# I testi del web: una proposta di classificazione sulla base del corpus PAISÀ

Claudia Borghetti (Università di Bologna)
Sara Castagnoli (Università di Trento)
Marco Brunello (CNR di Pisa)

## 1. Introduzione

Il presente contributo si propone di condividere le finalità, la metodologia di sviluppo e gli esiti delle prime ricerche condotte sul corpus PAISÀ, un corpus di testi in italiano contemporaneo scaricati dal web, ideato con finalità glottodidattiche e di ricerca nell'ambito del progetto omonimo (§1.1). Presenteremo come il progetto si inserisce nel panorama, sempre più vasto, dei *web-derived corpora*, gli accorgimenti che sono risultati necessari in fase di creazione per evitare la spinosa questione del copyright (§1.2), e le ripercussioni che ciò ha avuto sui contenuti (§1.3). Ci concentreremo poi sui diversi livelli di annotazione che arricchiscono il corpus PAISÀ, soffermandoci in particolare sullo sforzo di classificazione dei testi per argomento, intenzione comunicativa e genere testuale (§2 e §3), tre parametri che, una volta trasformati in criteri di ricerca e esplorazione del corpus, permetteranno agli utenti – insegnanti di lingua in primis – una consultazione estremamente mirata e raffinata dei testi.

### *1.1 Il progetto PAISÀ: una visione d'insieme*

Obiettivo principale del progetto PAISÀ – *Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati* è la creazione di una risorsa a supporto dell'apprendimento/insegnamento dell'italiano come lingua straniera e/o etnica: il progetto si rivolge infatti principalmente a italiani di seconda generazione residenti all'estero, che mantengono l'italiano come lingua nativa in ambiti d'uso fortemente ridotti, e quelli di terza generazione per i quali l'italiano è lingua seconda (L2).[1] Pur essendo tale finalità didattica prioritaria, il corpus sviluppato potrà altresì essere messo a disposizione di studiosi di italianistica, linguistica applicata, computazionale ecc. per scopi di ricerca, anche statistica/quantitativa, sulla lingua italiana.

[1] Al progetto PAISÀ (2009-2012), cofinanziato dal Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) tramite il Fondo per gli Investimenti della Ricerca di Base (FIRB), collaborano quattro partner: Università di Bologna, CNR di Pisa, Accademia Europea di Bolzano e Università di Trento. Maggiori informazioni sono disponibili sul sito web del progetto http://www.corpusitaliano.it/.

Se da un lato è possibile affermare che l'utilizzo di corpora per l'insegnamento delle lingue rappresenta ormai una metodologia relativamente condivisa (come attestato dai contributi in volumi quali Sinclair 2004 e Aston et al. 2004), è pur tuttavia vero che ciò rimane appannaggio quasi esclusivo dell'insegnamento dell'inglese come lingua seconda, straniera o di specialità, e che per le altre lingue – tra cui l'italiano – strumenti e risorse *corpus-related* rimangono assai scarse. In aggiunta, la diffusione dei principali corpora di italiano attualmente esistenti – ad es. La Repubblica (Baroni et al., 2004), itWaC (Baroni et al., 2009) – può essere limitata da questioni o dubbi relativi alla possibilità di condivisione/distribuzione dei testi in essi contenuti (per copyright o privacy), nonché dal fatto di non essere specificamente concepiti per finalità didattiche e di mancare di interfacce adatte a utenti non esperti dell'utilizzo di corpora.

La novità introdotta dal progetto PAISÀ risiede nell'utilizzo di documenti non soggetti al classico vincolo di copyright "Tutti i diritti riservati" (implicito, se non diversamente indicato), ma distribuiti con licenze *Creative Commons* che permettono una maggiore libertà e flessibilità nella gestione della ridistribuzione dell'opera (§1.2). I testi sono stati reperiti sul web in maniera (semi)automatica utilizzando strumenti sviluppati essenzialmente nell'ambito della ricerca su *web as corpus* (§1.2), annotati linguisticamente e con informazioni quali argomento, intenzione comunicativa e genere testuale (indicati nel testo con l'espressione "metadati"); il corpus sarà reso accessibile attraverso un'interfaccia di ricerca multidimensionale specificamente concepita per studenti e insegnanti di italiano sul sito web del progetto (http://www.corpusitaliano.it), dal quale sarà inoltre possibile scaricare gratuitamente l'intero corpus.

### *1.2 Creazione e composizione del corpus PAISÀ*

La realizzazione del corpus PAISÀ ricalca la metodologia messa a punto nell'ambito del progetto WaCky (http://wacky.sslmit.unibo.it; Baroni et al., 2009), con il quale condivide l'obiettivo di creare corpora di lingua generale (non specialistici), di grandi dimensioni, utilizzando il web come fonte di dati linguistici. L'approccio prevede che si identifichino le url dei documenti da scaricare effettuando ricerche per combinazioni casuali di parole su un motore di ricerca; per PAISÀ, le parole utilizzate sono state tratte dal *Vocabolario di Base della Lingua Italiana* (De Mauro 1989),[2] organizzate in una lista di 50.000 coppie. Rispetto

[2] Disponibili online all'indirizzo http://ppbm.paravia.it/dib_lemmario.php.

ad altri *web corpora*, tuttavia, per avere la certezza che PAISÀ fosse legalmente ridistribuibile e sfruttando la possibilità offerta da Yahoo! di restringere la ricerca a pagine in lingua italiana rilasciate con licenza *Creative Commons* (CC), si è scelto di scaricare dal web documenti per cui fosse prevista la possibilità di riutilizzo e condivisione all'interno di un'altra opera.[3] Una volta ottenuta la lista di url, si è proceduto con l'eliminazione di pagine erroneamente riconducibili alle licenze CC, individuate sulla base di una *black list* di siti realizzata manualmente durante la realizzazione di precedenti e sperimentali versioni del corpus, nonché allo scaricamento e alla ripulitura dei documenti con il sistema *KrdWrd* (Steger, Stemle 2009).

Una seconda componente del corpus PAISÀ comprende documenti provenienti dalle versioni italiane di alcuni dei progetti web di Wikimedia Foundation – Wikipedia, Wikinews, Wikisource, Wikibooks, Wikiversity  in questo caso, anziché scaricare i singoli documenti, si sono utilizzati i *dump* ufficiali rilasciati da Wikimedia Foundation[4] assicurandosi di includere nel corpus soltanto quei progetti che garantissero porzioni consistenti di testo continuo.

Una volta ottenuti tutti i testi, si è effettuata una scrematura sull'intera collezione allo scopo di eliminare i documenti vuoti o con quantità di testo inferiori alle 150 parole. In Tab.1 è indicato il 'peso relativo' delle due componenti del corpus PAISÀ per quanto riguarda il numero complessivo di documenti e *token*.

