Risorse e applicazioni computazionali per l'accesso ai beni culturali: il Corpus CHerIDesCo

Gloria Gagliardi°, Massimo Guarino°[1]
°Università di Napoli "L'Orientale"

The paper presents CHerIDesCo – *Cultural Heritage - Italian Description Corpus*, a domain-specific linguistic resource designed for the training and testing of novel NLP tools in the Cultural Heritage field. The corpus has been developed by the UNIOR NLP Research group as a part of the SMACH project, a three-year project funded by the National Operative Program to pursue the Smart Specialization Strategy defined by the EU. The project aims at improving language-based human-computer interaction in the Cultural Heritage domain through the development of innovative applications for multilingual access to the contents based on semantic language technologies. In particular, the paper describes the design of the CHerIDesCo corpus, the annotation procedures, and the platforms where the resource has been uploaded. As pointed out in the conclusion, this linguistic resource can be exploited in several NLP tasks (e.g., NER – Named-Entity Recognition, NEL – Named-Entity Linking, and Topic Modeling).

Keywords: BBCC, Cultural Heritage, NLP – Natural Language Processing

1. Introduzione

Il termine *Cultural Heritage* (CH, in italiano 'Patrimonio Culturale') designa l'insieme dei beni tangibili (es. siti o reperti archeologici, opere pittoriche o scultoree) o intangibili (es. feste e riti religiosi, cerimonie folkloristiche, gare sportive e agonistiche storiche) che, per particolare rilievo storico-culturale ed estetico, sono di interesse pubblico (UNESCO 1972). In particolare, secondo

[1] Sia il corpus CHerIDesCo, sia questo articolo che ne descrive struttura e finalità sono il frutto del lavoro congiunto dei due autori. Tuttavia, data la diversità delle reciproche competenze, Gloria Gagliardi si è occupata del *corpus design*, della raccolta dei testi e della compilazione dei metadati; Massimo Guarino è invece responsabile dell'annotazione della risorsa. Alla prima si attribuiranno §1, §2, §3.1, §3.3 e §4.1, al secondo §3.2, §4.2 e §5.

CHIMERA. Romance Corpora and Linguistic Studies 8 (2021), 25-43. ISSN 2386-2629
© 2021 The Author(s); distributed under the Creative Commons Attribution License

l'ordinamento giuridico italiano (Decreto Legislativo 22 gennaio 2004, n. 42, *Codice dei beni culturali e del paesaggio*, ai sensi dell'articolo 10 della legge 6 luglio 2002, n. 137) nel patrimonio culturale sono inclusi beni culturali e beni paesaggistici.

Negli ultimi decenni le istituzioni che si occupano di CH (gallerie, librerie, archivi e musei, indicate di solito con l'acronimo inglese GLAM – *Galleries, Libraries, Archives and Museums*) hanno iniziato a utilizzare su larga scala risorse e applicazioni di tipo informatico per la conversione in formato digitale, la consultazione e la conservazione del patrimonio.

La progressiva digitalizzazione delle risorse ha però posto il tema dell'accesso a tali informazioni in cima all'agenda delle istituzioni, sia sul versante teorico che sul piano applicativo: infatti, per essere recuperati, condivisi e utilizzati dagli utenti, i documenti elettronici devono essere descritti secondo "schemi" e "regole" comuni, ovvero essere conformi a standard, ed essere accompagnati da metadati adeguati che ne descrivano le caratteristiche strutturali e tecniche.

Come evidenziato da Doerr & Stead (2011), dunque, non è suffiente mettere in rete i contenuti, ma è necessario integrarli in rappresentazioni formalmente coerenti, intregrabili e riutilizzabili, ovvero ricorrere a:

> [...] conceptual models or core ontologies of relationships for the digital world that are completely integrated and cover, in a complementary way, a vast spectrum of key conceptualizations for memory institutions and the management of digital content. Such core ontologies of relationships are fundamental to schema integration and play a vital role in practical knowledge management completely different to the role played by specialist terminologies. The vision is not merely to aggregate content with finding aids, as current DLs do, but to integrate digital information into large scale, trans-disciplinary networks of knowledge. These networks support not only accessing source documents, but also using and reusing the integrated knowledge embedded in the data and metadata themselves while managing the increasingly complex digital data aggregates and their derivatives.

