (escatocollo) nelle lettere, es. *Bye, Ugo* </**ecoll**> [TTM]

2.5.1.4 <**versi**>eventuali parti versificate</**versi**> [TTM]

2.5.1.5 <**nota**>testo della nota</**nota**> [TTM]

2.5.1.6 <**marginale**>interpolazione nel margine</**marginale**> [TTM]

2.5.1.7 <**interlinea**>interpolazione nell'interlinea</**interlinea**> [TTM]

2.5.1.8 <**calce**>interpolazione nell'interlinea</**calce**> [TTM]

2.5.1.9 Avvertenze per l'uso dei tag <**marginale**>**,** <**interlinea**> e <**calce**> .

2.5.1.9.1 Questi tag sono di natura più testuale che paleografica e devono essere usati per indicare porzioni di testo di relativa autonomia ed estensione, non semplici parole singole portate fuori dal rigo in un processo correttorio, come specificato anche in 2.2.3.3. [TTM]

2.5.1.9.2 Marginalia ed interlinea così definiti sono markuppati propriamente solo in TTM, in TD vengono trascritti tra quadre, nella riga dopo il punto in cui l'inserzione è stata esplicitamente richiesta o comunque chiaramente intesa dall'autore, od altrimenti in calce al testo.

Immaginiamo, ad esempio, che nel seguente testo, completo e di due sole righe, vi siano due marginalia, di cui il marginale 1 sia introdotto con un asterisco nel margine superiore ed il marginale 2 sia posto nel margine sinistro senza riferimenti nel testo. In TD avremmo semplicemente:

```
Il tram è uscito da rotaie e caduto sul fianco . Nessuno morto .
[Sopra una macchina di polizia e tre carretta di angurie .]
Io ho divertito molto .
[Nonna non daccordo, ma me troppo ridere .] [TD].
```

2.5.1.9.3 In TTM il sistema è il medesimo, salvo che si ricorre al tag `<marginale>` al posto delle quadre, e che in presenza di un segno di richiamo esplicito nel testo sarà possibile usare anche una notazione HTML-like di "anchor" `<A>` e "name" (`name="___"`) per meglio precisare l'inserzione.

Lo stesso esempio di prima sarà pertanto rappresentato così (il `name` dell'ancora `A` può naturalmente essere scelto liberamente caso per caso):

```
Il tram è uscito da rotaie e caduto sul fianco
    <href="#star_1">*</href> . Nessuno morto .
<marginale> <A name=″star_1″>*</A> sopra una macchina di polizia
    e tre carretta di angurie . </marginale>
Io ho divertito molto .
<marginale> Nonna non d¬ accordo, ma me troppo ridere .
    </marginale> [TTM].
```

2.5.1.9.4 Ad analogo trattamento saranno sottoposte anche le eventuali note (a fondo pagina od a fondo testo). In TD saranno semplicemente riprodotte a fondo testo, in TTM saranno ancora riprodotte a fondo testo, ma riceveranno anche il trattamento html-like illustrato in 2.5.1.9.2-3.

2.5.2 Vanno poi marcate zone del **testo del docente**, in quanto di diverso autore dallo scrivente - apprendente, come tipicamente le domande poste dall'insegnate in questionari od esercizi di *comprehension*:

`<docente>__</docente>` [TTM]

Si noti che il tag `<docente>` può essere embricato nel tag `<turno>` (cfr. § 2.5.5), per testi dialogici docente - allievo.

2.5.3 La **citazione**, anch'esso testo propriamente di diverso autore dall'apprendente, viene contrassegnata con un tag apposito:

`<citaz>__</citaz>` [TTM]

2.5.4 Il **discorso diretto** viene contrassegnato con il tag

`<ddir>__</ddir>` [TTM]

Ad esempio avremo:

```
Mentre Egidio stava finalmente tornando a casa dopo una faticosa giornata passata a mettere a posto la cantina della suocera, incontrò la prozia Amalasunta : <ddir>« Caro Egidio ! Stavo proprio per venire da te ! Non è che mi daresti una mano a mettere a posto la soffitta ? »</ddir> , gli disse . [TTM]
```

2.5.5 Si marcano i **turni** del dialogo, con indicazione convenzionale o con il nome del dialogante assegnato nel testo, come nell'esempio seguente (con markup incompleto, ridotto a quanto qui pertinente):

```
<turno_Archimede>Oggi ho visto Topolino e Basettoni al Carrefour delle Gru</turno>
<turno_Pippo>Già , dovevano comprare il regalo per il compleanno di Minni</turno> [TTM]
```

2.5.5.1 Come ulteriore esempio contenente sia i turni del dialogo che il discorso diretto, riportiamo il medesimo dialogo che avevamo visto, con markup incompleto ed in versione ridotta, in 2.1.2.1:

```
<turno_A>Io : <ddir>Buongiorno , potrebbe
aiutarmi ? </ddir></turno>
<turno_B>Commesso : <ddir>Buongiorno Signor , cosa potrei fare
<blank_2>per lei ? </blank></ddir></turno>
<turno_A>Io : <ddir>Oggi è il compleanno di mia amica .
<blank_1>Ho preparato una torta buona ma il
mio cane la ha mangiata e devo
procurarla da qualche mezzi . </blank></ddir></turno> [TTM]
```

Nel caso di e-mail con quoting, i quoting saranno chiamati `turno_quote`, eventualmente numerati nel caso di molteplicità di fonti `turno_quote1-n` :

2.6 ***Markup di pre-tagging***.

Sono alcune categorie introdotte nella TTM che propriamente apparterrebbero piuttosto al POS-tagging, ma che, praticamente, risulta utile introdurre prima. [TTM].

2.6.1 Sono in primo luogo i **nomi propri** che dovrebbero essere marcati come tali anche prima del POS-tagging; in particolare 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>
```

2.6.2 Gli eventuali **indirizzi web** presenti nel testo saranno marcati con il tag `<url>`. Ad es.:

`andate su <url>www.pippo.it</url> e guardate che bbbello`

2.6.3 Anche le **espressioni numerico-matematiche** o comunque in cifre, ad esclusione dei semplici numerali "linguistici" espressi in cifre anziché in lettere e dei punti-elenco, saranno adeguatamente contrassegnate col tag `<mat>`. [TTM]

Avremo quindi marcati con `<mat>` esempi come

`<mat>15 + 3 / 2 = 9</mat>`

ma non marcati esempi come i seguenti:

`voglio 15 giorni di vacanza`

2.6.4 Un apposito tag è invece previsto per le **espressioni di datazione**, siano esse numeriche o frasali; quando la data è sufficientemente determinabile il tag può essere fornito di un attributo per specificarla in formato standard:

```
<date>__</date>
<date_yyyy-mm-dd>__</date>
```

Ad es. (tanto il primo esempio quanto il formalismo sono adattati dalla TEI):

```
Given on the <date_1977-06-12">Twelfth Day of June in the Year of Our Lord One Thousand Nine Hundred and Seventy-seven</date>
Il giorno di <date_0000-12-25>Natale</date>
```

2.6.4.1 Si noti che le espressioni tempo, generiche (*oggi*) o puntuali (*16:47*), non vanno marcate con il tag "`date`".

2.6.5 Le zone del testo in **lingue diverse** dall' italiano vanno contrassegnate al modo seguente:

```
<lng_nomelingua>__</lng>
```

Se non si sa quale altra lingua sia, mettere *altralingua* come nome lingua. [TTM]

2.6.5.1 Saranno markuppati con `<lng>` i sintagmi, le frasi od i paragrafi effettivamente non in italiano e le parole straniere che siano chiaramente distinguibili come prestiti non adattati. Si noti che molte espressioni straniere sono ormai state lessicalizzate anche in italiano, tali espressioni non verranno dunque marcate con il tag `<lng>`, perchè ormai entrate nell'uso comune e presenti nel dizionario italiano. Avremo quindi

```
Che fare , mi chiedeva . <lng_inglese>It 's up to you</lng> , le dico io . [TTM]
ho appoggiato la testa sul <lng_inglese>pillow</lng> e mi sono addormentato [TTM]
```

ma

```
vai nella subdirectory e apri il file " Valico " [TTM]
```

senza marca

2.6.5.2 Nel caso in cui il testo in lingua "altra" sia in una lingua scarsamente conosciuta in Europa (quindi non inglese, ma ad esempio lusaziano inferiore), si creerà un file ad esso collegato con la consueta sintassi HTML di `href` contenente la traduzione. Ad es:

```
<lng_finnico><A href="nomefile.txt">Vatanen nousi Heinolan linja-autoon, silla mukavassakaan kylässä ei pidä iättömiin joutilaana asua.</A></lng_finnico>
```

Il cui file-traduzione (che nella fattispecie potrebbe essere `paasilinna-01.txt`) conterrà:

```
Vatanen salì sul pulmann per Heinola: non poteva certo fare l'eterno sfaccendato, sia pure in un villaggio ospitale.
```

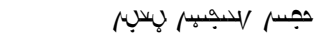
2.6.5.2.1 Nel caso in cui la lingua "altra" sia in caratteri non latini (per es. la hindi) i file di base del corpus (TD e TTM) conterranno il testo straniero in traslitterazione scientifica. Analogamente a quanto sopra, però, la TTM presenterà un collegamento HTML-like ad un file esterno contenente la traduzione, che in questo caso conterrà a sua volta un altro link html ad un file unicode (od alla peggio PDF) con il testo in caratteri originali. Ad esempio il testo (TTM)

```
la scrittura ufficiale mancese, la <lng_manju> <A href="tongki_trad.txt">tongki fuka sindaha hergen</A> </lng_manju>, fu introdotta nel <date_1632-00-00>1632</date>. [TTM]
```

punterà al file `tongki_trad.txt` con la traduzione:

