Perugia Corpus

Per citare il *Perugia corpus*, e per maggiori informazioni sulla sua composizione e annotazione:

Spina S. *Il Perugia Corpus: una risorsa di riferimento per l'italiano. Composizione, annotazione e valutazione.* In: (a cura di): R. Basili, A. Lenci, B. Magnini, Proceedings of the First Italian Conference on Computational Linguistics CLiC-it 2014. vol. 1, p. 354-359. Pisa: Pisa University Press, 2014.

Il *Perugia Corpus* (PEC) è un corpus di riferimento dell'italiano contemporaneo, scritto e parlato, creato **all'Università per Stranieri di Perugia** all'interno di vari progetti di ricerca, coordinati da Stefania Spina.
Il PEC è composto da oltre 26 milioni di parole, distribuite in **10 generi testuali**:

1. scritto accademico
2. amministrazione
3. film
4. letteratura
5. parlato
6. saggistica
7. temi scolastici
8. stampa
9. televisione
10. web

Ogni genere testuale è a sua volta suddiviso in **43 tipi di testi differenti** (il numero corrisponde a quello dei dieci macrogeneri elencati sopra):

1 articolo
1 dispensa
1 manuale
1 tesi di laurea
2 legge
2 regolamento
3 dialogo filmico
4 romanzo
5 canzone
5 conferenza
5 conversazione faccia a faccia
5 conversazione telefonica
5 conversazione istituzionale (processuale, medica, scolastica)
5 discorso istituzionale
5 discorso politico
5 discorso processuale
5 discorso religioso
5 intervista
5 lezione
6 saggio
7 tema scuole medie
7 tema scuole superiori
8 articolo cronaca
8 articolo cultura
8 articolo economia
8 editoriale
8 articolo estero
8 lettere
8 articolo politica
8 articolo spettacolo
8 articolo sport
9 fiction
9 pubblicità
9 spettacolo
9 sport commento
9 sport cronaca
9 talkshow
9 telegiornale
10 blog
10 chat
10 forum
10 social network
10 Wikipedia

Il PEC è dotato di una annotazione multilivello:

- **l'annotazione in Xml**, finalizzata ad una descrizione della struttura interna dei testi;
- il ***pos-tagging* + lemmatizzazione** (questo è il *tagset* utilizzato; qui c'è una lista di espressioni composte da più di una parola annotate come parole singole).

N.B.: Per le **liste di frequenza** del PEC, si consiglia di usare quelle disponibili qui: *CQPweb* produce infatti degli errori, non distinguendo a dovere caratteri accentati e non accentati.

Perugia, giugno 2015
Stefania Spina