Inoltre, per aumentare il numero dei potenziali visitatori dei luoghi di interesse culturale e dei fruitori delle risorse online sarebbe auspicabile rendere disponibili i contenuti mediante portali digitali multilingui.

Un forte impulso allo sviluppo di tali infrastrutture è stato dato dal progressivo affermarsi, in questo dominio, di tecniche proprie del NLP (*Natural Language Processing*, in italiano TAL – *Trattamento Automatico del*

Linguaggio) e delle *Digital Humanities*, in grado di garantire non solo la possibilità di definire modelli per la creazione, rappresentazione, estensione e mantenimento di lessici, repertori terminologici e ontologie, ma anche di accedere e gestire in maniera "intelligente" queste basi documentali, rispondendo in maniera flessibile ai bisogni informativi degli utenti.

Il presente lavoro si inserisce in questa linea di ricerca, presentando una risorsa linguistica sviluppata *ad hoc* per applicazioni computazionali nel dominio dei BBCC: il corpus CHerIDesCo.[2]

Il contributo è organizzato come segue: il §2 illustra SMACH – *Semantic Multilingual Access to Cultural Heritage*, progetto di ricerca all'interno del quale è stata sviluppata la risorsa. Il §3 è dedicato alla descrizione del corpus: *design*, annotazione e modalità di accesso. Il §4 presenta alcune informazioni di natura quantitativa sulle proprietà del lessico della risorsa, ponendola a paragone con un corpus di riferimento per l'italiano scritto, CORIS. Nel §5, infine, vengono indicate alcune possibili applicazioni e delineate alcune ipotesi di sviluppo futuro.

2. SMACH – Semantic Multilingual Access to Cultural Heritage

Il progetto SMACH, acronimo di *Semantic Multilingual Access to Cultural Heritage*[3], finanziato nell'ambito del Programma Operativo Nazionale "Ricerca e Innovazione" 2014-2020, è in linea con le esigenze di accessibilità e interoperabilità descritte nel §1. Si propone infatti di sviluppare tecnologie linguistiche volte a migliorare l'interazione tra l'uomo e il sistema informativo del dominio dei Beni Culturali (BBCC), mettendo a punto applicazioni per il recupero, l'estrazione e l'accesso multilingue alle basi di conoscenza del settore (es. archivi e database digitali del patrimonio museale) e sfruttando tecnologie semantiche innovative, quali i LLOD – *Linguistic Linked Open Data* (Chiarcos, Nordhoff & Hellman 2012; Chiarcos *et al.* 2013) e gli *embedding* (es. rappresentazioni di parole/sequenze di parole in spazi numerici di tipo vettoriale), nonché strumenti per la traduzione automatica/assistita. Le attività di ricerca del progetto si articolano intorno a due nodi principali:

– rappresentazione semantica cross-linguistica dei contenuti, raggiunta attraverso la conversione delle informazioni catalografiche dei BBCC in un'annotazione multilingue e multi-livello, compatibile con le ontologie e i

[2] https://cheridesco.altervista.org/
[3] https://sites.google.com/view/smach-project/home

formalismi standardizzati nel dominio dei beni culturali e delle *Digital Humanities*, come ad esempio CIDOC – *Conceptual Reference Model* (Doerr 2003, 2009) ed *Europeana data model* (Aloia, Concordia & Meghini 2011);
– analisi e *testing* di tecniche di *annotation projection* (Padó & Lapata 2009; Ehrmann, Turchi & Steinberger 2011; Tiedemann 2014) e di *Multilingual Machine Learning* (Spohr, Hollink & Cimiano 2011), finalizzati alla costruzione di un prototipo di sistema informativo multilingue.

Ad oggi, la possibilità di applicare approcci NLP ai BBCC in lingua italiana è però fortemente limitata dalla quasi completa assenza di risorse linguistiche e corpora di dominio, monolingui e multilingui, da utilizzare per il *training* e il *testing* degli algorimi.