```
scrittura con punti e cerchi.
[<A href="tongki_char.txt">originale</A>]
```

che conterrà a sua volta un puntatore ad un file con il testo in caratteri, ossia nel nostro es. al file `tongki_char.txt`, contenente finalmente:



2.7 ***Etichette embricate.***

Nel caso in cui si richieda più di un tag [TTM] per descrivere un particolare stato del testo, è possibile **includere un'etichetta dentro l'altra**, sempre seguendo la gerarchia indicata nell'originale, come si sarà già d'altra parte intuito da parecchi casi precedenti.

2.7.1 Un primo esempio, molto semplice, potrebbe essere il seguente:

```
<ent>Parco di<topn>Abruzzo</topn></ent> [TTM]
```

dove *Parco di Abruzzo* è il nome del parco , ma *Abruzzo* è un toponimo (cfr. § 2.6.1).

2.7.2 Un esempio più complesso, che coinvolge inserzione (cfr. § 2.2.3.2) e correzione (cfr. § 2.2.3.1): in questo caso l'apprendente ha dapprima scritto *ieri ho avuto freddo*, poi ha inserito sul rigo tra *avuto* e *freddo* la lezione *davvero grande*, quindi ha corretto *grande* in *tanto*; l'iter è rappresentabile al modo seguente (dopo l'embricatura TTM è mostrata contrastivamente anche la soluzione in TD):

```
Ieri ho avuto davvero tanto <INS>davvero tanto
    <CORR>grande</CORR></INS> freddo   [TTM]
Ieri ho avuto davvero tanto {grande} {00} freddo   [TD]
```

3. Il dopo.

3.0 Terminata la fase di immisione manuale dei dati secondo i criteri esposti nel capitolo precedente, si rendono ancora necessarie alcune operazioni sui files TTM prima di raggiungere l'assetto definitivo (i files TD resteranno invece come sono, per documentazione filologica). Solo alcune di queste fasi, tuttavia, saranno sommariamente descritte in queste Guidelines.

3.1 La prima, e più semplice, di queste operazioni consiste in un ulteriore perfezionamento del markup fino a raggiungere quello che chiamiamo **formato di scambio**, attuata questa volta automaticamente con uno script in Perl.

3.1.1 La principale modifica riguarda l'assetto delle righe, prima con l'introduzione dei tag `<tLn>` "text line" ed `<eLn>` "empty line", attuata dallo script `linee.pl` (preparato da Simona Colombo):

```
#!/usr/local/bin/perl -w
while(<>){
if (/^</)
{
    print;
}
elsif (/^\s+</)
{
    print;
}
elsif ( /^\s*$/ )
    {  ## empty line
    print "\n<eLn/>\n";
    }
else
    {
    print "\n<tLn>\n$_</tLn>\n";
    }
}
```

e poi con la loro numerazione progressiva interna ad ogni testo, attuata dallo script `8_contarighe.pl`, originariamente sviluppato da Simona Colombo per i corpora di newsgroups in allestimento:

```
#!/usr/bin/perl
# conta le tline qline e sostituisce eline con quante ne ha so-
stituite
$numrighe=0;
$numeline=0;
$eline=0;
while (<>){
        if (/^<head>/)
        {
            $numrighe=0;
            $numeline=0;
            print;
        }
        elsif (/^(\<tLn|\<qLn|\<pl)/)
        {
```

```
                    if ($eline==1)
                    {
                        $eline=0;
                        print "<eLn>$numeline<\/eLn>\n";
                        $numeline=0;
                    }
                    $numrighe++;
                    s/(<qLn)(.+)(>)/$1$2 nr\=$numrighe $3/;
                    s/(<tLn)(>)/$1 nr\=$numrighe $2/;
                    s/(<pl)(>)/$1 nr\=$numrighe $2/;
                    print;
                }
                elsif (/^<eLn\/>/)
                {
                    $eline=1;
                    $numeline++;
                }
                elsif (/^<\/body>/)
                {
                    if ($eline==1)
                    {
                        $eline=0;
                        print "<eLn>$numeline<\/eLn>\n";
                        $numeline=0;
                    }
                    print;
                }
                elsif (/^(news\:)(.+\w+)(\.\.\.|$)/)
                {
                    #s/\n//;
                    #print $1;
                    print "\n<news>$1$2<\/news>$3\n";
                }
                elsif (/(.+)(news\:)(.+\w+|.+\>)(\.\.\.|$)/)
                {
                    print $1;
                    print "\n<news>$2$3<\/news>$4\n";
                }
                else
                {
                    print;
                }
        }
```

3.1.2 Le ragioni di questo riassetto sono duplici: da un lato una più facile trasformabilità in standard XML e CQP Format, dall'altro la uniformazione agli altri corpora italiani (e non) in fase di creazione da parte del nostro gruppo (tanto in sede di bmanuel.org come di corpora.unito.it). La piena compatibilità di VALICO con altri corpora e la possibilità di poter attuare ricerche incrociate su più corpora usando la medesima sintassi di ricerca rappresenta infatti uno dei punti di forza di questo progetto.

3.1.3 Alcuni esempi di documenti in formato di scambio si trovano nell'Appendice 3.

3.2 A partire dal formato di scambio, con procedure ancora una volta completamente automatizzate, previo eventualmente un perfezionamento dell'assetto XML, si perverrà ad una versione in formato CQP già complertamente gestibile ed interrogabile come corpus, di cui è prevista la messa online come beta sul sito di corpora.unito.it.

3.3 A partire da questa prima versione semplice (solo markuppata e tokenizzata), si approntera una versione annotata; sono previsti in particolare un **POS-tagging** (basato sul tagset EAGLES e sull'esperienza del CT), da attuare usando il Tree Tagger dell'IMS Stuttgart, ed un **error-tagging**, secondo strategie ancora in parte da definire.

4. Appendice 1: questionari.

4.0 In questa appendice riproduciamo i cinque questionari che (cfr. § 0.2.2) i fornitori di testi dovranno produrre accanto alle copie meccaniche degli originali (quando non agli originali stessi) in quella che definivamo come "ipotesi minima".

Tali questionari saranno distribuiti ai "fornitori" anche in file separati od in formato cartaceo.

4.1 Il primo questionario è il ***questionario-autore***, che può essere somministrato in alternativa alla compilazione delle Headers (cfr. § 1.2 e sgg.).

QUESTIONARIO AUTORE

Caro studente d'italiano,

per favore compila la prima parte del questionario.

Nella versione pubblica del corpus in cui finiranno i tuoi scritti saranno introdotte misure per la tutela della riservatezza dei dati che ci fornisci e gli scritti saranno anonimi.

Ti chiediamo pertanto di firmare e di darci così l'autorizzazione

(a) ad usare i testi da te prodotti col fine di svolgere su tali opere attività di:

- analisi computazionale del testo (tokenizzazione, mark-up, tagging)
- inserimento del testo in uno o più corpora linguistici;
- pubblicazione anche in rete del testo nel contesto dei suddetti corpora;
- estrazione del testo o di parti di esso dai suddetti corpora;

(in particolare cedi, ai fini di cui sopra, ogni diritto sul tuo testo, ivi compresi i diritti di riproduzione diretta ed indiretta, di diffusione e comunicazione al pubblico, in qualsiasi modo forma e modo, di distribuzione, noleggio o prestito; il diritto di ripubblicare i contenuti del tuo testo nel suo formato originario rimane comunque tuo).

(b) al trattamento dei tuoi dati personali conformemente al DL 196/03 sulla privacy.

(firma dell'autore del testo) _____

QUESTIONARIO

A cura dell'autore del testo

Nome:

Cognome:

Sesso: m ☐ f ☐ gruppo ☐

Età:	1-7	8-13	14-18	19-25	26-30	30-40	40-50	oltre
	☐	☐	☐	☐	☐	☐	☐	☐

Status sociale (in base al reddito: modesto (1), medio (2), alto (3)): 1 ☐ 2 ☐ 3 ☐

Da quanti anni studi italiano: 1 ☐ 2 ☐ 3 ☐ 4 ☐ + ☐

Lingua madre:

Lingua 1 di comunicazione, se diversa dalla lingua madre:

Altre lingue conosciute (metterle in ordine, partendo dalla più conosciuta):

Grado di scuola frequentato in lingua madre: nessuno ☐ elementare ☐ medio ☐ superiore ☐ universitario ☐

Permanenza in Italia: dove _____ _____ _____
quanto _____ _____ _____

Dove e quando ascolti/leggi e parli/scrivi italiano: a scuola ☐ con amici ☐ in famiglia ☐ radio TV internet ☐

A cura del docente o di chi fornisce il testo scritto

Nome (del docente):

Cognome (del docente):

Data di produzione:

Luogo:

Istituzione:

Tipo di testo: composizione: libera ☐ tesina ☐ testo dialogico ☐
questionario/comprensione ☐
traduzione ☐ dettato ☐ riassunto ☐ e.mail ☐ lettera ☐

Prova data come: prova di fine anno ☐ prova in itinere ☐ senza valutazione ☐

4.1.1 La prima parte va compilata dall'autore del testo, la seconda invece dal docente o fornitore (nel caso in cui l'autore coincida con il fornitore, dovrà compilare anche i campi Data, Luogo e Istituzione.).

4.1.2 Ribadiamo il fatto che nella versione pubblica del corpus saranno introdotte misure per la tutela della privacy e chiediamo esplicitamente l'autorizzazione al trattamento dei dati conformamente alla legge 675/96.

4.2 Il secondo questionario è il **questionario-docente**, che contiene le generalità e il curriculum del docente (ossia del somministratore degli esercizi e pertanto di solito fornitore dei testi).

QUESTIONARIO DOCENTE

Nome: _____ Cognome: _____
Indirizzo: _____
Tel.: _____
e-mail: _____
Istituzione di appartenenza:
Università _____
Scuola superiore _____
Istituto di cultura italiana _____
Altro _____
Ruolo: _____

Curriculum di studi:
(es :dal Sett 1990 al Giugno 1995 Laurea in Lettere moderne- Università degli studi di Torino)
dal ..... al ..... _____
dal ..... al ..... _____
dal ..... al ..... _____
dal ..... al ..... _____

Curriculum lavorativo:
(es :dal Ott 2000 al Luglio 2003 Docente di Italiano L2- Università del Piemonte Orientale)
dal ..... al ..... _____
dal ..... al ..... _____
dal ..... al ..... _____
dal ..... al ..... _____

Altre attività:
(es.: collaborazione con vari organismi in attività di educazione interculturale)
_____
_____
_____
_____

Interessi di ricerca:
(es.Linguistica acquisizionale (in particolare la sintassi) – glottodidattica (rapporto apprendente – testo)

Pubblicazioni principali:
_____

_____
_____
_____

4.3 Il terzo questionario è il **questionario-esercizio**, che contiene le caratteristiche dell'esercizio: consegna, scopo, informazioni sullo svolgimento dell'esercizio.

QUESTIONARIO ESERCIZIO

Consegna*:

(1) fornire una fotocopia della consegna stessa (preferibile), oppure
(2) trascrivere

_____
_____
_____

(es: Leggete l'incipit del racconto e scrivete una breve storia (200 parole max.) Marta non era una ragazza qualunque...)

Scopo dell'esercizio†:

(es: Verificare l'acquisizione della struttura logica del testo narrativo e la capacità di usare correttamente i tempi verbali studiati)

_____
_____

Contesto:

esercizio in classe □
esercizio a casa □

Tipo di testo:

descrittivo □ narrativo □ argomentativo □ regolativo □ articolo □
tesina □ dialogo □ questionario □ traduzione □ dettato □
riassunto □ lettera □ altro_____

Extra info‡:

interventi del docente:

_____

lessico suggerito:

_____

strutture morfosintattache suggerite:

_____

altro:
(es:l'esercizio è stato svolto a gruppi di due senza l'ausilio del dizionario)
_____

NOTE
* Per *consegna* si intende il compito assegnato allo studente
† Per *scopo dell'esercizio* intendiamo le abilità testate dall'esercizio
‡ Per *extra info* intendiamo qualsiasi tipo di informazione che non sia contenuta nelle voci precedenti. Es. modalità di svolgimento dell'esercizio, materiali a disposizione...

4.4 Il quarto questionario è il **questionario-test**, che contiene le caratteristiche della prova sottoposta agli studenti. È necessario compilare tale questionario solo nel caso in cui i testi forniti siano delle prove di verifica, siano esse in itinere o di fine anno.

QUESTIONARIO TEST

Tipo di prova:
esame di fine corso □
verifica in classe □

Tempo a disposizione: _____

Dizionari lasciati a disposizione:
monolingue □ _____
bilingue □ _____
nessuno □

Testi di riferimento*:
_____
_____
_____

NOTA
* Per *testi di riferimento* si intendono i testi usati dai ragazzi per la preparazione del test

4.5 Il **questionario-scuola**, infine, contiene le generalità dell'istituzione presso la quale sono stati prodotti gli elaborati e alcune informazioni sul corso preso in considerazione. Nel caso in cui i gradi di istruzione italiani non corrispondano con quelli stranieri è necessario fornire il nome dell'eventuale grado estero corrispondente.

QUESTIONARIO SCUOLA

Nome della Scuola: _____
Indirizzo: _____
Grado:
elementare □
(_____)*
media inferiore □
(_____)*

media superiore □
(_____)*
università □ Corso di laurea_____
(_____)*
altro: _____

Informazioni sul corso preso in considerazione:
corso per principianti □
corso intermedio □
corso avanzato □
principali argomenti trattati:
_____
_____

NOTA
* Indicare l'eventuale nome del grado estero corrispondente.

5. Appendice 2: stelloncini.

5.0 Forniamo qui i modelli, accompagnati da brevi descrizioni di riferimento, degli stelloncini che il trascrittore (od il fornitore nell'ipotesi "massima") dovrà allestire.

5.1 **Stelloncino-fornitore.** Nome della persona che ha materialmente raccolto il testo; ogni fornitore di testi dovrà fornire le proprie generalità (rispondendo al "questionario docente" o direttamente compilando questo stelloncino), anche istituzionali e scientifiche (cfr. § 1.2.1.4). Il "nome" di tale stelloncino sarà dato dal "nomecognome" del fornitore accompagnato dalla sigla `F`, ed il suo formato sarà lo stesso `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per Tanya Roy avremo

`tanyaroy_F.txt`

5.1.1 Il template (fornito anche in file a parte per il riempimento) è pertanto il seguente:

