



**Charles University, Faculty of Arts**  
**Institute of the Czech National Corpus**

nam. Jana Palacha 2, 116 36 Prague 1, Czech Republic  
tel.: +420 2 21 619 357, ucnk@ff.cuni.cz

---

## Il Progetto

### Corpus Nazionale Ceco e corpora di altre lingue

progetto n. 0021620823, finanziato dal Ministero dell'Istruzione, della gioventù e dell'educazione fisica della Repubblica ceca e realizzato presso l'Università "Carlo" di Praga, Facoltà di Filosofia, direttore responsabile prof. PhDr. František Čermák, DrSc.

si compone di alcuni sottoprogetti, tra cui

#### Il progetto *Intercorp*

Il progetto *Intercorp* si propone di creare dei corpora paralleli, costituiti cioè da traduzioni dal ceco verso tutte le lingue europee di maggiore diffusione e verso alcune tra le meno diffuse, sulla base dei dati del Corpus Nazionale Ceco e in collegamento con esso. Il progetto comprende in particolare l'inglese, il tedesco, il francese, il russo, lo slovacco, lo spagnolo e l'italiano. Tra le lingue meno diffuse saranno rappresentate l'arabo, il bulgaro, il danese, il finlandese, il lituano, l'ungherese, il macedone, il norvegese, il polacco, il portoghese, lo sloveno, il serbo, lo svedese ed eventualmente altre, come ad es. il giapponese, e il cinese (saranno i singoli esperti delle aree linguistiche citate, attivi presso l'Università "Carlo" a garantire la realizzazione dei vari sottoprogetti). I corpora paralleli avranno diverse funzioni, dalla ricerca individuale, soprattutto contrastiva, alla creazione e al perfezionamento di dizionari, alla pratica didattica in ambito universitario in diverse discipline linguistiche. I compiti delle singole unità, tra loro collegate, saranno i seguenti:

- a) La progressiva creazione dei singoli corpora paralleli, inclusa l'attuazione di strategie linguistiche sia complessive che specifiche, la soluzione del problema dei copyright, le forme di catalogazione e di registrazione, e la visualizzazione sul Corpus Nazionale Ceco e la connessione a esso, la formazione e il sostegno dei partner del progetto per i problemi di carattere informatico, il coordinamento del gruppo e dei sottogruppi;
- b) la progressiva acquisizione di testi (di dimensioni variabili tra 400.000 e 1.000.000 di parole, a seconda della accessibilità, dell'importanza culturale della lingua e del volume dei dati accessibili in traduzione), la loro catalogazione, l'acquisizione e il testaggio del software destinato all'elaborazione bilaterale nella prima fase ecc...;
- c) l'elaborazione dei testi, soprattutto l'impegnativo allineamento e in alcuni casi anche il basso livello della lemmatizzazione (in base alle possibilità e all'effettiva necessità), l'assistenza software ai singoli gruppi, il coordinamento complessivo dei gruppi indicati, l'estrazione degli equivalenti;
- d) lo svolgimento di progetti di ricerca correlati (in una tappa successiva, una volta che saranno stati creati dei corpora paralleli sufficientemente ampi), che dipenderà dai temi attuali e dalle necessità delle discipline rappresentate, inclusa la pubblicazione dei loro risultati, in workshop e riviste;
- e) la realizzazione dell'accesso al pubblico dei corpora paralleli (in base alle possibilità e al grado di disponibilità dei fornitori), soprattutto per lo studio e almeno per uso interno, eventualmente sul web.