**QUI INSERIRE TAB. 1**

| | N° documenti | N° token |
|---|---|---|
| **Non-Wikis** | 200.521 | 506.638.863 |
| **Wikis** | 268.567 | 193.393.072 |
| **TOTALE** | 469.088 | 700.031.935 |

**Tabella 1:** Composizione del corpus PAISÀ (settembre 2010)

### *1.3 Conseguenze del metodo di scaricamento e ripulitura sulla tipologia e la qualità dei testi*

Abbiamo più volte enfatizzato, nei paragrafi precedenti, come la scelta di includere nel corpus PAISÀ testi distribuiti con licenze *Creative Commons* sia dettata dalla volontà di produrre una risorsa condivisibile sia con la comunità scientifica che al di fuori di essa (per finalità

[3] Escludendo quindi le licenze CC contenenti l'opzione *Non opere derivate*, le formule considerate, che prevedono che il corpus sia ridistribuito alle stesse condizioni, comprendono: CC-Attribuzione, CC-Attribuzione-Condividi allo stesso modo, CC-Attribuzione-Non commerciale, CC-Attribuzione-Non commerciale-Condividi allo stesso modo. Per approfondimenti si veda http://creativecommons.it/Licenze.

[4] http://dumps.wikimedia.org/backup-index.html.

educative, ad esempio); parimenti, si è accennato al fatto che la procedura semi-automatica di compilazione del corpus permette la creazione di risorse di notevoli dimensioni in tempi brevi e a fronte di un intervento umano estremamente limitato. Al fine di descrivere correttamente le caratteristiche del corpus e di giustificare alcune sue mancanze, è tuttavia necessario rendere conto degli svantaggi che tali decisioni hanno comportato, in termini di varietà e completezza testuale.

In primo luogo, nonostante il web si caratterizzi per una ricchissima varietà testuale (tanto che, secondo Crystal (2006: 229), «it offer[s] a home to all linguistic styles within a language»), si è riscontrato come l'utilizzo delle licenze CC sia in realtà limitato a un numero relativamente ristretto di tipologie di siti e di testi; nello specifico, ad esempio, il 63% circa dei documenti scaricati tramite motore di ricerca proviene da blog personali o istituzionali che tipicamente trattano argomenti connessi al sociale e alla politica, o ancora all'informatica e al mondo dei motori. Inoltre, sono inevitabilmente escluse dal corpus molte forme di comunicazione elettronica nate proprio con la rete – siano esse esempi di comunicazione personale (ovvero "uno-a-uno") come e-mail e messaggistica istantanea, o con multi-destinatari come chat, forum, social network ecc. In altre parole, la presenza della licenza CC restringe considerevolmente la varietà dei documenti in termini di genere e argomento, tanto che risulta impossibile affermare che PAISÀ sia rappresentativo della comunicazione elettronica nel suo complesso.

In secondo luogo, oltre al sistema di scaricamento, anche quello di ripulitura ha avuto conseguenze sulla qualità dei testi: il sistema *KrdWrd* consente infatti di ottenere testi continui molto puliti perché, partendo dalla pagina web completa, separa le aree del documento in cui si concentra l'informazione linguistica da quelle a basso contenuto testuale (menu, immagini, in generale *boilerplate*) producendo un file in formato solo testo che verrà analizzato e eventualmente – fatti salvi i limiti specificati in §1.2 – incluso nel corpus. Ciò comporta che l'unità di base inclusa nel corpus è inferiore alla pagina web completa. Metodologicamente parlando, accettare di adeguarsi ad esigenze/conseguenze essenzialmente computazionali non ha rappresentato, a fronte dei vantaggi ottenuti, uno svantaggio, specialmente se si considera che l'intrinseca granularità dei testi web ha portato autori diversi a individuare come unità di studio alternativamente segmenti di pagina web (Waltinger et al., 2009), pagine web intere (Santini, 2007), siti (Lindemann, Litting, 2011) o addirittura network di siti tra loro collegati (Björneborn, 2011). D'altra parte, come vedremo, la rimozione di elementi caratterizzanti delle pagine web come menu e immagini si traduce nella

disponibilità di un minor numero di *feature* per la classificazione automatica dei generi web (cfr. 4).

## 2. Annotazione del corpus

Nei paragrafi che seguono illustreremo il duplice livello di annotazione di cui sono corredati i documenti inclusi in PAISÀ: da un lato un'annotazione di tipo linguistico/morfo-sintattico, dall'altro un insieme di metadati su argomento, intenzione comunicativa e genere testuale. Poiché la prima si è rivelata meno problematica di quanto ci si potesse attendere, fornendo anzi indicazioni interessanti sulla natura dei testi contenuti nel corpus (§2.1), in §2.2 e nel resto dell'articolo ci concentreremo prevalentemente sulla metodologia e le problematiche riguardanti la classificazione dei testi in funzione dei tre parametri metatestuali.

### *2.1 L'annotazione linguistica: alcune osservazioni*

L'annotazione di primo livello di cui è arricchito il corpus PAISÀ, quella linguistica, comprende attualmente divisione in frasi, tokenizzazione, lemmatizzazione e Part-Of-Speech tagging, e verrà ampliata entro la fine del progetto con informazioni di natura sintattica ricavate da un *parsing* a dipendenze.

Come accennato al paragrafo §2, ci si attendeva che – data la presunta particolarità della lingua del web, che alcuni autori accostano ad esempio alla lingua dell'oralità (es. Crystal, 2006) – gli strumenti a disposizione, addestrati su corpora più "tradizionali" (specie di testi giornalistici), avrebbero registrato performance sensibilmente peggiori. In realtà il POS-tagger utilizzato nell'ambito del progetto (Dell'Orletta, 2009), che rappresenta lo stato dell'arte per la lingua italiana con un'accuratezza che raggiunge il 97.10%, ha dimostrato alla sua prima applicazione al corpus PAISÀ un'accuratezza del 95.10%. Dopo due cicli di correzione manuale di 20.000 token ciascuno[5] (che hanno messo in evidenza errori di segmentazione del testo, abbreviazioni/sigle precedentemente non considerate, emoticon e altre sequenze 'anomale' di caratteri) e due conseguenti fasi di riaddestramento, si è ottenuta un'accuratezza del 96.03%, difficilmente incrementabile se si considera che essa è prevalentemente dovuta alla presenza di estratti di lingua non-standard come:

**QUI INSERIRE FIG. 1**

[5] Il primo ciclo è stato condotto su un campione casuale di testi, il secondo su un insieme di testi selezionati in base alle difficoltà riscontrate dal POS-tagger nelle fasi precedenti dell'annotazione.

- Approposito del processo "Cogne bis"; ma siamo seri ; Boja Fauss…!!!
- …quann è arrivat l'ora, s'add calà o sipario …

**Figura 1:** Due esempi di italiano non-standard all'interno del corpus PAISÀ

Il confronto tra le due percentuali suggerisce che i testi contenuti in PAISÀ si caratterizzano per una netta prevalenza dell'uso tipicamente scritto della lingua italiana standard, non distante da quella utilizzata al di fuori del web; questo è probabilmente dovuto all'assenza di tipologie testuali per le quali è lecito presupporre una vicinanza maggiore alla lingua parlata, come e-mail, chat e social network (cfr. §1.3).

### *2.2 Perché una classificazione dei testi per metadati*

L'inserimento di metadati relativi all'argomento, al genere testuale e all'intenzione comunicativa dei testi che compongono PAISÀ rappresenta un'operazione decisamente più complessa dell'annotazione di tipo linguistico, soprattutto in relazione ai documenti scaricati dalla rete senza alcun controllo o pre-selezione (cfr. §1.2): la scelta di scaricare documenti in base alla ricerca di parole della lingua generale mirava infatti a ridurre il rischio di selezionare testi appartenenti a determinati domini o generi, al fine di non influenzare troppo la composizione del corpus. Ciò nonostante, l'annotazione per metadati è considerato un aspetto essenziale del corpus PAISÀ, per due ragioni principali.

