European Language Resource Coordination
Connecting Europe Facility

Marcello Soffritti

- Bio: Studi universitari, laurea e ricerche a Bologna, Kiel e Monaco di Baviera. Già Preside di Facoltà, Direttore di Dipartimento e Coordinatore di Dottorato. Dagli anni 90 partecipazione a numerosi progetti di ricerca in ambito linguistico e glottodidattico.
- Ruolo:
  - Dal 1991 Professore ordinario al Dipartimento di Interpretazione e Traduzione, Università di Bologna (Sede di Forlì) – Co-Responsabile del centro **CoLiTec: Corpora, Linguistica, Tecnologia**
  - Dal 2001 Direttore dell'Istituto di Comunicazione Specialistica e Multilinguismo dell' EURAC (Accademia Europea, Bolzano)
- Responsabilità: In entrambe le istituzioni, attività scientifica, didattica, progetti e pubblicazioni nei seguenti campi:
  - Traduzione multimediale e tecnica (in particolare per testi giuridici e istituzionali)
  - Corpora e Linguaggi specialistici
  - Terminologia
  - Lessicologia e Lessicografia
- Abstract: la quantità di materiale validato disponibile presso le due Istituzioni è altissima, ma ai fini della traduzione automatica ancora molto frammentata e disparata. Per un recupero si richiede un'azione complessa: verifica tecnica, conversione dei formati, infrastruttura tecnica e gestionale, ecc.. Per una messa a sistema occorre una pianificazione radicalmente nuova di tutti i flussi di lavoro da cui possa derivare materiale linguistico in versione bilingue o plurilingue. Allo stesso tempo, è indispensabile uno sforzo di collaborazione tra istituzioni partner in diversi paesi europei.

SOMMARIO

- Quali sono i tipi di dati/risorse che possono essere messi a disposizione
  - Corpora giuridici it-ted – (codici italiani e tedeschi, raccolte di leggi nazionali e regionali in italiano e in tedesco, in parte allineati). (EURAC)
  - Portali terminologici (terminologia giuridica e amministrativa per il diritto civile, penale, amministrativo, universitario, ecc. in italiano e in tedesco, insieme ai corpora da cui è stata estratta). (EURAC)
  - Applicazioni di linguistica computazionale (annotazione, visualizzazione dei risultati delle query, Tool per la disambiguazione dei nomi propri, ecc). (EURAC)
  - Corpora di riferimento di grandi dimensioni per varie lingue europee (WaCky corpora). (Colitec UNIBO)
  - Corpora e glossari relativi al linguaggio istituzionale universitario. (Colitec UNIBO)
  - Corpora, database terminologici e memorie di traduzione frutto di attività didattica, in domini vari (Colitec UNIBO)
  - Software user-friendly per la costruzione semi-automatica di corpora. (Colitec UNIBO)
- Come sono questi dati rispetto a parametri:
  - Monolingue e multilingue, in parte annotati, in parte allineati, in diverse combinazioni linguistiche comprendenti l'italiano
- Quali sono i formati tipici?
  - Vari (Multiterm, XML, TXT)
- Come sono prodotti?
  - Metodi automatici e manuali, verifiche semiautomatiche e manuali da parte di staff e talvolta esperti esterni (tesi di laurea)
- Come sono distribuiti i dati?
  - Per quanto possibile pubblicati, ma talvolta sottoposti a vincoli di proprietà intellettuale e aziendale
- Chi li usa già?
  - Partner istituzionali, progetti collegati, docenti e studenti, professionisti (nei limiti previsti da regolamenti e accordi specifici)
- Come sono prodotti, trattati e distribuiti i dati?
  - In stretta osservanza delle disposizioni nazionali
- Dove sono collocati, depositati, archiviati e mantenuti i dati?
  - Server dedicati di proprietà delle relative istituzioni

PRODOTTI E RISORSE - EURAC

- Corpora giuridici it-ted – (codici italiani e tedeschi, raccolte di leggi nazionali e regionali in italiano e in tedesco, in parte allineati).
- Portali terminologici (terminologia giuridica e amministrativa per il diritto civile, penale, amministrativo, universitario, ecc. in italiano e in tedesco, insieme ai corpora da cui è stata estratta): BISTRO, LexAlp – circa 50.000 termini.

