***CODE:* C**ataloghi dell’**O**fferta **D**idattica in **E**uropa

**Resoconto dell’incontro del 26/03/2014**

**PARTECIPANTI**

* M. Soffritti (coordinatore progetto)
* S. Bernardini
* E. Dalan
* A. Ferraresi
* E. Wiesmann
* E. Zanchetta
* A. Zingaro

**ORDINE DEL GIORNO**

1) Verifica della fattibilità del corpus monolingue paragonabile secondo i criteri stabiliti nell’incontro precedente;

2) Implementazione e utilizzo di un’applicazione web per l’inserimento dei testi e la costruzione di corpora a partire da essi;

3) Incontro di fine primo anno del progetto.

**DISCUSSIONE**

***1) Verifica della fattibilità del corpus monolingue paragonabile secondo i criteri stabiliti nell’incontro precedente***

I membri del progetto hanno riportato le loro impressioni sull’effettiva fattibilità del corpus di guide web italiane e di area anglofona, francofona e germanofona secondo i criteri stabiliti durante l’incontro del 26/02/2014. Questi criteri prevedevano un bilanciamento del numero di testi in base a aree e sotto-aree disciplinari, università e paesi di provenienza.

Per quanto riguarda il corpus di inglese nativo (guide web dal Regno Unito/Irlanda), la raccolta di testi secondo i parametri citati non si è presentata particolarmente problematica, e si è riusciti a raggiungere il numero stabilito di 180 testi, suddivisi equamente per paese/università/(sotto-)area disciplinare, come anche mostrato nella Tabella 1.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **N. TESTI / TOKEN** | | | | | | | | |
| **University** | PHYSICAL SCIENCES | | | SOCIAL SCIENCES | | | LIFE SCIENCES | | |
|  | math | computing | astrophysics | economics | literature | history | physiology | vet | psychiatry |
| **Cambridge** | 2/640 | 2/1057 | 2/625 | 2/538 | 0/0 | 2/867 | 2/547 | 2/579 | 1/221 |
| **Oxford** | 3/1110 | 2/671 | 0/0 | 3/544 | 0/0 | 3/1275 | 0/0 | 0/0 | 1/106 |
| **UCL** | 2/282 | 2/786 | 3/1243 | 2/414 | 3/815 | 3/882 | 2/303 | 0/0 | 0/0 |
| **Edinburgh** | 2/391 | 2/423 | 2/449 | 0/0 | 2/387 | 2/487 | 3/1050 | 2/930 | 2/886 |
| **Nottingham** | 1/470 | 2/801 | 2/1211 | 1/490 | 3/1706 | 0/0 | 2/856 | 3/2235 | 3/2614 |
| **Glasgow** | 0/0 | 0/0 | 1/213 | 2/369 | 2/412 | 0/0 | 1/263 | 3/749 | 3/672 |
| **TCD** | 3/588 | 3/696 | 3/598 | 3/1183 | 3/1050 | 3/872 | 3/765 | 1/255 | 2/204 |
| **UCD** | 3/1026 | 3/1047 | 3/1223 | 3/1363 | 3/1370 | 3/1343 | 3/1100 | 3/1282 | 3/1226 |
| **UCC** | 3/968 | 3/916 | 3/944 | 3/1251 | 3/1031 | 3/1160 | 3/1101 | 3/1014 | 3/1125 |
| **DCU** | 1/595 | 1/472 | 1/470 | 1/678 | 1/615 | 1/856 | 1/459 | 0/0 | 2/1161 |
| **Limerick** | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 3/1189 | 0/0 |

Tabella 1. Dati riassuntivi sul corpus nativo di guide web

Non si può dire lo stesso per la raccolta di testi per il corpus di inglese ELF (guide web dall’Italia, Francia/Belgio e Germania/Austria), che si è rivelata un processo lungo e poco produttivo. In particolare, i membri del progetto hanno riportato numerose difficoltà nel bilanciare il numero di testi secondo i parametri stabiliti, e in generale nel reperire testi in ELF di qualità accettabile. Nella maggior parte dei casi, infatti, si è appurato che il campione di università preso in considerazione in questa prima fase offre un numero esiguo di testi in inglese, e anche quelli presenti sono molto brevi, oppure presentano solo alcuni frammenti in lingua inglese, solitamente il *template* pre-impostato dall’università, mentre i contenuti rimangono nella lingua nazionale.

Per questa ragione si è deciso un cambio piuttosto radicale nella tipologia di testi da campionare, prendendo in considerazione esclusivamente guide web pubblicate da corsi di laurea internazionali interamente insegnati in inglese, che nella prima fase si era invece deciso di escludere. Tale scelta presenta numerosi vantaggi, sia dal punto di vista teorico/metodologico che applicato. Per loro natura, i corsi internazionali sono fortemente orientati ad attrarre studenti dall’estero e le loro guide web in inglese riflettono pertanto un’esigenza comunicativa concreta; questo non è necessariamente vero per le guide web in inglese di altri corsi di laurea, che invece in molti casi sembrano esistere soltanto per adempiere a obblighi istituzionali. Sembra inoltre lecito supporre che l’importanza delle guide web all’interno della comunicazione dei corsi di laurea internazionali sia anche garanzia di maggiore qualità dei testi, il che ai fini del progetto dovrebbe velocizzare il reperimento di testi accettabili per il corpus.