Innanzitutto, dal punto di vista dello sviluppo e del miglioramento della risorsa, acquisire una maggiore consapevolezza sulla composizione del corpus – ovvero capire quanto esso è (s)bilanciato rispetto a determinati argomenti e generi testuali – permette di avere un quadro di riferimento più chiaro per interpretare i dati da esso estratti e, eventualmente, intervenire per tentare di bilanciarlo attraverso lo scaricamento mirato di ulteriori testi.

Ma è dal punto di vista dei destinatari finali del corpus, insegnanti e apprendenti di lingua italiana in primis, che tale operazione acquisisce un'importanza ancora maggiore: la possibilità di visualizzare ogni testo congiuntamente alla triplice etichetta usata per descriverlo e, ancor più, di filtrare le ricerche eseguite sul corpus in base ad ognuno dei tre parametri (creando, eventualmente, sottocorpora con determinate caratteristiche), consentirà loro di usufruire di modalità di ricerca e visualizzazione avanzate e di avere una maggiore consapevolezza sulla provenienza dei fenomeni osservati.

Date le dimensioni del corpus, che determinano l'impossibilità di inserire manualmente un'annotazione così dettagliata, la sfida di PAISÀ consiste nel determinare quale metodo utilizzare per ottenere una classificazione automatica affidabile dei testi (§4), soprattutto in relazione ai parametri 'intenzione comunicativa' e 'genere' per i quali non si può fare affidamento su aspetti evidenti quali il lessico.

## 3. La triplice tassonomia di PAISÀ

In questa sezione presentiamo la tassonomia ideata per l'annotazione dei metadati nel corpus PAISÀ – triplice perché articolata nei tre parametri 'argomento', 'intenzione comunicativa' e 'genere' – nonché l'iter seguito per metterla a punto.

Per ciascuno dei parametri si è cercato di definire una *palette* di categorie (o classi) a disposizione dell'annotatore – persona fisica e/o classificatore automatico – per la classificazione dei singoli documenti; si è fatto ciò sulla base delle riflessioni emerse dall'analisi della letteratura e dall'osservazione diretta – ovviamente a campione – dei testi inseriti nel corpus, al fine di rispettarne le peculiarità: se infatti, da un lato, sarebbe sempre auspicabile, ai fini della condivisione delle risorse, utilizzare classificazioni quanto più possibile standardizzate e/o condivise, le caratteristiche specifiche di PAISÀ fanno sì che non vi siano rappresentati tutti i possibili argomenti e/o generi identificati in altri corpora (cfr. §1.3). Nel definire la tassonomia si è quindi stabilito di limitare il numero delle categorie per evitare di produrre *palette* troppo ampie e, al contempo, la presenza di classi non utilizzate.

A tale scopo, dopo aver abbozzato una prima tassonomia sulla base della letteratura (i principali riferimenti bibliografici presi in considerazione per ciascuno dei tre parametri saranno commentati nelle sezioni corrispondenti), abbiamo condotto quattro cicli di annotazione manuale su campioni casuali di testi, confrontando ogni volta le proposte di classificazione di tre annotatori. Al termine di ogni ciclo si sono apportate modifiche alla tassonomia sulla base delle incoerenze e delle problematiche emerse (eliminazione progressiva di classi non utilizzate, inserimento di nuove categorie, ridefinizione delle classi per rendere più netti i confini tra le stesse), aumentando progressivamente il tasso di *agreement* tra i tre annotatori.

### *3.1 L'argomento*

Dei tre parametri presi in esame, l'argomento è senz'altro quello a cui più spesso è stata accordata considerazione non solo nell'ambito dell'*Information Retrieval*, ma anche nello specifico della linguistica dei corpora. D'altra parte, nonostante il generale interesse, allo stato attuale non esiste una classificazione per *topic* valida in termini generali: ogni tentativo risponde a criteri diversi dettati dalla natura dei testi raccolti, dall'uso cui servono e, in ultima analisi, da fattori soggettivi (Sinclair et al., 1996). Non a caso, già negli anni '90 si contavano venti sistemi di classificazione diversi per la compilazione di corpora nella sola Europa (Calzolari et al., 1995) a cui vanno aggiunte le numerose tassonomie più recenti (tra cui, ad es., Lee, 2001). Oltre a essere l'una diversa dall'altra, le *palette* esistenti sono difficilmente comparabili non solo perché «the boundaries between the topics are ultimately blurred» (Sinclair et al., 1996), ma anche in riferimento al grado di generalità delle categorie considerate: laddove ad esempio il British National Corpus considera la categoria 'Leisure', altri entrano più nel dettaglio distinguendo tra 'people' e 'daily activities' (Biber, 1994).

### 3.1.1 PAISÀ: la classificazione per argomento

Come quelle citate, la tassonomia degli argomenti pensata per PAISÀ è stata creata sulla base di criteri esterni, non linguistici; non ha pretese di generalità, intende piuttosto aderire alle peculiarità dei testi contenuti nel corpus (§ 1.3) e rispondere alle esigenze particolari dei suoi possibili utenti (§ 1.1 e §2.2).

Perseguendo tali scopi, si è deciso innanzitutto di includere il minor numero possibile di categorie evitando *palette* ampie e estremamente dettagliate. Se infatti quella dell'argomento è sempre una categoria aperta (Biber, 1994: 44) tanto che «any classification of topics is not complete» (Sharoff, 2004), nel caso dei testi di PAISÀ tale tendenza è accentuata dalla particolarità e imprevedibilità della lingua del web che appare creativa in relazione a tutti gli elementi della testualità, argomento compreso. Nel caso delle pagine web si accentua ad esempio la tendenza alla sovrapposizione degli argomenti, caratteristica che, come si evince dall'esempio che segue (Fig. 2), rende talvolta complesso il lavoro di annotazione manuale.

**QUI INSERIRE FIG. 2**

L'Urlo di Munch che brucia, l'uomo che grida e grida mentre il fuoco annerisce e buca la sua tela, la arriccia e la sbriciola in un grumo di poltiglia bollente e appiccicosa, è una immagine che mi insegue da quando ieri si è saputo. L'Urlo l'ho visto anni fa ad Olso, una delle quattro copie, in una stanza informale e spopolata, com'è poi è tutta la Norvegia, informale e spopolata, questo quadro al centro di una parete spoglia, senza pretese di drammaticità. […] Io non ho quella che tutti chiamano memoria fotografica, quel discorso che si sente sempre fare "quando all'esame rispondo ad una domanda vedo la pagina del libro", io non vedo proprio niente, figuriamoci poi una delle novecento pagine della storia critica della letteratura inglese di David Daiches, che sono tutte uguali in quel font britannico sottile e stampate su carta povera che già intravedi la pagina suggessiva scritta fitta fitta e senza immagini. […] La fiducia entusiastica e smisurata che Anthony Bourdain mi ispira, cuoco e scrittore newyorkese, non si spiega se non con i miei trascorsi nella sottocultura delle cucine professionali, nel mio caso una umile cucina di pub malfamato, ma ricordate che una cucina di pub può essere sorgente di delizie come di morte, al pari di qualsiasi altra cucina. Essendo io una donnicciuola un po' schifiltosa, ed essendo Bourdain un scafato cuoco di scuola francese, lo snodo più interessante della sua visione del mondo, dal mio punto di vista, è il rapporto con le frattaglie. **[ccmod.443]**

**Figura 2:** Esempio di sovrapposizione di argomenti all'interno di uno stesso documento

Per ovviare a casi come quello proposto, in sede di annotazione manuale si è stabilito di attenersi strettamente a criteri quantitativi, e cioè di annotare in base all'argomento prevalente del testo; in tal modo dovrebbe essere possibile anticipare le decisioni dell'annotatore automatico.

