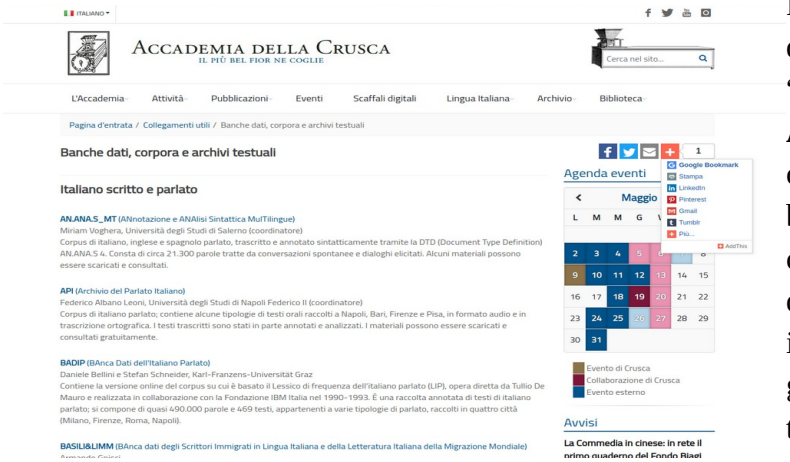
Benchmark

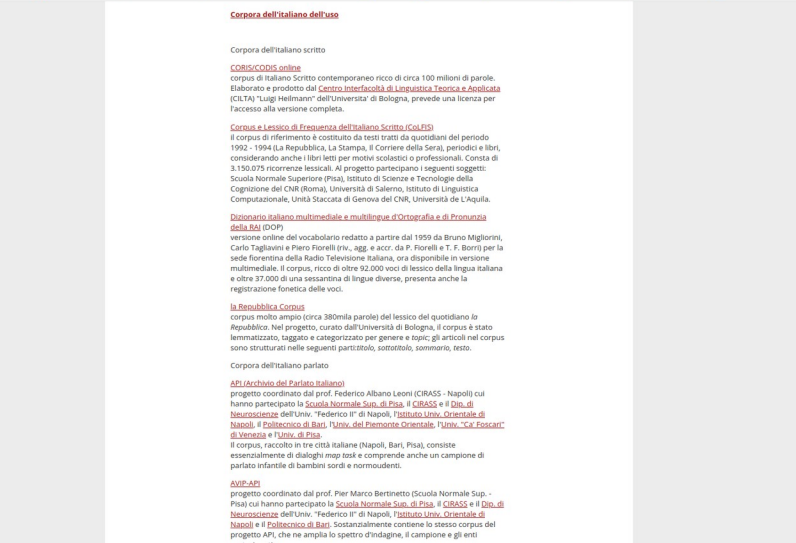
Tramite l’analisi dell’esistente possiamo notare che esistono già alcuni siti che svolgono la funzione di “aggregatori”, fornendo in un solo spazio una raccolta dei maggior corpus di italiano, fornendone una breve descrizione e il link alle piattaforme originali.



Primo fra tutti come risultato fornito da Google con stringa di ricerca “**corpus di italiano**” troviamo: **Accademia della Crusca.** Qui, tra i collegamenti utili, troviamo una bacheca di raccolta banche dati, corpora e archivi testuali. Seppur distinti fra italiano scritto e parlato, italiano antico, italiano di apprendenti, giornalistico, di internet, tecnico e trasmesso, nella bacheca troviamo assieme, come si evince dal titolo, banche dati, corpora e archivi testuali. In questo senso non può’ essere utile ai fini dell’utilizzo dei corpora in classe perché comporterebbe un passaggio ulteriore da parte dell’insegnante/studente che è quello di selezionare i corpora e distinguerli da archivi e banche dati.

Secondo risultato proposto dal motore di ricerca Google è la definizione di corpus di **Treccani** che pero non si limita alla stessa. Treccani dopo la definizione fornisce un elenco di alcuni corpus conosciuti, con link agli stessi e una breve descrizione.

Oltre a ciò questa voce è correlata da alcuni esempi di uso come possiamo osservare. Questo è certamente uno strumento utile per chi si sta approcciando per la prima volta ai corpora, ma non esaustivo per una classe di giovani studenti o studenti di italiano L2 che devono utilizzare i corpora come strumento didattico.



Corpora dell'italiano dell'uso

Corpora dell'italiano scritto

CORIS/CODIS online
corpus di Italiano Scritto contemporaneo ricco di circa 100 milioni di parole. Elaborato e prodotto dal Centro Interfacoltà di Linguistica Teorica e Applicata (CILTA) "Luigi Heilmann" dell'Universita' di Bologna, prevede una licenza per l'accesso alla versione completa.

Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS)
il corpus di riferimento è costituito da testi tratti da quotidiani del periodo 1992 - 1994 (La Repubblica, La Stampa, Il Corriere della Sera), periodici e libri, considerando anche i libri letti per motivi scolastici o professionali. Consta di 3.150.075 ricorrenze lessicali. Al progetto partecipano i seguenti soggetti: Scuola Normale Superiore (Pisa), Istituto di Scienze e Tecnologie della Cognizione del CNR (Roma), Università di Salerno, Istituto di Linguistica Computazionale, Unità Staccata di Genova del CNR, Università de L'Aquila.