3. Il corpus CHerIDesCo

Il corpus CHerIDesCo si propone di fornire una prima soluzione alla scarsità di risorse linguistiche di dominio per applicazioni NLP nell'ambito dei Beni Culturali: è infatti una raccolta bilanciata di testi descrittivi, prodotti dalle istituzioni e riferiti a musei, monumenti e siti archeologici statali italiani.

Per *testo descrittivo*, o più semplicemente *descrizione*, intendiamo in questa sede, sulla scorta di Roggia (2011):

> il risultato di un macro atto linguistico […], che consiste nel costruire un corrispondente linguistico di una porzione di mondo considerata da un punto di vista statico e atemporale.

La porzione di mondo cui può applicarsi una descrizione, definita *oggetto descrittivo*, può essere di vario tipo, purché si presti a essere considerata da un punto di vista statico: individui (oggetti, persone) o stati di cose (Bertinetto & Ossola 1982; Manzotti 2009). Nel caso in oggetto, i testi sono classificabili, per lo più, come *topografie*, ovvero descrizioni di luoghi (Mortara Garavelli 1988).

3.1 Corpus Design

Il corpus si compone, nella sua versione 1.0, di 680 testi in formato *plain-text* di varia lunghezza (da un minimo di 8 a un massimo di 6337 token; media = 436,57), per un totale di 296871 token. La risorsa è organizzata in 17 subcorpora regionali (Tabella 1); al momento sono stati esclusi dal campionamento i siti

delle regioni a statuto speciale Valle d'Aosta, Trentino-Alto Adige e Sicilia, che gestiscono e coordinano autonomamente musei, aree e parchi archeologici e monumenti del proprio territorio.

Tabella 1. Corpus Design

Subcorpus	n. di siti censiti	n. di token
CHerIDesCo–Abruzzo_subcorpus	23	26115
CHerIDesCo–Basilicata_subcorpus	21	6515
CHerIDesCo–Calabria_subcorpus	25	17627
CHerIDesCo–Campania_subcorpus	103	48582
CHerIDesCo–EmiliaRomagna_subcorpus	42	16902
CHerIDesCo–FriuliVeneziaGiulia_subcorpus	18	6336
CHerIDesCo–Lazio_subcorpus	135	59383
CHerIDesCo–Liguria_subcorpus	15	4417
CHerIDesCo–Lombardia_subcorpus	26	9643
CHerIDesCo–Marche_subcorpus	31	9281
CHerIDesCo–Molise_subcorpus	14	11005
CHerIDesCo–Piemonte_subcorpus	22	8837
CHerIDesCo–Puglia_subcorpus	21	9762
CHerIDesCo–Sardegna_subcorpus	60	18315
CHerIDesCo–Toscana_subcorpus	72	31706
CHerIDesCo–Umbria_subcorpus	14	6634
CHerIDesCo–Veneto_subcorpus	19	5811

I siti sono stati individuati a partire dalla lista messa a disposizione *online* dalla Direzione Generale Musei del Ministero per i beni e le attività culturali e per il turismo (MiBACT), con l'obiettivo di favorire la ricerca e la diffusione delle conoscenze riguardanti il patrimonio culturale italiano custodito nei musei e rappresentato nei luoghi della cultura italiani[4].

La risorsa è però stata progettata per crescere nel tempo, aumentando i punti di raccolta e il numero di testi riferiti ai siti censiti. A ogni testo è infatti associato un ID univoco, così strutturato:

Regione_IDSito_IDtesto

es. Abruzzo_001_01

[4] http://musei.beniculturali.it/musei

Museo archeologico nazionale di Campli – presentazione sito MiBACT

es. Abruzzo_001_02
Museo archeologico nazionale di Campli – Depliant

es. Abruzzo_002_01
Chiesa di San Pietro ad Oratorium, Capestrano – presentazione sito MiBACT

I testi che compongono il corpus sono stati raccolti e catalogati manualmente, non ricorrendo al *web-crawling* (Baroni & Ueyama 2006), per assicurare la massima qualità, adeguatezza e pulizia del dato.

3.2 Annotazione

Il corpus CHerIDesCo è stato tokenizzato e annotato dal punto di vista morfosintattico ricorrendo ai diversi *tool* contenuti in *Stanza* (Qi *et al.* 2018, 2020), un *package* Python *open-source* di strumenti per l'analisi computazionale del linguaggio verbale predisposto dallo Stanford NLP Group[5].