La classificazione per argomenti nella sua forma attuale è la seguente:

**QUI INSERIRE TAB. 2**

| ARGOMENTI | DESCRIZIONE |
|---|---|
| **Business** | Economia, commercio, finanza, lavoro, ecc. |
| **Arti** | Arti visive, letteratura, architettura, cinema, musica, ecc. |
| **Hi-tech** | Informatica, computer, web, telefonia, elettronica, ecc. |
| **MSP** | Medicina - Salute - Psicologia, ecc. |
| **Leisure** | Tv, moda, astrologia, sport, videogiochi, viaggi, cucina, ecc. |
| **FLERS** | Filosofia - Lingua - Educazione/formazione - Religione - Sociologia, ecc. |
| **S.naturali** | Scienze naturali: meteorologia, astronomia, biologia, fisica, chimica, matematica, geografia, ecc. |
| **PSS** | Politica - Società - Storia: istituzioni, amministrazione (trasporti, esercito, ecc.), legge, geopolitica, ecologia, etica, ecc. |

**Tabella 2:** La tassonomia degli argomenti del corpus PAISÀ

Le otto categorie fissate sono il risultato di una serie di aggiustamenti a partire dalla proposta di Sharoff (2004) che distingue tra 'natsci' (scienze naturali); 'appsci' (scienze applicate, 'sosci' (scienze sociali), 'politics', 'commerce', 'life', 'arts' e 'leisure'. Tutte le modifiche apportate mirano ad adeguare la tassonomia alle peculiarità dei testi di PAISÀ. Dopo ad esempio un primo tentativo di mantenere la distinzione tra 'natsci', 'socsci' e 'appsci', si è

constatato che la natura intrinsecamente accademica di tale categorizzazione non rispecchiava quella più libera e leggera di PAISÀ. Nelle *palette* di riferimento scaturite dai successivi cicli di annotazione, si è preferito quindi introdurre 'hi-tech' e raggruppamenti del tutto originali come 'MSP', 'FLERS' e 'PSS'. Queste novità hanno reso l'annotazione senz'altro più agevole. Quanto all'accuratezza, l'unico caso in cui la facilità di annotazione non si è accompagnata a un altrettanto alto grado di precisione è stato quello di 'Politica, Società, Storia': l'alta incidenza di documenti di argomento 'PSS' e la conseguente perdita di informatività dell'etichetta, potrebbe dover richiedere in futuro un'ulteriore suddivisione interna.

È interessante notare come il rating di accordo degli annotatori sia rimasto piuttosto alto e pressoché invariato nel corso dei quattro cicli di annotazione (86.61% al primo e 86.66% al quarto). Nel complesso questo fa pensare che la nuova proposta di classificazione, almeno per l'annotatore specialista, sia intuitiva non meno di quelle più "tradizionali".

### 3.1.2 Alcune conferme: un esperimento di *clustering* sui testi di PAISÀ

Conferme sulla validità della tassonomia nel descrivere appropriatamente gli argomenti di PAISÀ arrivano da un esperimento di *clustering* statistico condotto da Serge Sharoff (Leeds University, Centre for Translation Studies), visiting scholar all'Università di Trento nell'autunno 2010. Il modello adottato (un *topic model* conosciuto come *Latent Dirichlet Allocation*) ha raggruppato i testi del corpus in 20 *cluster*/argomenti, caratterizzati dalla presenza di parole chiave emerse come altamente tipiche dei diversi argomenti. Si è quindi cercato di mappare la tassonomia degli argomenti sviluppata per PAISÀ (Tab. 2) con i 20 *cluster* emersi dall'esperimento, ottenendo i risultati presentati in Tab. 3 (per ogni *cluster* sono riportate alcune delle *keyword* risultate significative per il gruppo):

**QUI INSERIRE TAB. 3**

| ARGOMENTI | CLUSTER ASSOCIABILI |
|---|---|
| **Business** | Cluster 11: *euro milione mercato mese prezzo costo circa paese pagare miliardo* |
| **Arti** | Cluster 4: *film libro cinema personaggio protagonista pubblicare opera regista*<br>Cluster 6: *musica canzone cantante mare cantare musicale brano concerto* |
| **Hi-tech** | Cluster 19: *sito informazione internet sistema utente rete software contenuto* |
| **MSP** | Cluster 13: *medico legge ricerca umano diritto malattia salute scientifico studio* |
| **Leisure** | Cluster 1 [sport]: *titolo torino cronaca calcio contenuto squadra partita giocatore*<br>Cluster 2 [TV]: *registrare programma fratello mare isola show famosi pupa*<br>Cluster 10 [TV]: *ascolto canale rete share agosto telespettatore milione mentana*<br>Cluster 14 [TV]: *serie stagione episodio puntare settembre puntata pubblicare*<br>Cluster 5 [motori]: *motore moto gara pilota auto giro vettura versione modello*<br>Cluster 18 [cucina]: *acqua animale prodotto piccolo colore usare bianco vino* |
| **FLERS** | Cluster 3 [religione]: *signore figlio autem uomo chiesa terra super dominus gesù* |
| **S.naturali** | - |
| **PSS** | Cluster 0: *lavoro comune sociale progetto politico lavoratore scuola centro*<br>Cluster 8 [politica interna]: *politico legge berlusconi presidente governo partito*<br>Cluster 15 [politica estera]: *guerra politico stato paese governo americano*<br>Cluster 17: *polizia forza carcere piazza manifestazione carabiniere gruppo gente* |

**Tabella 3:** Corrispondenze tra tassonomia PAISÀ e risultati *clustering* statistico

Dal confronto emerge che sette delle otto categorie stabilite sulla base di criteri esterni appaiono confermate anche una volta adottati criteri essenzialmente lessicali. L'accordo è quanto mai evidente in classi come 'Business' o 'Hi-tech'. In tre casi ('Arti', 'Leisure' e 'PPS') a una classe corrispondono più *cluster*: ad 'Arti', ad esempio, ne corrisponde uno incentrato sulla musica e l'altro su cinema e letteratura; a 'Leisure' invece fanno eco sei *cluster*, di cui tre dedicati alla televisione, e uno rispettivamente a sport, motori e cucina. Solo una classe della tassonomia, quella dedicata alle 'Scienze naturali', non trova riscontri nei risultati del *clustering*: tale dato sembra confermare il carattere non accademico-disciplinare dei testi di PAISÀ e sembra suggerire l'opportunità di un'ulteriore revisione della tassonomia.