```
Indirizzo:
Tel:
e-mail:
Istituzione di appartenenza e ruolo:
corso di studi:
              percorso lavorativo;
              attività;
              interessi di ricerca;
              pubblicazioni.
```

5.2 **Stelloncino-trascrittore**. Nome della persona che ha materialmente trascritto il testo, nel caso che questa sia distinta da chi lo ha raccolto; anche in questo campo si dovrà compilare uno stelloncino con le proprie generalità con i criteri di cui sopra (cfr. anche § 1.2.1.5). Analogamente, il "nome" dello stelloncino sarà dato dal "nomecognome" del trascrittore accompagnato dalla sigla `T`, ed il suo formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per Francesca Minozzi avremo

`francescaminozzi_T.txt`

5.2.1 Nel caso fornitore e trascrittore coincidano, l'indicazione sarà ripetuta più volte, e la sigla nel nome del file sarà `FT`, ad es.

`silviacamarca_FT.txt`

5.2.2 Il template (fornito anche in file a parte per il riempimento) è pertanto il seguente:

```
Indirizzo:
Tel:
e-mail:
Istituzione di appartenenza e ruolo:
corso di studi:
              percorso lavorativo;
              attività;
```

```
interessi di ricerca;
pubblicazioni.
```

5.3 **Stelloncino-istituzione**. Generalità e caratteristiche dell'istituzione in cui sono stati prodotti e raccolti i testi (cfr. § 1.2.1.8.1). Lo stelloncino dovrà avere una lunghezza di max 720 battute (cioè c. 8 righe di 90 battute). Ogni stelloncino dovrà essere posto in un file separato, avente per nome una forma sintetica del nome dell'istituzione medesima accompagnato dalla sigla `I`, ed il suo formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per i documenti raccolti presso l'Università di New Delhi avremo

`Delhi-BA_I.txt`

5.3.1 Il template (fornito anche in file a parte per il riempimento) è pertanto il seguente:

```
Nome ufficiale:
Indirizzo:
Dipartimento / Corso laurea ecc.
Grado
Extra info (es. corso preso in considerazione)
```

5.4 **Stelloncino-gruppo**. Nel caso di gruppi di esercizi (cfr. §§ 1.2.2.3-4 e soprattutto 1.2.2.4.1), bisognerà indicare su uno stelloncino in max 900 battute (cioè c.10 righe di 90 battute) le caratteristiche dell'esercizio (scopo, abilità coinvolta, contesto, consegna…). Ogni stelloncino dovrà essere posto in un file separato, avente per nome lo stesso nome assegnato al gruppo accompagnato dalla sigla `G`, ed il suo formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio per il gruppo `rane` si avrà un file dal nome

`rane_G.txt`

5.4.1 Il template (fornito anche in file a parte per il riempimento) è pertanto il seguente:

```
Consegna (descrizione sintetica):
Scopo dell'esercizio:
Abilità coinvolta:
Tipo di testo (narrativo, descrittivo, argomentativo,
regolativo, articolo...)
Contesto:
Extra info:
```

5.5 **Stelloncino-prova**. Lo stelloncino illustrerà le caratteristiche della prova (cfr. § 1.2.5.4), chiarendo anche le sue condizioni di svolgimento (tempo dato, possibilità di consultare dizionari monolingui o bilingui di italiano o altri testi di riferimento). Il "nome" dello stelloncino sarà dato da una forma convenzionalmente abbreviata del nome della prova accompagnato dalla sigla `P`, ed il suo formato sarà il solito `.txt` degli altri files del corpus (cfr. §§ 2.0.1 e 2.0.3). Ad esempio avremo

`3dDegree_P.txt.`

5.5.1 Il template (fornito anche in file a parte per il riempimento) è pertanto il seguente

```
Tipo di prova
Tempo a disposizione
Dizionari a disposizione
Testi di riferimento
Extra info
```

5.6 **Stelloncino-consegna**. Riproduce la consegna materialmente assegnata dal docente. Questa va trascritta integralmente (secondo i criteri della TD) su file separato, il cui "nome" sarà dato da una forma convenzionale abbreviata del titolo della consegna (perlopiù la stessa del nome del gruppo, quando presente: cfr. § 1.2.2.3) accompagnato dalla sigla `C`, ed il suo formato sarà il solito `.txt` (cfr. §§ 2.0.1 e 2.0.3).

5.6.1 Trattandosi di una riproduzione fedele dei materiali approntati dal docente, la struttura del suo contenuto sarà prevedibilmente assai varia, e non è possibile fornirne un template.

6 Appendice 3: esempi di trascrizioni.

Forniamo qui di seguito alcuni esempi tra le prime trascrizioni effettuate, scelte per essere rappresentative dei principali problemi di trascrizione illustrati in queste Guidelines.

6.1 Un primo esempio, abbastanza semplice, è il seguente, fornito da Stefania Ferraris.

6.1.1 versione TD, `stefania001_TD.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Alessandra,Vogels</aut_NC>
       <fornitore>Stefania,Ferraris</fornitore>
       <trascr>Stefania,Ferraris</trascr>
       <data>?,12,02</data>
       <luogo>Vercelli,IT</luogo>
       <ist>scuola</ist>
       <ist_nome>Università del Piemonte Orientale</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>2</status>
       <annualita>1</annualita>
       <lingua1>tedesco,0</lingua1>
       <lingue>inglese,francese</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>6,Vercelli</permanenza>
       <esposizione>sc,am,fam,med</esposizione>
   </autore>
   <testo>
       <tipo_forma>c-lib_descr</tipo_forma>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>orig</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>stefaniaferraris_FT.txt,0,0</stel>
       <cons>Italia_C.txt</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ Quello che mi è piaciuto e quello che non mi è piaciuto in Italia

Sono tante le cose che mi hanno piaciuto e che mi
piaciono anche adesso dell'Italia ma altretante
quelle che non mi piaciono.
Inanzitutto mi piace la lingua italiana, cosa che
mi ha portato da venir qui - penso che sia
molto armonica e più musicale
in contrasto con la lingua tedesca.
Poi adoro la cucina sopratutta la facilità
con chi si può preparare una cosa
così gustosa.
Naturalmente con dei cibi così buoni non
può mancare il vino che, secondo me, è
eccezionale.
Mi piace tanto anche la regione in cui vivo,
Piemonte, perché in poco tempo sono al
mare e se mi annoio posso cambiare e
sono velocissimo nella montagna dove posso
$002$ sciare.
Inoltre mi piaciono tantissimo i miei amici,
le persone con cui vivo insieme qui
a Vercelli e i ragazzi spagnoli chi ho incontrato
al corso di italiano.
Essi sono sicuramente le cose migliore
che sono successe a me in Italia fino adesso!
```

```
La prima cosa che ho notato venendo in Italia
e non mi piace è il fatto di dover pagare
per le autostrade italiane.
Penso anche che il caos dell'Università
qui a Vercelli insieme alla burocrazia italiana
sono veramente faticosi.
La cosa che mi manca di più della Germania
è il nostro Natale!
È tutto diverso con le abitudine, i colori
che adesso non sto sentendo veramente
come i giorni di Natale sono vicini.

Ma sono sicurissima che prenderò
tanti bellissimi pensieri a tutto che ho
visto, ho fatto e chi ho incontrato
quando tornerò in Germania!
</BODY>
```