– Applicazioni di linguistica computazionale (annotazione, visualizzazione dei risultati delle query, Tool per la disambiguazione dei nomi propri, ecc).
– Studi sull'organizzazione della comunicazione aziendale bilingue e plurilingue.
– ACCESSIBILITA' ONLINE

PRODOTTI E RISORSE – COLITEC UNIBO

- Corpora di riferimento di grandi dimensioni per varie lingue europee (WaCky corpora, con corpus query system interno (CWB/CQP).
- Corpora, database terminologici e memorie di traduzione frutto di attività didattica, in domini vari.
- Corpora e glossari relativi al linguaggio istituzionale universitario
- Software user-friendly per la costruzione semi-automatica di corpora.

CORPORA PARALLELI COLITEC UNIBO

- Corpora paralleli: corpora di generi specializzati, principalmente descrizioni di moduli accademici, ritenuti testi chiave per favorire la mobilità studentesca a livello europeo;
- 2 set di corpora
  - Corpus UNIBO: le versioni italiane e inglesi delle descrizioni dei corsi di UNIBO nell'a.a. 2013/2014 (2 milioni di parole)
  - Corpus CODE parallelo: le versioni italiane e inglesi delle descrizioni dei corsi di varie università italiane (ca. 20.000 parole)
- Caratteristiche
  - Corpora bilingui italiano / inglese
  - Arricchiti con metadati (ad es. nome della disciplina, del corso di laurea, ecc.) e con informazioni su lemmi e parti del discorso
  - Allineati
- Formati disponibili: testo semplice + testo annotato + memoria di traduzione (TMX)

TERMINOLOGIA COLITEC UNIBO

- Circa 16.000 schede.
- Domini principali (> 10 progetti ognuno): alimentazione, edilizia/architettura, ecologia, energia, lavoro/immigrazione, manualistica, medicina, trasporti.
- Lingue: bilingue o trilingue, con italiano + (in ordine decrescente di frequenza) inglese, francese, spagnolo, tedesco, russo.
- Formati disponibili: formati proprietari (ad es. mdb) o di interscambio (TBX)
- Metodo per la costruzione: manuale

MEMORIE DI TRADUZIONE COLITEC UNIBO

- ca. 220.000 segmenti allineati
- Bilingue (di solito includendo l'italiano).
- Formati disponibili: TMX
- Metodi per la costruzione: allineamento automatico, solitamente con controllo manuale

CORPORA PER TERMINOLOGIA COLITEC UNIBO

- ca. 190 milioni di parole.
- Monolingue (544) o paralleli (92)
- Se paralleli, di solito allineati a livello di frase
- Formati disponibili: (per la maggior parte) testo semplice, non annotato
- Metodi per la costruzione: manuali (301) o semi-automatici (BootCaT: 258)

CARATTERISTICHE GENERALI DEL MATERIALE

- Materiale monolingue e multilingue.
- Corpora in parte annotati, in parte allineati (TXT, XML).
- Glossari terminologici in diversi formati (MDB, XML, EXCEL, ecc.)
- Materiale bi- e multilingue in diverse combinazioni linguistiche, comprendenti sempre l'italiano.

CREAZIONE E VALIDAZIONE

- Metodi automatici e manuali, elaborati e aggiornati nel corso di 10 – 15 anni.
- Verifiche semiautomatiche e manuali da parte di staff e talvolta esperti esterni (ad es. tesi di laurea).
- In alcuni casi terminologia validata da parte di apposite commissioni.

European Language Resource Coordination
Connecting Europe Facility

CONCLUSIONI

- La quantità di materiale validato disponibile presso le due Istituzioni è altissima, ma ai fini della traduzione automatica ancora molto frammentata e disparata.
- Per un recupero si richiede un'azione complessa: verifica tecnica, conversione dei formati, infrastruttura tecnica e gestionale, ecc..
- Per una messa a sistema occorre una verifica o ripianificazione di tutti i flussi di lavoro da cui possa derivare materiale linguistico in versione bilingue o plurilingue.
- Allo stesso tempo, è indispensabile uno sforzo di collaborazione tra istituzioni partner in diversi paesi europei.