Un secondo punto critico appare la mancata corrispondenza di alcuni *cluster* (in Tab. 4) con le classi previste dalla tassonomia PAISÀ:

**QUI INSERIRE TAB. 4**

| ARGOMENTO | CLUSTER ASSOCIABILI |
|---|---|
| ?? | Cluster 7: *politico paolo luca marco antonio carlo giovanni roberto blog alberto andrea poesia giorgio francesco verso franco scienza alessandro* |
| ?? | Cluster 9: *città euro centro roma fino metro zona viaggio milano strada mare piazza parco notte partire lungo luogo circa volo* |
| ?? | Cluster 12: *ragazzo figlio uomo bambino sentire amico lasciare famiglia ragazza mano vivere morte piccolo padre morto momento tornare madre* |
| ?? | Cluster 16: *credere scrivere piacere sentire problema gente guardare bello niente davvero leggere buono parere magari vostro lasciare ciò visto sperare* |

**Tabella 4:** *Cluster* senza corrispondenze nella tassonomia PAISÀ

Oltre a non avere rispondenza nella tassonomia, le parole chiave dei *cluster* riportati in Tabella 4 sono difficilmente riconducibili ad argomenti specifici. E' interessante però ipotizzare che alcuni raggruppamenti che risultano poco significativi per la classificazione degli argomenti possano fornire altre indicazioni: il *cluster* 16, ad esempio, presenta un numero interessante di *keyword* che rimandano alla sfera della percezione (*credere, piacere, sentire, sperare, bello, buono*, ecc.) e, per tale ragione, potrebbe essere indicativo di un genere come il diario o il commento personale. D'altra parte, sarà interessante verificare se tali discrepanze scompaiono abbassando il numero di *cluster* richiesti in uscita come parametro del modello statistico utilizzato, per appurare se tale accorgimento è sufficiente ad aumentare l'aderenza alla tassonomia.

Un vantaggio dello specifico *topic model* utilizzato consiste nella possibilità di evidenziare in quale proporzione ogni documento del corpus appartiene a diverse categorie (nel caso degli argomenti ad esempio: 44% 'Arti', 25% 'Leisure', 10% 'Business'): tale informazione consente di attribuire con maggiore certezza ogni testo alla classe predominante, ma anche di rappresentare la natura molteplice di molti documenti web, qualora si preveda la possibilità di utilizzare un'etichetta multipla all'interno dei metadati.

### *3.2 L'intenzione comunicativa*

Senza addentrarsi nell'analisi delle numerose proposte tassonomiche avanzate per le funzioni comunicative di testi "tradizionali" e web (Biber, 1994; Santini, 2005; Sharoff, 2004, 2007), è opportuno sottolineare come dedicarsi allo studio delle intenzioni comunicative di testi scaricati dal web possa avere diversi vantaggi: innanzitutto, sembra che le intenzioni o funzioni comunicative possano rivelarsi utili per lo studio dei generi web se assunte come parametro in base al quale confrontare *web genres* e generi testuali tradizionali (Sharoff, 2007); inoltre sono facilmente riconosciute dagli utenti del web (Santini, 2005), il che apre scenari interessanti per nuove forme di *query* sul web o su corpora linguistici come nel caso di PAISÀ.

#### 3.2.1 PAISÀ: la classificazione per intenzione comunicativa

Anche se stabilire quale sia la funzione comunicativa di un testo non è semplice, data l'intrinseca valenza multifunzionale di gran parte dei documenti, si è deciso di inserire tra i metadati in PAISÀ un'informazione di tipo funzionale (anche se solo in termini di funzione

prevalente) perché essa diventa particolarmente rilevante se messa in correlazione con gli altri due parametri relativi ad argomento e genere. Come risulta dall'esempio che segue (Fig. 3), talvolta solo l'annotazione dello scopo che un testo persegue aiuta a disambiguare la sua natura:[6]

**QUI INSERIRE FIG. 3**

Il WWF Italia valuta, che visti i rilevi riguardanti in particolare le aree di Capo Peloro - Laghi di Ganzirri e la Costa Viola (aree dove dovrebbero sorgere i piloni del ponte) e i Monti Peloritani (su cui impattano le strutture aeree del ponte), se la procedura si concludesse con il deferimento alla Corte di Giustizia europea, l'Italia sarebbe obbligata a mettere in un cassetto l'attuale progetto, che è stato posto a base di gara, e ri-elaborare un proposta radicalmente diversa da quella attuale. **[ccmod.1112]**
**Annotazione: [pss - informare - gio.cronaca]***

Finanziamenti pubblici, licenze edilizie su aree destinate dal Piano Regolatore ad altri scopi e agevolazioni per la costruzione di case di cui solo una piccola parte sara' in affitto "calmierato" ed il resto messo in vendita dagli stessi costruttori. Il piano "20.000 abitazioni in affitto" a Firenze e', in sintesi, una perdita di risorse collettive a vantaggio dei profitti privati. **[ccmod.1859]**
**Annotazione: [pss - argomentare - gio.cronaca]***

**Figura 3:** Esempio di due testi simili con intenzioni comunicative differenti

Per la classificazione delle intenzioni comunicative, ci siamo ispirati di nuovo al lavoro di Sharoff (2004) e stavolta in maniera più fedele: alla sua proposta sono stati apportati infatti solo minimi aggiustamenti terminologici nei casi di 'argomentare' e 'intrattenere':

**QUI INSERIRE TAB. 5**

| INTENZIONE | DESCRIZIONE |
|---|---|
| **Raccomandare** | Raccomandare, consigliare, convincere, persuadere, ecc. |
| **Informare** | Informare, descrivere, presentare e raccontare, esprimere se stessi/raccontarsi, ecc. |
| **Argomentare** | Argomentare, discutere, commentare e valutare |
| **Intrattenere** | Intrattenere, divertire |
| **Istruire** | Dare istruzioni, insegnare |

**Tabella 5:** La tassonomia delle intenzioni comunicative del corpus PAISÀ

### 3.2.2 L'ambiguità tra testo informativo e testo argomentativo

Il peggior dato relativo all'accordo tra annotatori è proprio quello inerente le intenzioni comunicative: laddove, durante il quarto ciclo di annotazione, si è arrivati a etichettare in

[6] Le etichette poste in calce ai due estratti saranno del tutto trasparenti solo una volta affrontato il terzo e ultimo parametro dedicato al genere (§ 3.3). Per il momento anticipiamo che la presenza dell'asterisco indica la provenienza del testo da un blog.

maniera omogenea circa l'86.7% dei testi in relazione agli argomenti e il 78.7% in relazione ai generi, sulle intenzioni comunicative non si è superato il 73.3%. Nella quasi totalità dei casi il disaccordo è stato dovuto all'intrinseca ambiguità tra testo informativo e testo argomentativo. E' spesso infatti difficile attribuire un peso prevalente a una delle due, come mostra l'esempio in Fig. 4, in cui a un incipit di chiaro carattere cronistico (v. indicazione oraria) seguono caratteristiche altrettanto chiare di segno questa volta argomentativo ("*cosa piuttosto singolare e assurda*", "*indicibile*"):

**QUI INSERIRE FIG. 4**

Ad andare sul posto è il direttore di Italymedia.it Antonello De Pierro, nonché voce storica di Radio Roma e presidente del movimento nazionale "L'Italia dei diritti". Il noto giornalista giunge nei locali della struttura esattamente alle ore 18,42, e quindi 18 minuti prima dell'orario canonico di chiusura al pubblico, con l'intenzione di effettuare un pagamento tramite bollettino postale. All'ingresso viene fermato da un impiegato, che con grande naturalezza lo avvisa del fatto che l'ufficio è già chiuso, cosa piuttosto singolare e assurda vista l'ora. Ignorando l'azzardato avvertimento si reca comunque presso la macchina erogatrice dei numeri progressivi che regolano l'affluenza agli sportelli, e qui lo attende un'indicibile sorpresa: dalla fessura esce un biglietto con la scritta "IL SERVIZIO NON E' ATTIVO". **[ccmod.4168]**

**Figura 4:** Esempio di testo polifunzionale (informativo e argomentativo)

*3.3 Il genere*

L'emergere del web e, con esso, dei generi web, ha rappresentato l'occasione per ripensare al concetto di genere testuale in generale e, almeno in una certa misura, ha comportato il riemergere di un tradizionale problema di definizione. Innumerevoli adesso come in passato sono infatti i tentativi di inquadrare il genere testuale: per alcuni costituisce un'azione sociale poiché, come categoria convenzionale di discorso, connette le intenzioni private alle esigenze pubbliche (Miller, 1984); per altri è «a recurring type or category of text, as defined by structural, thematic and/or functional criteria» (Duff, 2000: XIII) e, per questa ragione, è predittivo dei contenuti e delle finalità di un documento (Bateman, 2008).