Dizionario italiano multimediale e multilingue d'Ortografia e di Pronunzia della RAI (DOP)
versione online del vocabolario redatto a partire dal 1959 da Bruno Migliorini, Carlo Tagliavini e Piero Fiorelli (riv., agg. e accr. da P. Fiorelli e T. F. Borri) per la sede fiorentina della Radio Televisione Italiana, ora disponibile in versione multimediale. Il corpus, ricco di oltre 92.000 voci di lessico della lingua italiana e oltre 37.000 di una sessantina di lingue diverse, presenta anche la registrazione fonetica delle voci.

la Repubblica Corpus
corpus molto ampio (circa 380mila parole) del lessico del quotidiano *la Repubblica*. Nel progetto, curato dall'Università di Bologna, il corpus è stato lemmatizzato, taggato e categorizzato per genere e *topic*; gli articoli nel corpus sono strutturati nelle seguenti parti:*titolo, sottotitolo, sommario, testo*.

Corpora dell'italiano parlato

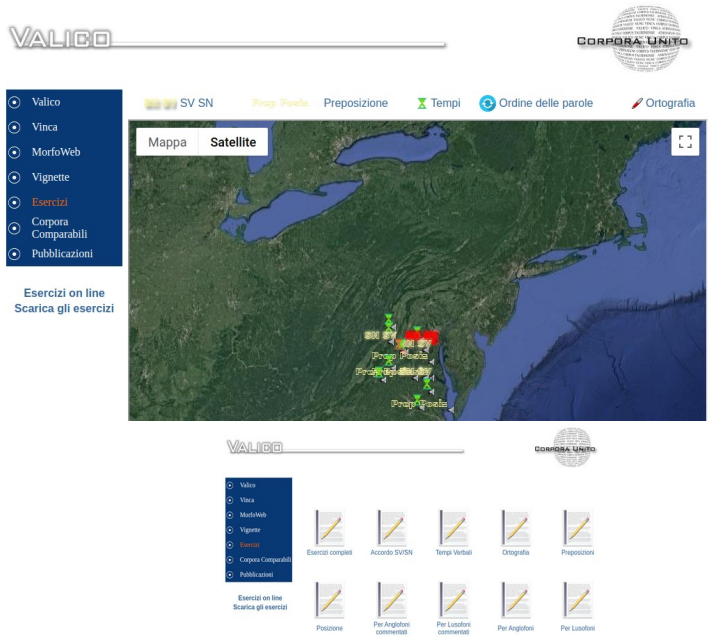
API (Archivio del Parlato Italiano)
progetto coordinato dal prof. Federico Albano Leoni (CIRASS - Napoli) cui hanno partecipato la Scuola Normale Sup. di Pisa, il CIRASS e il Dip. di Neuroscienze dell'Univ. "Federico II" di Napoli, l'Istituto Univ. Orientale di Napoli, il Politecnico di Bari, l'Univ. del Piemonte Orientale, l'Univ. "Ca' Foscari" di Venezia e l'Univ. di Pisa.
Il corpus, raccolto in tre città italiane (Napoli, Bari, Pisa), consiste essenzialmente di dialoghi *map task* e comprende anche un campione di parlato infantile di bambini sordi e normoudenti.

AVIP-API
progetto coordinato dal prof. Pier Marco Bertinetto (Scuola Normale Sup. - Pisa) cui hanno partecipato la Scuola Normale Sup. di Pisa, il CIRASS e il Dip. di Neuroscienze dell'Univ. "Federico II" di Napoli, l'Istituto Univ. Orientale di Napoli e il Politecnico di Bari. Sostanzialmente contiene lo stesso corpus del progetto API, che ne amplia lo spettro d'indagine, il campione e gli enti

Terzo risultato fornito da Google con query “corpus di italiano” è un ulteriore sito web **“Biblioteca Scuola Normale Superiore”**che riunisce in un unica bacheca dizionari e corpora dell’italiano. Distingue fra corpora dell’italiano scritto e parlato, settoriale e antico fornendone una breve descrizione e link alla risorsa originale.

Da questi esempi si può quindi comprendere che esistono vari siti che forniscono una bacheca da cui collegarsi a vari corpora **ma non forniscono delle istruzioni d’uso per gli stessi.** Nota positiva che va ovviamente mantenuta è la distinzione fra italiano d’uso e antico e italiano scritto e parlato, essenziale per poter aiutare insegnanti e studenti ad orientarsi nella scelta del corpus giusto per la lezione o gli esercizi.

Proseguendo nell’analisi del mercato a livello di contenuti esistenti, proponendo come stringa di ricerca “**esercizi corpus italiano**” sul motore Google otteniamo soprattutto dei risultati **PDF** di vari articoli circa l'utilità dell’utilizzo dei corpora a fini didattici, per la creazione di esercizi, ma non una raccolta di eventuali esercizi da proporre.



Non si può ignorare in questo momento di analisi l’apporto fornito dalla piattaforma **Valico**, che ha una sezione esercizi, in cui è possibile osservare e scaricare alcuni esercizi proposti a studenti italiano l2, ed è inoltre possibile scaricare i risultati ottenuti in formato excel. Sebbene questi esercizi hanno un origine diversa, ovvero, vengono progettati dopo aver osservato i risultati ottenuti dal learner corpus e costruiti per valutare correttamente gli “errori” commessi dall’apprendente, è senz’altro interessante la scelta di mostrare i risultati ottenuti distinguendoli per tipo di esercizi tramite legenda e ponendoli a seconda della provenienza degli apprendenti su una mappa geografia. Questo modo di esporre i risultati può senz’altro essere implementato su Corpora in Classe nella sezione learner corpora, per fornire agli studiosi uno strumento di visualizzazione ulteriore dei risultati.