6.1.2 versione TTM, `stefania001_TTM.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Alessandra,Vogels</aut_NC>
       <fornitore>Stefania,Ferraris</fornitore>
       <trascr>Stefania,Ferraris</trascr>
       <data>?,12,02</data>
       <luogo>Vercelli,IT</luogo>
       <ist>scuola</ist>
       <ist_nome>Università del Piemonte Orientale</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>2</status>
       <annualita>1</annualita>
       <lingua1>tedesco,0</lingua1>
       <lingue>inglese,francese</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>6,Vercelli</permanenza>
       <esposizione>sc,am,fam,med</esposizione>
   </autore>
   <testo>
       <tipo_forma>c-lib_descr</tipo_forma>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>orig</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>stefaniaferraris_FT.txt,0,0</stel>
       <cons>Italia_C.txt</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ <titolo>Quello che mi è piaciuto e quello che non mi è piaciuto in <topn>Italia</topn></titolo>

#001# Sono tante le cose che mi hanno piaciuto e che mi
piaciono anche adesso dell' <topn>Italia</topn> ma altretante
quelle che non mi piaciono .#
Inanzitutto mi piace la lingua italiana , cosa che
mi ha portato da venir qui - penso che sia
molto armonica e più musicale
in contrasto con la lingua tedesca .#
Poi adoro la cucina sopratutta la facilità
con chi si può preparare una cosa
così gustosa .#
Naturalmente con dei cibi così buoni non
può mancare il vino che , secondo me , è
```

```
eccezionale.#
Mi piace tanto anche la regione in cui vivo ,
<topn>Piemonte</topn> , perché in poco tempo sono al
mare e se mi annoio posso cambiare e
sono velocissimo nella montagna dove posso
$002$ sciare .#
Inoltre mi piaciono tantissimo i miei amici ,
le persone con cui vivo insieme qui
a <topn>Vercelli</topn> e i ragazzi spagnoli chi ho incontrato#
al corso di italiano .#
Essi sono sicuramente le cose migliore
che sono successe a me in <topn>Italia</topn> fino adesso !
La prima cosa che ho notato venendo in <topn>Italia</topn>
e non mi piace è il fatto di dover pagare
per le autostrade italiane .#
Penso anche che il caos dell' Università
qui a <topn>Vercelli</topn> insieme alla burocrazia italiana
sono veramente faticosi .#
La cosa che mi manca di più della <topn>Germania</topn>
è il nostro <date_0000,12,25>Natale</date> !#
È tutto diverso con le abitudine , i colori
che adesso non sto sentendo veramente
come i giorni di <date_0000,12,25>Natale</date> sono vicini .#

#002# Ma sono sicurissima che prenderò
tanti bellissimi pensieri a tutto che ho
visto , ho fatto e chi ho incontrato
quando tornerò in <topn>Germania</topn> !#
</BODY>
```

6.1.3 versione FS, `stefania001_TTM_02.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Alessandra,Vogels</aut_NC>
       <fornitore>Stefania,Ferraris</fornitore>
       <trascr>Stefania,Ferraris</trascr>
       <data>?,12,02</data>
       <luogo>Vercelli,IT</luogo>
       <ist>scuola</ist>
       <ist_nome>Università del Piemonte Orientale</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>2</status>
       <annualita>1</annualita>
       <lingua1>tedesco,0</lingua1>
       <lingue>inglese,francese</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>6,Vercelli</permanenza>
       <esposizione>sc,am,fam,med</esposizione>
   </autore>
   <testo>
       <tipo_forma>c-lib_descr</tipo_forma>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>orig</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>stefaniaferraris_FT.txt,0,0</stel>
       <cons>Italia_C.txt</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
<tLn nr=1>
$001$ <titolo>Quello che mi è piaciuto e quello che non mi è piaciuto in <topn>Italia</topn></titolo>
```

```
</tLn>
<eLn>1</eLn>
<tLn nr=2>
#001# Sono tante le cose che mi hanno piaciuto e che mi
</tLn>
<tLn nr=3>
piaciono anche adesso dell’ <topn>Italia</topn> ma altretante
</tLn>
<tLn nr=4>
quelle che non mi piaciono .#
</tLn>
<tLn nr=5>
Inanzitutto mi piace la lingua italiana , cosa che
</tLn>
<tLn nr=6>
mi ha portato da venir qui - penso che sia
</tLn>
<tLn nr=7>
molto armonica e più musicale
</tLn>
<tLn nr=8>
in contrasto con la lingua tedesca .#
</tLn>
<tLn nr=9>
Poi adoro la cucina sopratutta la facilità
</tLn>
<tLn nr=10>
con chi si può preparare una cosa
</tLn>
<tLn nr=11>
così gustosa .#
</tLn>
<tLn nr=12>
Naturalmente con dei cibi così buoni non
</tLn>
<tLn nr=13>
può mancare il vino che , secondo me , è
</tLn>
<tLn nr=14>
eccezionale.#
</tLn>
<tLn nr=15>
Mi piace tanto anche la regione in cui vivo ,
</tLn>
<topn>Piemonte</topn> , perché in poco tempo sono al
<tLn nr=16>
mare e se mi annoio posso cambiare e
</tLn>
<tLn nr=17>
sono velocissimo nella montagna dove posso
</tLn>
<tLn nr=18>
$002$ sciare .#
</tLn>
<tLn nr=19>
Inoltre mi piaciono tantissimo i miei amici ,
</tLn>
<tLn nr=20>
le persone con cui vivo insieme qui
</tLn>
<tLn nr=21>
a <topn>Vercelli</topn> e i ragazzi spagnoli chi ho incontrato#
</tLn>
<tLn nr=22>
al corso di italiano .#
</tLn>
<tLn nr=23>
Essi sono sicuramente le cose migliore
</tLn>
<tLn nr=24>
che sono successe a me in <topn>Italia</topn> fino adesso !
</tLn>
<tLn nr=25>
La prima cosa che ho notato venendo in <topn>Italia</topn>
</tLn>
<tLn nr=26>
e non mi piace è il fatto di dover pagare
</tLn>
<tLn nr=27>
per le autostrade italiane .#
</tLn>
<tLn nr=28>
Penso anche che il caos dell' Università
</tLn>
```

```
<tLn nr=29>
qui a <topn>Vercelli</topn> insieme alla burocrazia italiana
</tLn>
<tLn nr=30>
sono veramente faticosi .#
</tLn>
<tLn nr=31>
La cosa che mi manca di più della <topn>Germania</topn>
</tLn>
<tLn nr=32>
è il nostro <date_0000,12,25>Natale</date> !#
</tLn>
<tLn nr=33>
È tutto diverso con le abitudine , i colori
</tLn>
<tLn nr=34>
che adesso non sto sentendo veramente
</tLn>
<tLn nr=35>
come i giorni di <date_0000,12,25>Natale</date> sono vicini .#
</tLn>
<eLn>1</eLn>
<tLn nr=36>
#002# Ma sono sicurissima che prenderò
</tLn>
<tLn nr=37>
tanti bellissimi pensieri a tutto che ho
</tLn>
<tLn nr=38>
visto , ho fatto e chi ho incontrato
</tLn>
<tLn nr=39>
quando tornerò in <topn>Germania</topn> !#
</tLn>
</BODY>
```

6.2 Un secondo esempio, appena più complesso è il seguente, sempre contribuito da Stefania Ferraris.

6.2.1 versione TD, `stefania002_TD.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Estefania,Pleite</aut_NC>
       <fornitore>Stefania,Ferraris</fornitore>
       <trascr>Stefania,Ferraris</trascr>
       <data>????,12,02</data>
       <luogo>Vercelli,IT</luogo>
       <ist>scuola</ist>
       <ist_nome>Università del Piemonte Orientale</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,gn</gruppo_num>
       <gruppo_nome>finecorso</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>2</status>
       <annualita>?</annualita>
       <lingua1>spagnolo,0</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>6,Vercelli</permanenza>
       <esposizione>sc,am,fam,med</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</topics>
       <test>provadifine</test>
       <qualita>orig</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>stefaniaferraris_FT.txt,finecorso_G.txt,provadifine_P.txt</stel>
       <cons>temaitalia_C.txt</cons>
       <txtext>0</txext>
       <imgext>0</imgext>
```

```
       <txtint>0</txint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ Quello che mi è piaciuto e quello che non mi è piaciuto in Italia