Anche nello specifico dei generi web, il problema si ripresenta immutato, seppur riproposto in termini diversi. Se infatti «cybergenres may be extant (i.e., based on existing genres) or they may be novel (i.e., not like any existing genre in any other medium» (Shepherd, Watters, 1998), resta da chiarire cosa significhi "genere", vecchio o nuovo che sia. Parimenti affermare che «genres are cultural products, linked to a culture, a society, a community' e che «the Web is a new, large and heterogeneous community» (Santini, 2005) sposta semplicemente il problema dalle comunità "tradizionali" a quelle nuove sorte con esso.

Nonostante i molti sforzi definitori, quando di fatto si tenta di isolare gli elementi che accomunano i testi appartenenti a un medesimo genere, le indicazioni si fanno nebulose. Purtroppo la mancanza di chiarezza e la rinuncia a esplicitare le caratteristiche linguistiche e non linguistiche dei generi sul piano teorico costituisce il maggior freno alla ricerca empirica in ambito computazionale perché «without a theoretical definition and charaterisation underpinning the concept of genre, it is not clear how to select the members belonging to a genre class and in which way the genre labels 'represent' a selected genre class» (Santini et al., 2011: 8).

In mancanza di una definizione ultima, la classificazione dei testi di PAISÀ è basata sull'idea – come abbiamo visto largamente condivisa – che il genere sia una categoria convenzionale di discorso, un fenomeno sociale tipico di una comunità di parlanti e, per questo, parte integrante della competenza testuale dei singoli. Lo studio dedicato a questo parametro non intende quindi giungere a una nuova definizione di genere, quanto muoversi su un piano più prettamente empirico contribuendo a una caratterizzazione o parametrizzazione delle caratteristiche linguistiche e non linguistiche dei singoli generi web in italiano.

Per la definizione della *palette* dei generi di PAISÀ (§3.3.1), a nostra conoscenza la prima dettagliata per corpora di italiano di dimensioni paragonabili, si è fatto riferimento almeno inizialmente alla ricca bibliografia dedicata ai corpora inglesi, web e non-web, in particolare: Santini (2005, 2011), Lee (2001), Rehm et al. (2008).

### 3.3.1 PAISÀ: la classificazione per genere

Le principali novità della tassonomia per genere di PAISÀ rispetto alle *palette* esistenti sono la struttura gerarchica su due livelli e il trattamento riservato al blog. In questa sede, dopo la visualizzazione completa della *palette* (Tab. 6), ci soffermeremo soprattutto sulla prima delle due, mentre il ruolo del blog verrà descritto e commentato nel prossimo paragrafo (§ 3.3.2).

**QUI INSERIRE TAB. 6**

| BLOG ? | GENERE 1° livello | GENERI 2° livello |
|---|---|---|
| | Fiction | Prosa - Poesia - Sceneggiatura |
| √ | Guida | Tutorial - FAQ - Turismo - Ricetta |
| √ | Giornalismo | Cronaca - Editoriale - Intervista - Reportage - Recensione |
| | Accademia | Prosa - Lezione - Abstract |
| | Doc. ufficiale | Legge - Relazione - Contratto |
| | Scheda | Prodotto – Curriculum Vitae - About page |
| | Annuncio | |
| | Commento | |
| | Lemma | |

**Tabella 6:** La tassonomia dei generi del corpus PAISÀ

La strutturazione gerarchica della classificazione ha come primo e più immediato vantaggio il fatto di rendere l'idea di come i generi si definiscano a seconda dei casi a livelli diversi di generalità (ad esempio l'annuncio al primo livello, il curriculum al secondo). Questa soluzione consente una categorizzazione piuttosto dettagliata senza però comportare una lista eccessivamente lunga di generi come se ne trovano in letteratura, dove talvolta si arriva a individuare classi quali ad esempio 'liste di progetti', 'liste di prodotti' e 'citazioni' (Rehm et al., 2008). Inoltre, essa può favorire l'annotazione automatica perché, nel caso in cui gli strumenti non fossero in grado di individuare il massimo livello di specificità del genere, è pensabile che possano quantomeno segnalarne l'appartenenza a una data macrocategoria. Infine, la classificazione così concepita sembra avere un buon potenziale di comparabilità con altre tassonomie esistenti poiché è possibile svilupparla e/o ridurla nella singole parti senza stravolgerne per questo l'impianto di base.

In generale, per delineare la *palette* ci si è avvalsi di criteri misti. Prevale (soprattutto nei casi delle macrocategorie 'Fiction', 'Giornalismo', 'Accademia' e 'Documenti ufficiali') un sistema di raggruppamento dei generi legato agli ambiti d'uso delle relative comunità di parlanti (Berruto, 1981: 43). Più minute e fedeli alla tipologia dei testi di PAISÀ sono invece 'Guida' (che tra l'altro anticipa il prevalere dell'intenzione comunicativa 'Istruire' nei generi che comprende), 'Lemma' (che segnala i numerosi documenti tratti dai progetti web di *Wikimedia Foundation*) e 'Scheda' (un'etichetta utile a marcare il formato anche grafico di generi diversi come i prodotti commerciali o tecnici, i curricula e le pagine di presentazione di aziende, organizzazioni, ecc.).

### 3.3.2 Il blog: contenitore di generi

Nella classificazione il blog assume un ruolo particolare: non è un genere né di primo né di secondo livello ma un formato, un contenitore in cui possono essere pubblicati tutti i generi individuati. Essendo quindi concettualizzabile come un attributo opzionale da aggiungere all'annotazione del genere, è incluso come descrittore separato nella tassonomia (prima colonna della Tab. 6) e sottoforma di asterisco nella triplice etichetta assegnata a ogni documento (ad esempio: [flers - argomentare - gio.intervista]*).

