**Proposta di definizione della struttura e dei contenuti del corpus CODE**

Incontro FARB del 26/02/2014

**Definizione del tipo di risorsa**

Corpus paragonabile di testi (guide web) in **lingua italiana** e in **lingua inglese ELF**, questi ultimi prodotti in Italia, nell'area francofona (Francia e Belgio) e nell'area germanofona (Germania e Austria) dell'UE.

I testi in lingua italiana e il loro equivalente in lingua inglese ELF formeranno un sub-corpus potenzialmente allineabile; per questo motivo, e per il solo caso dell'italiano, è necessario far corrispondere i testi in lingua italiana ai loro equivalenti inglesi (v. “Modalità di salvataggio dei testi”). Il reperimento di testi nelle lingue francese e tedesco, invece, è a discrezione dei rispettivi responsabili.

**Criteri di costruzione del corpus**

Ambito disciplinare: nove discipline totali selezionate secondo il seguente grado di granularità (il codice a fianco di ciascuna disciplina è ripreso dalla suddivisione effettuata dal Consiglio Europeo della Ricerca e disponibile alla pagina web <http://en.wikipedia.org/wiki/European_Research_Council>):

PHYSICAL SCIENCES

1. matematica (PE1)
2. informatica (PE6)
3. astrofisica (PE9)

SOCIAL SCIENCES

1. economia (SH1)
2. letteratura (SH5)
3. storia (SH6)

LIFE SCIENCES

1. fisiologia (LS4)
2. veterinaria (LS6)
3. psichiatria (LS5)

Grandezza del corpus: facendo una stima approssimativa del numero di token contenuti nelle guide web si presuppone siano necessari 20 testi per ciascuna disciplina. Ciascun sub-corpus (inglese-IT, inglese-FR, inglese-DE e italiano) sarebbe quindi costituito da 180 testi, per un totale di 540 testi in inglese ELF e 180 in italiano.

Atenei: la lista degli Atenei da cui selezionare le guide web è quella stilata da Webometrics (<http://www.webometrics.info/>) secondo criteri legati al grado di visibilità e alla presenza nel web.

Procedura di selezione degli Atenei:

* *per l'Italia*, visualizzare il ranking degli Atenei italiani (<http://www.webometrics.info/en/Europe/Italy%20>) e procedere in ordine di classifica dal primo verso l'ultimo, fino a quando si è giunti a 180 guide web in inglese e 180 in italiano (possibilmente equivalenti). L’Ateneo di Bologna va escluso dalla selezione visto che il Cesia si è dichiarato disponibile a fornire il database completo delle guide web.
* *Per i due paesi francofoni e i due paesi germanofoni*, visualizzare il ranking degli Atenei in Europa (<http://www.webometrics.info/en/Europe>) e procedere in ordine di classifica selezionando gli Atenei di entrambi i paesi (riconoscibili dalle bandiere nazionali), fino a quando si è giunti a 180 guide web per lingua.

Altri parametri di costruzione del corpus

* non includere nel corpus più di tre testi per disciplina per ciascun Ateneo
* per i due paesi francofoni e i due paesi germanofoni, equilibrare il numero di testi in modo che per ciascuna disciplina siano presenti dieci testi prodotti in uno Stato e dieci prodotti nell'altro (con una tolleranza di uno/due testi nel caso risultasse difficile raggiungere il numero auspicato di testi per paese per disciplina)
* non salvare i riferimenti bibliografici (eventualmente tenere solo il titolo della sezione, ad es. “References”/ “Bibliography” ecc.)
* evitare di includere guide web in ELF in cui il nome del docente responsabile suggerisce che non provenga dal paese che si sta considerando (ad es. se il nome suggerisce che sia madrelingua inglese)
* non copiare testo non rilevante (ad es. menu di navigazione ai margini delle pagine web)
* assicurarsi che le guide web selezionate non contengano parti di testo in una lingua diversa da quella cercata

**Modalità di salvataggio dei testi**

Salvare un file di testo (con estensione **.txt**) per ogni guida web. Il testo va fatto precedere da un header contenente i seguenti metadati (ciascuna informazione deve essere seguita da un ritorno a capo):

* n. progressivo (ad es. “001”)
* URL (ad es. “[http://www.eng.unibo.it/etcetc](http://www.eng.unibo.it/eccecc)”)
* nome breve dell’ateneo (ad es. “unibo”, “uni-graz”, “univ-bordeaux”)
* paese (ad es. “francia”)
* disciplina secondo la classificazione dell’ERC (ad es. “economia”)
* nome dell’attività formativa (ad es. “Finance”)
* UG, PG o NA (Undergraduate, Postgraduate, Not Available).

Relativamente ai testi in italiano e in inglese-IT, inserire dopo il testo della guida web l’eventuale n. di testo equivalente, seguito da EN/IT

Inserire un’altra riga vuota dopo la fine dell’header (di seguito un esempio della struttura del file).

|  |
| --- |
| id=001  url=http://master-mea.univ-lille1.fr/res/unit6\_a.pdf  uni=univ-lille1  paese=francia  disciplina=economia  nome=Finance  status=PG  UNIT 6: FINANCE  Course title: European monetary and financial system Teacher: Yannick Bineau  Unit code: Unit 6, Semester 4.  ECTS: 2  PREREQUISITES: Euro bachelor in economics or management or equivalent  COURSE DESCRIPTION:  The course focuses on the recent evolutions of the European Monetary System, on the challenges faced by the international financial integration within Europe, on the architecture of the European and International financial and monetary system, and on global imbalances.  AIMS:  The aims of this course are :  - To study the main evolutions of the European Monetary Systems,  - To quantify the importance of European financial markets  integration,  - To appreciate the situation of European Institutions within monetary  and financial markets,  - To be able to evaluate the dynamics of International Monetary  Systems and global imbalances.  INTENDED LEARNING OUTCOMES:  After completing this unit the student should be able to :  - Identify the main needs and constraints faced by Europe in term of  financial and monetary integration,  - Have a good grasp of what the European Monetary and Financial  markets functioning  - Propose main economic analysis’ tools to nowadays global  imbalances  TEACHING AND LEARNING ACTIVITIES:  Lectures and colloquia: 15 hours Self-tuition: 45 hours  Total student effort: 60 hours  ASSESSMENT:  Examination on completion of teaching period: written or oral  BIBLIOGRAPHY: |