In Italia mi è {h} piaciuto la gente; quando io faccio
una domanda qualcuno {altra gente} la risposta sempre è molto gentile. Il gelato
di Italia mi piace moltissimo anche perché è grande
e sta buonissimo. Vercelli mi sembra una città
molto tranquilla e bella. (Ma forse troppo tranqui|lla).
Adesso con le luce di Natale sta più bella
che prima. Vercelli si trova tra Torino e Milano;
questo è un'altra cosa che mi piace anche perché
così io posso visitare Torino e Milano. Non mi è
piaciuto il freddo, ma ancora penso che {00} non è arrivato
il vero freddo. In Febrario sarà peggiore ... Mamma mia!!
</BODY>
```

6.2.2 versione TTM, `stefania002_TTM.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Estefania,Pleite</aut_NC>
       <fornitore>Stefania,Ferraris</fornitore>
       <trascr>Stefania,Ferraris</trascr>
       <data>????,12,02</data>
       <luogo>Vercelli,IT</luogo>
       <ist>scuola</ist>
       <ist_nome>Università del Piemonte Orientale</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,gn</gruppo_num>
       <gruppo_nome>finecorso</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>2</status>
       <annualita>?</annualita>
       <lingua1>spagnolo,0</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>6,Vercelli</permanenza>
       <esposizione>sc,am,fam,med</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</topics>
       <test>provadifine</test>
       <qualita>orig</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>stefaniaferraris_FT.txt,finecorso_G.txt,provadifine_P.txt</stel>
       <cons>temaitalia_C.txt</cons>
       <txtext>0</txext>
       <imgext>0</imgext>
       <txtint>0</txint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ <titolo>Quello che mi è piaciuto e quello che non mi è piaciuto in <topn>Italia</topn></titolo>

In Italia mi è <CORR>h</CORR> piaciuto la gente ; quando io faccio
una domanda qualcuno <CORR>altra gente</CORR> la risposta sempre è molto gentile . Il gelato
di <topn>Italia</topn> mi piace moltissimo anche perché è grande
e sta buonissimo . <topn>Vercelli</topn> mi sembra una città
molto tranquilla e bella . ( Ma forse troppo tranqui|lla ) .
Adesso con le luce di <date_0000,12,25>Natale</date> sta più bella
che prima. <topn>Vercelli</topn> si trova tra <topn>Torino</topn> e <topn>Milano</topn> ;
questo è un' altra cosa che mi piace anche perché
così io posso visitare <topn>Torino</topn> e <topn>Milano</topn> . Non mi è
piaciuto il freddo , ma ancora penso che <INS>penso che</INS> non è arrivato
```

```
il vero freddo . In <date_0000,02,00>Febrario</date> sarà peggiore ... Mamma mia !!#
</BODY>
```

6.2.3 versione FS, `stefania002_TTM_02.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Estefania,Pleite</aut_NC>
       <fornitore>Stefania,Ferraris</fornitore>
       <trascr>Stefania,Ferraris</trascr>
       <data>????,12,02</data>
       <luogo>Vercelli,IT</luogo>
       <ist>scuola</ist>
       <ist_nome>Università del Piemonte Orientale</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,gn</gruppo_num>
       <gruppo_nome>finecorso</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>2</status>
       <annualita>?</annualita>
       <lingua1>spagnolo,0</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>6,Vercelli</permanenza>
       <esposizione>sc,am,fam,med</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</topics>
       <test>provadifine</test>
       <qualita>orig</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>stefaniaferraris_FT.txt,finecorso_G.txt,provadifine_P.txt</stel>
       <cons>temaitalia_C.txt</cons>
       <txtext>0</txext>
       <imgext>0</imgext>
       <txtint>0</txint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
<tLn nr=1>
$001$ <titolo>Quello che mi è piaciuto e quello che non mi è piaciuto in <topn>Italia</topn></titolo>
</tLn>
<eLn>1</eLn>
<tLn nr=2>
In Italia mi è <CORR>h</CORR> piaciuto la gente ; quando io faccio
</tLn>
<tLn nr=3>
una domanda qualcuno <CORR>altra gente</CORR> la risposta sempre è molto gentile . Il gelato
</tLn>
<tLn nr=4>
di <topn>Italia</topn> mi piace moltissimo anche perché è grande
</tLn>
<tLn nr=5>
e sta buonissimo . <topn>Vercelli</topn> mi sembra una città
</tLn>
<tLn nr=6>
molto tranquilla e bella . ( Ma forse troppo tranqui|lla ) .
</tLn>
<tLn nr=7>
Adesso con le luce di <date_0000,12,25>Natale</date> sta più bella
</tLn>
<tLn nr=8>
che prima. <topn>Vercelli</topn> si trova tra <topn>Torino</topn> e <topn>Milano</topn> ;
</tLn>
<tLn nr=9>
questo è un' altra cosa che mi piace anche perché
</tLn>
<tLn nr=10>
```

così io posso visitare <topn>Torino</topn> e <topn>Milano</topn> . Non mi è
</tLn>
<tLn nr=11>
piaciuto il freddo , ma ancora penso che <INS>penso che</INS> non è arrivato
</tLn>
<tLn nr=12>
il vero freddo . In <date_0000,02,00>Febrario</date> sarà peggiore ... Mamma mia !!#
</tLn>
</BODY>

6.3 Un terzo esempio è poi il seguente trascritto da Carla Marello.

6.3.1 Versione TD, `tanya_carla001_TD.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Yutrai,?</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Carla,Marello</trascr>
       <data>2003,01,09</data>
       <luogo>New Dehli,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-B.A.</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>m</specifiche>
       <eta>19-25</eta>
       <status>3</status>
       <annualita>1</annualita>
       <lingua1>hindi,inglese</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?,?</permanenza>
       <esposizione>?</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>?</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>tanyaroy_F.txt,carlamarello_T.txt,0,0</stel>
       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ Ho fatto molte cose in la vacanza.
Prima sono andato i parenti con
miei poi siamo andati al cinema.
il Film si chiama “Die another
day” è il Film di James Bond.
Sono andato in la Resturante con
miei {mio} amici per mangare
e bere. abbiamo {abbia,x}
mangato molte cose in La
Resturante i.e. carne, pane, pizza
Dolci {Dolce} e abbiamo benuto
Martini e Bacardi poi
siamo {sono} andati in discoteca
per ballare.

                vorrei spendere
molto tempo su studio {(durante
a capodano)} e sempre io {ix} {x}
sento contento durante a
capodano.
</BODY>
```

6.3.2 versione TTM, `tanya_carla001_TTM.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Yutrai,?</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Carla,Marello</trascr>
       <data>2003,01,09</data>
       <luogo>New Dehli,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-B.A.</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>m</specifiche>
       <eta>19-25</eta>
       <status>3</status>
       <annualita>1</annualita>
       <lingua1>hindi,inglese</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?,?</permanenza>
       <esposizione>?</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>?</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>tanyaroy_F.txt,carlamarello_T.txt,0,0</stel>
       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ #001# Ho fatto molte cose in la vacanza .
Prima sono andato i parenti con
miei poi siamo andati al cinema .
il Film si chiama " <lng_inglese>Die another
day</lng> " è il Film di <anth>James Bond</anth> .
Sono andato in la Resturante con
miei <CORR>mio</CORR> amici per mangare
e bere . abbiamo <CORR>abbia,x</CORR>
mangato molte cose in La
Resturante i.e. carne , pane , pizza
Dolci <CORR>Dolce</CORR> e abbiamo benuto
Martini e Bacardi poi
siamo <CORR>sono</CORR> andati in discoteca
per ballare .#

#002# <blank_3></blank> vorrei spendere
molto tempo su studio <CORR>( durante
a <date_0000,01,01>capodano</date> )</CORR> e sempre io <CORR>ix,x</CORR>
sento contento durante a
<date_0000,01,01>capodano</date> .
</BODY>
```

6.3.3 versione FS, `tanya_carla001_TTM_02.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Yutrai,?</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Carla,Marello</trascr>
```

```
        <data>2003,01,09</data>
        <luogo>New Dehli,IN</luogo>
        <ist>scuola</ist>
        <ist_nome>Delhi University-B.A.</ist_nome>
    </doc-id>
    <set-id>
        <corpus>valico</corpus>
        <gruppo_num>1,g1</gruppo_num>
        <gruppo_nome>0</gruppo_nome>
    </set-id>
    <autore>
        <specifiche>m</specifiche>
        <eta>19-25</eta>
        <status>3</status>
        <annualita>1</annualita>
        <lingua1>hindi,inglese</lingua1>
        <lingue>?</lingue>
        <scolarizzazione>un</scolarizzazione>
        <permanenza>?,?</permanenza>
        <esposizione>?</esposizione>
    </autore>
    <testo>
         <tipo_forma>c-lib_descr</tipo>
        <topics>   </topics>
        <keyw>(____,____,____,____)</keyw>
        <test>?</test>
        <qualita>origFC</qualita>
        <esecuzione>ms</esecuzione>
    </testo>

    <ref>
        <stel>tanyaroy_F.txt,carlamarello_T.txt,0,0</stel>
        <cons>0</cons>
        <txtext>0</txtext>
        <imgext>0</imgext>
        <txtint>0</txtint>
        <imgint>0</imgint>
    </ref>
</HEAD>
<BODY>
<tLn nr=1>
$001$ #001# Ho fatto molte cose in la vacanza .
</tLn>
<tLn nr=2>
Prima sono andato i parenti con
</tLn>
<tLn nr=3>
miei poi siamo andati al cinema .
</tLn>
<tLn nr=4>
il Film si chiama “ <lng_inglese>Die another
</tLn>
<tLn nr=5>
day</lng> ” è il Film di <anth>James Bond</anth> .
</tLn>
<tLn nr=6>
Sono andato in la Resturante con
</tLn>
<tLn nr=7>
miei <CORR>mio</CORR> amici per mangare
</tLn>
<tLn nr=8>
e bere . abbiamo <CORR>abbia,x</CORR>
</tLn>
<tLn nr=9>
mangato molte cose in La
</tLn>
<tLn nr=10>
Resturante i.e. carne , pane , pizza
</tLn>
<tLn nr=11>
Dolci <CORR>Dolce</CORR> e abbiamo benuto
</tLn>
<tLn nr=12>
Martini e Bacardi poi
</tLn>
<tLn nr=13>
siamo <CORR>sono</CORR> andati in discoteca
</tLn>
<tLn nr=14>
per ballare .#
</tLn>
<eLn>1</eLn>
<tLn nr=15>
```

```
#002# <blank_3></blank> vorrei spendere
</tLn>
<tLn nr=16>
molto tempo su studio <CORR>( durante
</tLn>
<tLn nr=17>
a <date_0000,01,01>capodano</date> )</CORR> e sempre io <CORR>ix,x</CORR>
</tLn>
<tLn nr=18>
sento contento durante a
</tLn>
<date_0000,01,01>capodano</date> .
</BODY>
```