Contrariamente a quanto si potrebbe essere portati a pensare, il blog non ha in sé un carattere informale, sicuramente almeno non a livello linguistico; ciò è banalmente dimostrato

dal fatto che articoli molto simili – se non addirittura lo stesso articolo – possono essere pubblicati sia su blog che su giornali online (esempio in Fig. 5):

**QUI INSERIRE FIG. 5**

Il Gruppo Angelucci ha versato 500.000 euro alla lista di Fitto in occasione delle elezioni regionali del 2005. Secondo il gruppo (Tosinvest), si tratta di un regolare finanziamento registrato a bilancio. Per la Procura di Bari si tratta invece di una tangente pagata per assicurarsi l'appalto da 198 milioni di euro con cui Angelucci ha ottenuto la gestione delle undici residenze sanitarie "assistite" dalla Regione Puglia[1]. Si tratta della stessa inchiesta per cui è indagato Francesco Storace. Il parlamento, tuttavia, ha respinto l'autorizzazione a procedere con l'arresto con 457 voti favorevoli (su 462 presenti), 1 contrario (Antonio Borghesi dell'Idv) e 4 astenuti". **[ccmod.1112]**
**Annotazione: [pss - informare - gio.cronaca]***

Il 10 di maggio, durante il suo discorso al Parlamento sullo Stato della Nazione, il Presidente Vladimir Putin ha annunciato che la Russia renderebbe il Rublo "internazionalmente convertibile", così da poterlo utilizzare nelle transazioni riguardanti petrolio e gas naturale. Al momento, il petrolio viene esclusivamente valutato in dollari. L'annuncio di Putin risuona come un annuncio di guerra. **[ccmod.3667]** - **Annotazione: [pss - informare - gio.cronaca]**

**Figura 5:** Due articoli simili pubblicati rispettivamente in un blog e in un giornale online

Sarebbe estremamente difficile capire quale dei due esempi riportati in Fig. 5 sia tratto da un blog sulla base di *feature* puramente linguistiche. Questo perché, non l'informalità della lingua, ma il layout (la struttura del post, la presenza di commenti, ecc.) contraddistingue il blog. Nel nostro caso, dato che abbiamo deciso di non avvalerci per il momento di *feature* estratte dal layout e dal codice HTML, è la presenza nell'url della parola chiave "blog" oppure di nomi di servizi di *blog hosting* (come "Splinder") a guidare l'annotazione.

Le caratteristiche peculiari del layout hanno tra l'altro una funzione sociale non trascurabile: rendono il blog immediatamente riconoscibile al lettore e, in tal modo, lo mettono in guardia del fatto che i contenuti, poiché non certificati e garantiti da enti pubblici, testate giornalistiche, ecc., potrebbero eventualmente essere non del tutto attendibili. Da questo punto di vista il blog può effettivamente dirsi informale, ma si tratta di un'informalità che non lascia traccia nella lingua, per quanto intuitivamente riconosciuta dagli utenti del web.

A questa 'informalità non linguistica', solo talvolta si accompagnano in documenti tratti da blog fenomeni linguistici legati al registro informale. È il caso del blog nel senso più classico, sinonimo di diario personale, di cui riportiamo un esempio nell'estratto in Fig. 6:

**QUI INSERIRE FIG. 6**

Oggi ho girato un po' di negozi di abbigliamento, e sembravano diventati tutti punti vendita di merchandising della Fiorentina. Sarà contenta mae, e lo sono pure io, ché il viola, specie scuro, è un colore che mi è sempre piaciuto. Al punto che, cosa che mi succede piuttosto di rado, ho comprato una camicia solamente perché mi piaceva, e non perché nel mio armadio non c'è più nulla che non stia cadendo a pezzi. Io non conosco i nomi dei colori che danno gli stilisti, ma direi che secondo la nomenclatura XII è Purple gessata di DarkViolet. **[ccmod.048] - Annotazione: [commento - leisure - argomentare]***

**Figura 6:** Esempio di blog inteso come "diario personale"

In alcune *palette* di generi web tale testo ricadrebbe nella categoria 'personal blog', spesso messa in contrapposizione a 'corporate blog - CLOG' (es. Rehm et al., 2008). Nella classificazione dei generi di PAISÀ, dato il ruolo attribuito al blog, tale distinzione non avrebbe senso e, difatti, il documento citato è etichettato come 'Commento' al pari di tanti altri brevi testi reperibili sul web su giornali online, portali turistici, ecc. Tale informazione non viene comunque persa dato che viene comunque espressa con la classificazione del documento per intenzione comunicativa.

### 3.3.3 I generi di PAISÀ: tra scritto e parlato, formale e informale

Date le peculiarità di PAISÀ, dove mancano tipi testuali nati con il web come e-mail, chat, social network, ecc., non possiamo affermare che il *Netspeak* presente nel nostro corpus sia «something fundamentally different from both writing and speech, as traditionally understood» (Crystal, 2006: 272); al contrario, vi si trovano molti dei fenomeni legati all'uso scritto della lingua (periodi lunghi e articolati, ipotassi, varietà lessicale, ricchezza di punteggiatura, rispetto delle norme ortografiche, ecc.), e solo in misura minore quelli che sembrano caratterizzare la lingua elettronica (abbreviazioni e neologismi, uso espressivo della punteggiatura, errori di battitura o ortografia, ecc.). Abbiamo già visto in precedenza come il prevalere dell'uso scritto della lingua sembri essere confermato dalle ottime performance del *POS-tagger* utilizzato per l'annotazione linguistica del corpus (§2.1).

Quanto al livello di formalità, la questione sembra essere più articolata rispetto al rapporto tra forme scritte e parlate. In rete, e quindi tendenzialmente in PAISÀ, si trovano infatti tre tipi di testo: generi che, direttamente prestati al web come articoli giornalistici, curricula ecc., presentano lo stesso uso dei registri linguistici che vige al di fuori della rete; nuovi tipi testuali nati con la rete stessa (blog, wiki, forum, ecc.) atti ad ospitare testi vari, da formali a molto informali; e infine generi tradizionali (ad esempio la ricetta o la guida

turistica) i quali, data l'estrema libertà e contaminazione del web, possono assumere registri più o meno formali, come mostra l'esempio che segue:

QUI INSERIRE FIG. 7

Le origini della città sono remote, risalgono infatti all'età neolitica. Si pensa che in seguito alla colonizzazione greca, nella località del Cozzo di Apollo, abbia avuto origine la misteriosa Kasmenai. Anche i Romani hanno lasciato testimonianze della loro invasione; particolarmente importante è il ritrovamento di un edificio termale nel centro della città. Con l'arrivo dei Bizantini si formò il casale di Comicio, chiamato in seguito Jhomiso, e finalmente Comiso, attorno a cui nasce il primo nucleo urbano. **[ccmod.2304] - Annotazione: [leisure - informare - guida.turismo]**

Se noleggiate una macchina presso l'aeroporto, come quello di Madrid, e incontrate una persona che vi chiede la cortesia di aiutarlo con il pneumatico a terra della sua macchina, NON CI CASCATE! Alcuni visitatori sono stati derubati quando si sono fermati in aiuto di conducenti di false macchine rotte. Nel centro della capitale è tipico imbattersi in saccheggiatori, per le strade e nel metro. Molti rubano mimetizzando il furto con una mappa della città e facendosi aiutare da uno o più "soci". **[ccmod.2402] Annotazione: [leisure - istruire - guida.turismo]**

**Figura 7:** Due guide turistiche di registro linguistico diverso

## 4. Conclusioni

Lo studio presentato segna un primo passo verso l'obiettivo di dotare il corpus PAISÀ di un livello di annotazione testuale che segnali argomento, intenzione comunicativa e genere dei documenti web contenuti al suo interno. Il risultato ottenuto, una tassonomia dei tre parametri suddetti rispettosa delle peculiarità dei testi del web e di PAISÀ in particolare, rappresenta infatti un prerequisito fondamentale per le fasi di lavoro successive.