Alla luce dei nuovi sviluppi, si è deciso di modificare i criteri per la selezione dei testi da inserire nel corpus, e di procedere nel seguente modo (a questo proposito si rimanda anche al file *definizione\_corpus\_code\_v2.doc*):

1. partire da una lista di corsi di laurea internazionali offerti da ciascuno Stato (ad es. per la Germania: <https://www.daad.de/deutschland/studienangebote/international-programs/en/>);

2. effettuare, per quanto possibile, un bilanciamento dei testi a) per macro-aree disciplinari, ossia *physical*, *social* e *life sciences* (circa un terzo dei testi per ognuna di queste categorie); b) per atenei (mirando a includere testi da più corsi di laurea/atenei diversi); il numero di testi da raggiungere è, idealmente, 150;

3. includere solo testi con almeno 150 parole di contenuto coeso;

4. includere guide web da corsi non internazionali solo se davvero interessanti;

5. salvare ogni guida web in formato *.pdf*(dal browser: “Stampa” > come file *pdf*), per tenere traccia di elementi eliminati nella conversione in formato solo testo, ad es. formattazione e immagini;

6. tenere nota delle fasi della ricerca (ad es. i motivi per cui sono stati esclusi determinati corsi), e dei criteri adottati per riconoscere il genere (ad es. a livello di terminologia).

Oltre alla tipologia di testi, si è deciso di modificare l’insieme di metadati da inserire per ogni testo (v. anche Sezione 2). In particolare si segnala l’aggiunta del metadato “*lingua\_autore*”: benché possa risultare difficile reperire questa informazione, sarebbe importante tenerne traccia, stabilendo soprattutto se un testo è stato scritto da un autore di madrelingua inglese oppure no. Inoltre, è stato aggiunto un campo “*qualità*”, in cui è possibile segnalare se il testo risulta intuitivamente di qualità bassa/media/alta (ad es. considerando la completezza del testo nelle sue parti normative, secondo i descrittori di Dublino). Infine, nel campo “*PDF*” andrà inserito un riferimento al file *pdf* della guida web (v. punto 5 dell’elenco sopra).

È qui riportato un esempio di *header* completo, con i metadati da registrare in fase di raccolta dei testi:

**id**=001

**lingua\_testo**=EN

**paese**=Francia

**lingua\_autore**=Francese

**url**=http://master-mea.univ-lille1.fr/res/unit6\_a.pdf

**uni**=univ-lille1

**macroarea**=life

**nome\_corso**=Finance

**status**=PG

**qualità**=Low/Medium/High

**corso internazionale**=BOOL (default yes)

**copyright\_status**=Unknown/Creative Commons/Other (default: Unknown)

**note**=

**date**=2014

**PDF**=file in PDF

Sulla base dei criteri descritti in questa sezione verrà effettuata una seconda verifica di fattibilità della costruzione del corpus, di cui si discuterà nel prossimo incontro.

***2) Implementazione e utilizzo di un’applicazione web per l’inserimento dei testi e la costruzione di corpora a partire da essi***

È in fase di sviluppo un’applicazione web (costituita da un’interfaccia utente e da un database) che permetterà ai membri del progetto di inserire i testi che confluiranno nel corpus CODE e di generare corpora formati da una selezione dei testi stessi. L’interfaccia di inserimento permette di separare il testo della guida web dai metadati relativi al testo stesso: fino a questo momento testo e metadati sono stati inseriti in un unico file *.txt*, ma così facendo aumentano le probabilità di introdurre errori formali, che renderebbero difficile il trattamento automatico dei dati. Inoltre, il database permetterà di controllare in ogni momento lo stato di avanzamento della raccolta dei testi e di selezionarli in base a parametri definiti dagli utenti (sfruttando i metadati). Tutto questo dovrebbe anche semplificare la collaborazione tra le persone coinvolte nella costruzione del corpus.

Si prevede che, una volta affinata, l’applicazione web potrà essere utilizzata anche da utenti esterni al progetto, anche per raccogliere corpora non necessariamente di guide web. In questo senso, l’applicazione potrebbe attrarre l’interesse di un numero ampio di utenti interessati alla costruzione collaborativa di corpora, e potrebbe costituire uno dei *deliverable* del progetto CODE.

***3) Incontro di fine primo anno del progetto***

Al termine del primo anno di progetto, indicativamente a giugno 2014, verrà organizzato un incontro per presentare i risultati raggiunti fino a quel momento e iniziare a riunire i portatori di interesse. A questo scopo si è pensato di invitare soggetti appartenenti a due categorie principali, ossia:

1. Accademici interessati all’inglese istituzionale e alla lingua inglese ELF, sia italiani (ad es. Giuseppe Palumbo, Marco Venuti, Francesca Vigo, personale dell’EURAC), sia stranieri (ad es. Anna Mauranen o Heidi Depraetere);

2. Figure istituzionali collegate all’internazionalizzazione delle università, sia interne all’Ateneo di Bologna (ad es. Giovanna Filippini, dirigente del DIRI, e Valentina Comba, responsabile della manutenzione della banca dati terminologica disponibile sulla Intranet di Ateneo), sia esterne, collegate ad es. alla Direzione Generale per l'Istruzione e la Cultura dell’Unione Europea (a questo scopo si potrebbe scrivere a Rodolfo Maslias come primo contatto).

**PROSSIMO INCONTRO: 16 APRILE 2014**

***Riassunto delle attività da svolgere prima dell’incontro:***

* individuare una lista di corsi internazionali in lingua ELF all’interno della propria area geografica di competenza (italiana, francofona o germanofona);
* partendo dalla lista di corsi individuata, verificare la reperibilità delle guide web in lingua ELF, l’accessibilità ai testi e la qualità degli stessi, al fine di effettuare una seconda verifica di fattibilità della costruzione del corpus;
* scaricare un campione di testi nei formati *txt* e *pdf*, secondo i criteri descritti nella Sezione 1, tenendo traccia delle ricerche svolte e registrando le informazioni da inserire come metadati.