6.4 Un quarto e più complesso esempio è il seguente trascritto da Valeria Saggiotto.

6.4.1 versione TD, `tanya_valeria001_TD.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Shipra,Kapoor</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Valeria,saggiotto</trascr>
       <data>2003,01,17</data>
       <luogo>New Dehli,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-B.A.</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g5</gruppo_num>
       <gruppo_nome>miamadre</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>?</status>
       <annualita>1</annualita>
       <lingua1>hindi,inglese</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?</permanenza>
       <esposizione>sc</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>tanyaroy_F.txt,valeriasaggiotto_T.txt,miamadre_G.txt,0</stel>
       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ Mia madre

#001# Il nome di mia madre è Harsh
Kapoor. Lei ha {è} fatto B.Ed e M.A. Mia
madre è casa linga *. Mia madre
è bellissima e molto simpatica. Lei
è di Punjab. Mia madre ha {è}
quaranta cinque anni. Mia madre
è molto {x} cura di me. Quando sono
venuta la sera si preparara il cibo
per noi. Ancora ogni giorni {00} si alza presto per
me. Lei è molto buona cucina.
Mia madre è a migliore amica a me
Io {0} condevido tutti i problemi e sentimenti {00} con mia
madre. Lei risolve gli certamente
il meglio. Lei è mi inspirazione.
Mia madre è migliore donna nel
```

```
mondo e amo mia madre molto
* era una professoressa ma non a
adesso.
</BODY>
```

6.4.2 versione TTM, `tanya_valeria001_TTM.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Shipra,Kapoor</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Valeria,saggiotto</trascr>
       <data>2003,01,17</data>
       <luogo>New Dehli,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-B.A.</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g5</gruppo_num>
       <gruppo_nome>miamadre</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>?</status>
       <annualita>1</annualita>
       <lingua1>hindi,inglese</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?</permanenza>
       <esposizione>sc</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>tanyaroy_F.txt,valeriasaggiotto_T.txt,miamadre_G.txt,0</stel>
       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ <titolo><emph_sc,u1>Mia Madre</emph></titolo>

Il nome di mia madre è <anth>Harsh
Kapoor</anth> . Lei ha <CORR>è</CORR> fatto B.Ed e M.A . Mia
madre è casa+linga <A href="#star_1">*</A> . Mia madre
è bellissima e molto simpatica. Lei
è di <topn>Punjab</topn>. Mia madre ha <CORR>è</CORR>
quaranta cinque anni . Mia madre
è molto <CORR>x</CORR> cura di me . Quando sono
venuta la sera si preparara il cibo
per noi . Ancora ogni giorni <INS>ogni giorni</INS> si alza presto per
me . Lei è molto buona cucina .
Mia madre è a migliore amica a me .
Io <CORR>x</CORR> condevido tutti i problemi e sentimenti <INS>e sentimenti</INS> con mia
madre . Lei risolve gli certamente
il meglio . Lei è mi inspirazione .
Mia madre è migliore donna nel
mondo e amo mia madre molto .#
<calce>< A name=star_1>*</A > era una professoressa ma non a
adesso </calce>.
</BODY>
```

6.4.3 versione FS, `tanya_valeria001_TTM_02.TXT`:

```
<HEAD>
```

```
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Shipra,Kapoor</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Valeria,saggiotto</trascr>
       <data>2003,01,17</data>
       <luogo>New Dehli,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-B.A.</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g5</gruppo_num>
       <gruppo_nome>miamadre</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>19-25</eta>
       <status>?</status>
       <annualita>1</annualita>
       <lingua1>hindi,inglese</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?</permanenza>
       <esposizione>sc</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
       <stel>tanyaroy_F.txt,valeriasaggiotto_T.txt,miamadre_G.txt,0</stel>

       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
<tLn nr=1>
$001$ <titolo><emph_sc,u1>Mia Madre</emph></titolo>
</tLn>
<eLn>1</eLn>
<tLn nr=2>
Il nome di mia madre è <anth>Harsh
</tLn>
<tLn nr=3>
Kapoor</anth> . Lei ha <CORR>è</CORR> fatto B.Ed e M.A . Mia
</tLn>
<tLn nr=4>
madre è casa+linga <A href="#star_1">*</A> . Mia madre
</tLn>
<tLn nr=5>
è bellissima e molto simpatica. Lei
</tLn>
<tLn nr=6>
è di <topn>Punjab</topn>. Mia madre ha <CORR>è</CORR>
</tLn>
<tLn nr=7>
quaranta cinque anni . Mia madre
</tLn>
<tLn nr=8>
è molto <CORR>x</CORR> cura di me . Quando sono
</tLn>
<tLn nr=9>
venuta la sera si preparara il cibo
</tLn>
<tLn nr=10>
per noi . Ancora ogni giorni <INS>ogni giorni</INS> si alza presto per
</tLn>
<tLn nr=11>
me . Lei è molto buona cucina .
</tLn>
<tLn nr=12>
Mia madre è a migliore amica a me .
</tLn>
```

```
<tLn nr=13>
Io <CORR>x</CORR> condevido tutti i problemi e sentimenti <INS>e sentimenti</INS> con mia
</tLn>
<tLn nr=14>
madre . Lei risolve gli certamente
</tLn>
<tLn nr=15>
il meglio . Lei è mi inspirazione .
</tLn>
<tLn nr=16>
Mia madre è migliore donna nel
</tLn>
<tLn nr=17>
mondo e amo mia madre molto .#
</tLn>
<calce>< A name=star_1>*</A > era una professoressa ma non a
<tLn nr=18>
adesso </calce>.
</tLn>
</BODY>
```

6.5 Un esempio di email è il seguente, fornito e trascritto da Silvia Camarca.

6.5.1 versione TD, `silvia001_TD.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Harolyn,Pinson</aut_NC>
       <fornitore>Silvia,Camarca</fornitore>
       <trascr>Silvia,Camarca</trascr>
       <data>2003,01,??</data>
       <luogo>Val della Torre,IT</luogo>
       <ist>0</ist>
       <ist_nome>0</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>oltre</eta>
       <status>1</status>
       <annualita>1</annualita>
       <lingua1>inglese,0</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>8,Val della Torre</permanenza>
       <esposizione>sc,med</esposizione>
   </autore>
   <testo>
        <tipo_forma>email</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>origCE</qualita>
       <esecuzione>wp</esecuzione>
   </testo>
   <ref>
       <stel>silviacamarca_FT.txt,0,0</stel>
       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ Ciao!

Buon Anno a te!!  I biscotti errano squisito - grazie a tua
mamma!

Becki ama Italia!  Non ha voluto tornare in America!  Grazie per il
tuo aiuto a Roma.

Si, fa freddo in Val Della Torre, ma non c'è neve - ancora!
```

```
Per la settimana sono malata.  Il mio stomaco non sta bene.  Forse
è meglio aspettare fino la prossima settimana per le lezioni.
Lunedi?

Ci sono le fotografie atoccato.  La chiesa è in Francia - siamo
andati li prima di Natale.  Abbiamo scalato fino la chiesa.
L'altre fotographie sono a casa - Becki e Katie.

Non so quando torneramo in America.  Ancora non abbiamo i documenti
da Italia.

Ci vediamo!

Harolyn
</BODY>
```

6.5.2 versione TTM, `silvia001_TTM.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Harolyn,Pinson</aut_NC>
       <fornitore>Silvia,Camarca</fornitore>
       <trascr>Silvia,Camarca</trascr>
       <data>2003,01,??</data>
       <luogo>Val della Torre,IT</luogo>
       <ist>0</ist>
       <ist_nome>0</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>oltre</eta>
       <status>1</status>
       <annualita>1</annualita>
       <lingua1>inglese,0</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>8,Val della Torre</permanenza>
       <esposizione>sc,med</esposizione>
   </autore>
   <testo>
        <tipo_forma>email</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>origCE</qualita>
       <esecuzione>wp</esecuzione>
   </testo>
   <ref>
       <stel>silviacamarca_FT.txt,0,0</stel>
       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
$001$ <pcoll>Ciao !</pcoll>

#001# Buon Anno a te !! I biscotti errano squisito - grazie a tua
mamma !#

#002#<anth>Becki</anth> ama <topn>Italia</topn> ! Non ha voluto tornare in <topn>America</topn> !
Grazie per il tuo aiuto a <topn>Roma</topn> .#

#003#Si, fa freddo in <topn>Val Della Torre</topn>, ma non c' è neve - ancora !#

#004#Per la settimana sono malata . Il mio stomaco non sta bene .
Forse è meglio aspettare fino la prossima settimana per le lezioni .
Lunedi ?#

#004#Ci sono le fotografie atoccato . La chiesa è in <topn>Francia</topn> -
siamo andati li prima di <date_2002,12,25>Natale</date> . Abbiamo scalato fino la chiesa .
```

```
L' altre fotographie sono a casa - <anth>Becki</anth> e <anth>Katie</anth> .#

#005#Non so quando torneramo in <topn>America</topn> . Ancora non abbiamo i documenti
da <topn>Italia</topn> .#

<ecoll>Ci vediamo !

<anth>Harolyn</anth></ecoll>
</BODY>
```