Nel proseguire la ricerca, per prima cosa ci concentreremo sull'annotazione automatica di generi e intenzioni comunicative: con un approccio di *machine learning* estrarremo alcune *feature* e, sulla base di quelle, addestreremo dei classificatori. Sebbene una combinazione di sole *feature* linguistiche a detta di molti non sia la più efficace per il riconoscimento automatico del genere (Lim et al., 2005), in una prima fase ci avvarremo esclusivamente di lunghezza media delle frasi, rapporto tra *content* e *function words*, punteggiatura, combinazioni di parti del discorso, frequenze di parole, ecc. (oltre alle informazioni estraibili dall'url per il riconoscimento dei blog). In un secondo momento, amplieremo l'uso dell'url e proveremo a impiegare informazioni estratte dal layout delle pagine web (immagini, formattazione, link, ecc.) e dal codice HTML. Inoltre, è nostra intenzione estendere a tali due parametri l'esperimento di *clustering* testato sugli argomenti.

Infine, a seconda dei risultati ottenuti dalle due suddette sperimentazioni, torneremo a riconsiderare la tassonomia per validarla o modificarla ulteriormente alla luce di quanto scoperto. A questo punto, una volta apportati gli opportuni cambiamenti e ripetuta la classificazione dei testi, sarà possibile avere un quadro completo della composizione del corpus in termini di argomenti, intenzioni comunicative e generi e, nel caso, provvedere a ulteriori scaricamenti dal web che aiutino a bilanciarlo il più possibile. Tale accorgimento sarà senz'altro d'aiuto all'utenza che, grazie anche a un'interfaccia *user-friendly* con opzioni avanzate di consultazione e visualizzazione, potrà consultare PAISÀ in maniera varia e personalizzata.

**Bibliografia**


Aston G., Bernardini S., Stewart D. (eds.) (2004), *Corpora and Language Learners*, John Benjamins, Amsterdam/Philadelphia.

Baroni M., Bernardini S., Comastri F., Piccioni L., Volpi A., Aston G., Mazzoleni M. (2004), *Introducing the "la Repubblica" Corpus: A Large, Annotated, TEI(XML)-compliant Corpus of Newspaper Italian*, in "Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004", ELDA, Lisbona, pp. 1771-1774.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009), *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*, in "Language Resources and Evaluation" 43(3), pp. 209-226.

Bateman J.A. (2008), *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*, Palgrave Macmillan, London.

Berruto G. (1981), *Tipologia dei testi e analisi degli eventi comunicativi tra sociolinguistica e "textthéorie"*, in D. Goldin (a cura di), *Teoria e analisi del testo*, Cleup, Padova, pp. 29-46.

Biber D. (1994), *An Analytical Framework for Register Studies*, in D. Biber, E. Finegan (eds.), *Sociolinguistic Perspectives on Register*, Oxford University Press, Oxford, pp. 31-56.

Björneborn L. (2011), *Genre Connectivity and Genre Drift in a Web of Genres*, in A. Mehler, S. Sharoff, M. Santini (eds.), *Genres on the Web: Computational Models and Empirical Studies*, Springer, Dordrecht, pp. 255-273.

Calzolari N., Baker M., Kruyt, T. (eds) (1995), *Towards a Network of European Reference Corpora*, in "Linguistica Computazionale", XI, pp. 127-175.

Crystal D. (2006), *Language and the Internet*, Cambridge University Press, Cambridge.

Dell'Orletta F. (2009), *Ensemble System for Part-of-Speech Tagging*, in "Proceedings of Evalita '09: Evaluation of NLP and Speech Tools for Italian" (Reggio Emilia, Dicembre 2009). http://www.evalita.it/2009/proceedings.

De Mauro T. (1989), *Guida all'uso delle parole*, Editori Riuniti, Roma.

Duff D. (2000), *Modern Genre Theory*, Pearson, Harlow.

Lee D. Y. W. (2001), *Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle*, in "Language Learning & Technology", 5(3), pp. 37-72.

Lim C. S., Lee, K. J., Kim, G. C. (2005), *Multiple Sets of Features for Automatic Genre Classification of Web Documents*, in "Information Processing and Management", 41, pp. 1263-1276.

Lindemann C., Littig L. (2011), *Classification of Web Sites at Super-genre Level*, in A. Mehler, S. Sharoff, M. Santini (eds.), *Genres on the Web: Computational Models and Empirical Studies*, Springer, Dordrecht, pp. 211-235.

Miller C. (1984), *Genre as Social Action*, in "Quarterly Journal of Speech", 70, pp. 151-167.

Rehm G., Santini M., Mehler A., Braslavski P., Gleim R., Stubbe A., Symonenko S., Tavosanis M., Vidulin V. (2008), *Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems*, in "Proceedings of the Sixth International Language Resources and Evaluation Conference (LREC'08)", http://www.lrec-conf.org/proceedings/lrec2008/.

Santini M. (2005), *Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis*, in "Proceedings of the 8th Annual Colloquium for the UK Special Interest Group for Computational Linguistics (CLUK05)", 11 Jan 2005, Manchester, UK.

Santini M. (2007), *Characterizing Genres of Web Pages: Genre Hybridism and Individualization*, in "Proceedings of the 40th Annual Hawaii International Conference on System Sciences".

Santini M. (2011), *Cross-Testing a Genre Classification Model for the Web*, in A. Mehler, S. Sharoff, M. Santini (eds.), *Genres on the Web: Computational Models and Empirical Studies*, Springer, Dordrecht, pp. 87-128.

Santini M., Mehler A., Sharoff S. (2011), *Riding the Rough Waves of Genre on the Web. Concepts and Research Questions*, in A. Mehler, S. Sharoff, M. Santini (eds.), *Genres on the Web: Computational Models and Empirical Studies*, Springer, Dordrecht, pp. 3-33.

Sharoff S. (2004), *Towards Basic Categories for Describing Properties of Texts in a Corpus*, in "Proceedings of Fourth International Conference on Language Resources and Evaluation, LREC 2004", ELDA, Lisbona.

Sharoff S. (2007), *In the Garden and in the Jungle: Comparing Genres in the BNC and Internet*, in A. Mehler, S. Sharoff, M. Santini (eds.), *Genres on the Web: Computational Models and Empirical Studies*, Springer, Dordrecht, pp. 149-166.

Shepherd M., Watters C.R. (1998), *The Evolution of Cybergenres*, in "Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences".

Sinclair J. (1996), *Preliminary Recommendations on Text Typology*, EAGLES Document EAG-TCWG-TTYP/P, http://www.ilc.cnr.it/EAGLES96/texttyp/texttyp.html [ultimo accesso 15.04.2011].

Sinclair J. (ed.) (2004), *How to use Corpora in Language Teaching*, John Benjamins, Amsterdam/Philadelphia.

Steger J.M., Stemle E. (2009), *KrdWrd - Architecture for Unified Processing of Web Content*, in I. Alegria, I. Leturia, S. Sharoff (eds), "Proceedings of the Fifth Web as Corpus Workshop", pp. 63-70. www.sigwac.org.uk/raw-attachment/wiki/WAC5/WAC5_proceedings.pdf

Waltinger U., Mehler A., Wegner A. (2009), *A Two-level Approach to Web Genre Classification*, in “Proceedings of the 5th International Conference on Web Information Systems and Technologies”.