6.5.3 versione FS, `silvia001_TTM_02.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Harolyn,Pinson</aut_NC>
       <fornitore>Silvia,Camarca</fornitore>
       <trascr>Silvia,Camarca</trascr>
       <data>2003,01,??</data>
       <luogo>Val della Torre,IT</luogo>
       <ist>0</ist>
       <ist_nome>0</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0</gruppo_nome>
   </set-id>
   <autore>
       <specifiche>f</specifiche>
       <eta>oltre</eta>
       <status>1</status>
       <annualita>1</annualita>
       <lingua1>inglese,0</lingua1>
       <lingue>?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>8,Val della Torre</permanenza>
       <esposizione>sc,med</esposizione>
   </autore>
   <testo>
        <tipo_forma>email</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>0</test>
       <qualita>origCE</qualita>
       <esecuzione>wp</esecuzione>
   </testo>
   <ref>
       <stel>silviacamarca_FT.txt,0,0</stel>
       <cons>0</cons>
       <txtext>0</txtext>
       <imgext>0</imgext>
       <txtint>0</txtint>
       <imgint>0</imgint>
   </ref>
</HEAD>
<BODY>
<tLn nr=1>
$001$ <pcoll>Ciao !</pcoll>
</tLn>
<eLn>1</eLn>
<tLn nr=2>
#001# Buon Anno a te !! I biscotti errano squisito - grazie a tua
</tLn>
<tLn nr=3>
mamma !#
</tLn>
<eLn>1</eLn>
<tLn nr=4>
#002#<anth>Becki</anth> ama <topn>Italia</topn> ! Non ha voluto tornare in <topn>America</topn> !
</tLn>
<tLn nr=5>
Grazie per il tuo aiuto a <topn>Roma</topn> .#
</tLn>
<eLn>1</eLn>
<tLn nr=6>
#003#Si, fa freddo in <topn>Val Della Torre</topn>, ma non c' è neve - ancora !#
</tLn>
<eLn>1</eLn>
```

```
<tLn nr=7>
#004#Per la settimana sono malata . Il mio stomaco non sta bene .
</tLn>
<tLn nr=8>
Forse è meglio aspettare fino la prossima settimana per le lezioni .
</tLn>
<tLn nr=9>
Lunedi ?#
</tLn>
<eLn>1</eLn>
<tLn nr=10>
#004#Ci sono le fotografie atoccato . La chiesa è in <topn>Francia</topn> -
</tLn>
<tLn nr=11>
siamo andati li prima di <date_2002,12,25>Natale</date> . Abbiamo scalato fino la chiesa .
</tLn>
<tLn nr=12>
L' altre fotographie sono a casa - <anth>Becki</anth> e <anth>Katie</anth> .#
</tLn>
<eLn>1</eLn>
<tLn nr=13>
#005#Non so quando torneramo in <topn>America</topn> . Ancora non abbiamo i documenti
</tLn>
<tLn nr=14>
da <topn>Italia</topn> .#
</tLn>
<ecoll>Ci vediamo !
<anth>Harolyn</anth></ecoll>
</BODY>
```

6.6 Un sesto e più complesso esempio è il seguente trascritto da Francesca Minozzi.

6.6.1 versione TD, `tanya_francesca001_TD.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Rengal,?</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Francesca,Minozzi</trascr>
       <data>????,??,??</data>
       <luogo>New Delhi,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-Diploma</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0<gruppo_nome>
   </set-id>
   <autore>
       <specifiche>?</specifiche>
       <eta>19-25</eta>
       <status>?</status>
       <annualita>3</annualita>
       <lingua1>Punjabi,Hindi</lingua1>
       <lingue>Inglese,?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?,?</permanenza>
       <esposizione>sc</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>?</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
        <stel>tanyaroy_F.txt,francescaminozzi_T.txt,0,0</stel>
        <cons>matrimonio_C.txt</cons>
        <txtext>0</txtext>
        <imgext>0</imgext>
        <txtint>0</txtint>
   </ref>
</HEAD>
<BODY>
$001$ 2) Descrivete la situazione indiana per quanto riguarda l'isti|tuzione
  del matrimonio.
```

```
-› In India l'istituzione del matrimonio è molto forte,
invece in Italia perché nel India ci sono molti religioni,
hanno molte culture , le tradizioni e i costumi. Ogni
religione ha i modi diversi per fare il matrimonio. Il
matrimonio è necessario per vivere insieme perché la società
non permette a nessuno vivere con una ragazza o un rag|azzo.
Le tradizini di matrimonio sono molte buoni. Il
matrimonio in India sembra come una festa perché tutte
le persone conosciute vengono ad attendere il matrimonio
in tutte le religioni. In India i parenti dei sposi scegli|ono
un ragazzo o una ragazza. Ma in questi giorni
il modo è un pò cambiato perché la cultura di occidentale
ha effetto i giovani indiani molto. Vogliono fare matrimonio
d'amore , ma la società indiana non permette questi tipi
di matrimonio.

3) Quale o quali lingue parli tu ? E quelli intorno a te ,
  i tuoi amici e parenti ? Commenta sulla realtà linguistica
  indiana.

-› Parlo 4 lingue , si chiamano l' inglese, l' indi , Punjabi
e l' italiano. Ma in questi giorni mi piace l'italiano
molto perché studio l' italiano all' università di Delhi
È una opportunità per conoscere le lingue stranieri . Si
può imparare molte lingue qui come spagnolo , francese
Germania e portoghese ecc. La mia madrelingua è
hindi. Le lingue ufficiali sono hindi e inglese. I
miei parenti parlano solo hindi ma mia sorella e mio
fratello parlano anche l'inglese . Ci sono molte lingue
locali in India come Garwali, Bhojpuri , Magdi , Avdi,
Rajasthani , Maithili , Telgu , Tanul , Urdu ecc.

$002$ 4) Chi non sogna di vincere una lotteria ? Alla televisione
  abbiamo visto “Kaum banega karorpati” in India. La televi|sione
  ti sembra un sogno o un incubo?

-› Oggi la televisione è una necessità per la gente.
Si può guardare la t.v per tre regioni -› La prima
è per imparare le nuove cose che successo nel
mondo . La seconda è per dimenticare i suoi problemi
e la terza è per divertirsi. Quando qualcuno {qualcumo} è triste ,
deve guardare la t.v. Ogni persona vuole vincere
una lotteria. In India ho visto “ Kaun Banega Karorpati”
nella t.v. perché è un modo giusto per ottenere la conosc|enza
e anche i soldi. Mi sembra la televisione come
un sogno .

[5] ) Ti ricordi di quando avevi due anni ? Scrivi della cosa
  che ti piaceva di più e quella che odiavi di più.

[-›] Quando ero bambino mi piaceva mangiare i cioccolati.
Tutto il tempo io mangiavo i cioccolati perché era molto
buoni . Il sapore era molto buono . Potevo fare tutto
per i cioccolati . Ogni giorno pensavo che abbia mangiato
4 O cinque (5) cioccolati.
                                        Ma non mi piaceva gioc [xxxxxxxx]
qualcuno perché volevo giocare solo  {0} con mia mad [xxxxxxxx]
mia madre . Tutto il tempo volevo [xxxtxxxxxxxxxx]
madre [xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx]
[xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx]
</BODY>
```

6.6.2 versione TTM, `tanya_francesca-001_TTM.txt`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Rengal,?</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Francesca,Minozzi</trascr>
       <data>????,??,??</data>
       <luogo>New Delhi,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-Diploma</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
```

```
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0<gruppo_nome>
   </set-id>
   <autore>
       <specifiche>?</specifiche>
       <eta>19-25</eta>
       <status>?</status>
       <annualita>3</annualita>
       <lingua1>Punjabi,Hindi</lingua1>
       <lingue>Inglese,?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?,?</permanenza>
       <esposizione>sc</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>?</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>
        <stel>tanyaroy_F.txt,francescaminozzi_T.txt,0,0</stel>
        <cons>matrimonio_C.txt</cons>
        <txtext>0</txtext>
        <imgext>0</imgext>
        <txtint>0</txtint>
   </ref>
</HEAD>
<BODY>
$001$ <mat>2 )</mat> <titolo><docente>Descrivete la situazione indiana per quanto riguarda l' isti|tuzione
<blank_1>del matrimonio .</blank></docente></titolo>

-› In <topn>India</topn> l' istituzione del matrimonio è molto forte ,
invece in <topn>Italia</topn> perché nel <topn>india</topn> ci sono molti religioni ,
hanno molte culture , le tradizioni e i costumi . Ogni
religione ha i modi diversi per fare il matrimonio . Il
matrimonio è necessario per vivere insieme perché la società
non permette a nessuno vivere con una ragazza o un rag|azzo .
Le tradizini di matrimonio sono molte buoni . Il
matrimonio in <topn>India</topn> sembra come una  festa perché tutte
le persone conosciute vengono ad attendere in matrimonio
in tutte le religioni . In <topn>india</topn> i parenti degli sposi scegli|ono
un ragazzo o una ragazza . Ma in questi giorni
il modo è un pò cambiato perché la cultura di occidentale
ha effetto i giovani indiani molto . Vogliono fare matrimonio
d' amore , ma la società indiana non permette questi tipi
di matrimonio .

<mat>3 )</mat> <titolo><docente>Quale o quali lingue parli tu ? E quelli intorno a te ,
<blank_1> i tuoi amici e parenti ? Commenta sulla realtà linguistica
indiana . </blank></docente></titolo>

-› Parlo 4 lingue , si chiamano l' inglese , l' indi , Punjabi
e l' italiano . Ma in questi giorni mi piace l' italiano
molto perché studio l' italiano all' università di <topn>Delhi</topn>
È una opportunità per conoscere le lingue stranieri . Si
può imparare molte lingue qui come spagnolo , francese
Germania e portoghese ecc. La mia madrelingua è
hindi . Le lingue ufficiali sono hindi e inglese . I
miei parenti parlano solo hindi ma mia sorella e mio
fratello parlano anche l' inglese . Ci sono molte lingue
locali in <topn>India</topn> come Garwali , Bhojpurii , Magdi , Avdi ,
Rajasthani ,Maithili , Telgu , Urdu ecc.

$002$ <mat>4 )</mat> <titolo><docente>Chi non sogna di vincere una lotteria ? Alla televisione
<blank_1> abbiamo visto <lng_hindi>“ Kaum banega karorpati ”</lng> in <topn>India</topn> . La televi|sione
ti sembra un sogno o un incubo ? </blank></docente></titolo>

-› Oggi la televisione è una necessità per la gente .
Si può guardare la t.v  per tre regioni -› La prima
è per imparare le nuove cose che successo nel
mondo . La seconda è per dimenticare i suoi problemi
e la terza è per divertirsi . Quando qualcuno <CORR<qualcumo></CORR> è triste ,
deve guardare la t.v. Ogni persona vuole vincere
una lotteria . In <topn>India</topn> ho visto <lng_hindi>“ Kaum Banega Karorpati ”</lng>
nella t.v. perché è un modo giusto per ottenere la conosc|enza
e anche i soldi . Mi sembra la televisione come
un sogno .

<mat><LAC>5</LAC>)</mat> <titolo><docente>Ti ricordi di quando avevi due anni ? Scrivi della cosa
```

```
<blank_1> che ti piaceva di più e quella che odiavi di più . </blank_1></docente></titolo>

<LAC>-›</LAC>Quando ero bambino mi piaceva mangiare i cioccolati
Tutto il tempo io mangiavo i cioccolati perché era molto
buoni . Il sapore era molto buono . Potevo fare tutto
per i cioccolati . Ogni giorno pensavo che abbia mangiato
4 o cinque ( <mat>5</mat> ) cioccolati .
<blank_3></blank> Ma non mi piaceva gioc <LAC>xxxxxxxx</LAC>
qualcuno perché volevo giocare solo  <INS>solo</INS> con mia mad <LAC>xxxxxxxx</LAC>
mia madre . Tutto il tempo volevo <LAC>xxxtxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</LAC>
madre <LAC>xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</LAC>
<LAC>xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</LAC>
</BODY>
```

6.6.3 versione FS, `tanya_francesca-001_TTM_02.TXT`:

```
<HEAD>
   <doc-id>
       <idN>-----</idN>
       <charset>ansi</charset>
       <lingua>italiano</lingua>
       <aut_NC>Rengal,?</aut_NC>
       <fornitore>Tanya,Roy</fornitore>
       <trascr>Francesca,Minozzi</trascr>
       <data>????,??,??</data>
       <luogo>New Delhi,IN</luogo>
       <ist>scuola</ist>
       <ist_nome>Delhi University-Diploma</ist_nome>
   </doc-id>
   <set-id>
       <corpus>valico</corpus>
       <gruppo_num>1,g1</gruppo_num>
       <gruppo_nome>0<gruppo_nome>
   </set-id>
   <autore>
       <specifiche>?</specifiche>
       <eta>19-25</eta>
       <status>?</status>
       <annualita>3</annualita>
       <lingua1>Punjabi,Hindi</lingua1>
       <lingue>Inglese,?</lingue>
       <scolarizzazione>un</scolarizzazione>
       <permanenza>?,?</permanenza>
       <esposizione>sc</esposizione>
   </autore>
   <testo>
        <tipo_forma>c-lib_descr</tipo>
       <topics>   </topics>
       <keyw>(____,____,____,____)</keyw>
       <test>?</test>
       <qualita>origFC</qualita>
       <esecuzione>ms</esecuzione>
   </testo>
   <ref>

        <stel>tanyaroy_F.txt,francescaminozzi_T.txt,0,0</stel>
        <cons>matrimonio_C.txt</cons>
        <txtext>0</txtext>
        <imgext>0</imgext>
        <txtint>0</txtint>
   </ref>
</HEAD>
<BODY>
<tLn nr=1>
$001$ <mat>2 )</mat> <titolo><docente>Descrivete la situazione indiana per quanto riguarda l' i-
sti|tuzione
</tLn>
<blank_1>del matrimonio .</blank></docente></titolo>
<eLn>1</eLn>
<tLn nr=2>
-› In <topn>India</topn> l' istituzione del matrimonio è molto forte ,
</tLn>
<tLn nr=3>
invece in <topn>Italia</topn> perché nel <topn>india</topn> ci sono molti religioni ,
</tLn>
<tLn nr=4>
hanno molte culture , le tradizioni e i costumi . Ogni
</tLn>
<tLn nr=5>
religione ha i modi diversi per fare il matrimonio . Il
</tLn>
```

<tLn nr=6>
matrimonio è necessario per vivere insieme perché la società
</tLn>
<tLn nr=7>
non permette a nessuno vivere con una ragazza o un rag|azzo .
</tLn>
<tLn nr=8>
Le tradizini di matrimonio sono molte buoni . Il
</tLn>
<tLn nr=9>
matrimonio in <topn>India</topn> sembra come una festa perché tutte
</tLn>
<tLn nr=10>
le persone conosciute vengono ad attendere in matrimonio
</tLn>
<tLn nr=11>
in tutte le religioni . In <topn>india</topn> i parenti degli sposi scegli|ono
</tLn>
<tLn nr=12>
un ragazzo o una ragazza . Ma in questi giorni
</tLn>
<tLn nr=13>
il modo è un pò cambiato perché la cultura di occidentale
</tLn>
<tLn nr=14>
ha effetto i giovani indiani molto . Vogliono fare matrimonio
</tLn>
<tLn nr=15>
d' amore , ma la società indiana non permette questi tipi
</tLn>
<tLn nr=16>
di matrimonio .
</tLn>
<mat>3)</mat> <titolo><docente>Quale o quali lingue parli tu ? E quelli intorno a te ,
<blank_1> i tuoi amici e parenti ? Commenta sulla realtà linguistica
<eLn>1</eLn>
<tLn nr=17>
indiana . </blank></docente></titolo>
</tLn>
<eLn>1</eLn>
<tLn nr=18>
-› Parlo 4 lingue , si chiamano l' inglese , l' indi , Punjabi
</tLn>
<tLn nr=19>
e l' italiano . Ma in questi giorni mi piace l' italiano
</tLn>
<tLn nr=20>
molto perché studio l' italiano all' università di <topn>Delhi</topn>
</tLn>
<tLn nr=21>
È una opportunità per conoscere le lingue stranieri . Si
</tLn>
<tLn nr=22>
può imparare molte lingue qui come spagnolo , francese
</tLn>
<tLn nr=23>
Germania e portoghese ecc. La mia madrelingua è
</tLn>
<tLn nr=24>
hindi . Le lingue ufficiali sono hindi e inglese . I
</tLn>
<tLn nr=25>
miei parenti parlano solo hindi ma mia sorella e mio
</tLn>
<tLn nr=26>
fratello parlano anche l' inglese . Ci sono molte lingue
</tLn>
<tLn nr=27>
locali in <topn>India</topn> come Garwali , Bhojpurii , Magdi , Avdi ,
</tLn>
<tLn nr=28>
Rajasthani ,Maithili , Telgu , Urdu ecc.
</tLn>
<eLn>1</eLn>
<tLn nr=29>
002 <mat>4)</mat> <titolo><docente>Chi non sogna di vincere una lotteria ? Alla televisione
</tLn>
<blank_1> abbiamo visto <lng_hindi>“ Kaum banega karorpati ”</lng> in <topn>India</topn> . La televi|sione
<tLn nr=30>
ti sembra un sogno o un incubo ? </blank></docente></titolo>
</tLn>
<eLn>1</eLn>
<tLn nr=31>
-› Oggi la televisione è una necessità per la gente .

```
</tLn>
<tLn nr=32>
Si può guardare la t.v  per tre regioni -› La prima
</tLn>
<tLn nr=33>
è per imparare le nuove cose che successo nel
</tLn>
<tLn nr=34>
mondo . La seconda è per dimenticare i suoi problemi
</tLn>
<tLn nr=35>
e la terza è per divertirsi . Quando qualcuno <CORR<qualcumo></CORR> è triste ,
</tLn>
<tLn nr=36>
deve guardare la t.v. Ogni persona vuole vincere
</tLn>
<tLn nr=37>
una lotteria . In <topn>India</topn> ho visto <lng_hindi>“ Kaum Banega Karorpati ”</lng>
</tLn>
<tLn nr=38>
nella t.v. perché è un modo giusto per ottenere la conosc|enza
</tLn>
<tLn nr=39>
e anche i soldi . Mi sembra la televisione come
</tLn>
<tLn nr=40>
un sogno .
</tLn>
<mat><LAC>5</LAC>)</mat> <titolo><docente>Ti ricordi di quando avevi due anni ? Scrivi della cosa
<blank_1> che ti piaceva di più e quella che odiavi di più . </blank_1></docente></titolo>
<LAC>-›</LAC>Quando ero bambino mi piaceva mangiare i cioccolati
<eLn>2</eLn>
<tLn nr=41>
Tutto il tempo io mangiavo i cioccolati perché era molto
</tLn>
<tLn nr=42>
buoni . Il sapore era molto buono . Potevo fare tutto
</tLn>
<tLn nr=43>
per i cioccolati . Ogni giorno pensavo che abbia mangiato
</tLn>
<tLn nr=44>
4 o cinque ( <mat>5</mat> ) cioccolati .
</tLn>
<blank_3></blank> Ma non mi piaceva gioc <LAC>xxxxxxxx</LAC>
<tLn nr=45>
qualcuno perché volevo giocare solo  <INS>solo</INS> con mia mad <LAC>xxxxxxxx</LAC>
</tLn>
<tLn nr=46>
mia madre . Tutto il tempo volevo <LAC>xxxtxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</LAC>
</tLn>
<tLn nr=47>
madre <LAC>xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</LAC>
</tLn>
<LAC>xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx</LAC>
</BODY>
```

7 Indice generale.