La *pipeline* utilizzata ha adottato i moduli *tokenize* (che divide il testo in frasi e, successivamente al loro interno, token), *POS* (che attribuisce la *Part of Speech*) e *lemma* (che provvede alla lemmatizzazione delle parole utilizzando come input i singoli token e i valori derivanti dall'annotazione della parte del discorso). In particolare, in relazione al *tagging* delle parti del discorso, la *pipeline* fornisce in output (Figura 1):

a. l'annotazione basata sul *tag-set* del *framework* Universal Dependencies (UD),[6] UPOS. Lo schema di annotazione, basato sui *tag* per l'annotazione morfosintattica di *Stanford dependencies* (de Marneffe *et al.* 2006, 2008, 2014), *Google* (Petrov *et al.* 2012)[7], e *Interset* (Zeman 2008)[8], include le seguenti categorie (Loos *et al.* 2003)[9]:

[5] https://nlp.stanford.edu/
[6] https://universaldependencies.org/. Citando gli stessi autori, lo scopo del progetto Universal Dependencies (UD) è di sviluppare "cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and parsing research from a language typology perspective".
[7] https://github.com/slavpetrov/universal-pos-tags
[8] http://ufal.mff.cuni.cz/interset
[9] http://www.sil.org/linguistics/GlossaryOfLinguisticterms/

- ADJ: aggettivo
- ADP: adposizione
- ADV: avverbio
- AUX: ausiliare
- CCONJ: congiunzione coordinante
- DET: determinante
- INTJ: interiezione
- NOUN: nome
- NUM: numerale
- PART: particella
- PRON: pronome
- PROPN: nome proprio
- PUNCT: punteggiatura
- SCONJ: congiunzione subordinante
- SYM: simboli
- VERB: verbo
- X: altro

b. l'annotazione in accordo con i *tag* PoS linguo-specifici della particolare *treebank* utilizzata, XPOS. Per l'italiano, il *tag-set* di riferimento è quello sviluppato dall'Istituto di Linguistica Computazionale A. Zampolli in collaborazione con l'Università di Pisa per l'annotazione del corpus ISST-TANL (Montemagni *et al.* 2003; Montemagni & Simi 2007), che include le seguenti etichette *fine-grained*[10]:

- A: aggettivo
  AP: aggettivo possessivo
- B: avverbio
  BN: avverbio di negazione
- CC: congiunzione coordinante
  CS: congiunzione subordinante
- D: determinante:
  DE: determinante esclamativo
  DI: determinante. indefinito
  DQ: determinante interrogativo
  DR: determinante relativo
  DD: determinante dimostrativo

[10] http://www.italianlp.it/docs/ISST-TANL-POStagset.pdf

- E: preposizione
  EA: preposizione articolata
- F: punteggiatura
  FB: punteggiatura "bilanciata" (es. « » – –)
  FC: punteggiatura che marca confine di clausola
  FF: virgola, trattino
  FS: punteggiatura che marca confine di frase
- I: interiezione
- N: numerale cardinale
  NO: numeroale ordinale
- P: pronome:
  PD: promome dimostrativo
  PE: pronome personale
  PI: pronome indefinito
  PP: pronome possessivo
  PQ: pronome interrogativo
  PR: pronome relativo
  PC: pronome clitico
- RD: articolo determinativo
  RI: articolo indeterminantivo
  S: sostantivo – nome comune
  SA: abbreviazione/acronimo
  SP: sostantivo – nome proprio
- V: verbo
  VA: ausiliare
  VM: modale
- X: altro

c. le caretteristiche lessicali e grammaticali non coperte dai PoS *tag*, dette *Universal morphological features*, UFeats.[11] Consentono di descrivere con maggior livello di dettaglio le proprietà della forma analizzata: ad esempio, di specificare, per una forma verbale, modo (*mood*), tempo (*tense*), aspetto (*aspect*), diatesi (*voice*), persona (*person*), evidenzialità (*evidentiality*), polarità (*polarity*).

L'annotazione è fornita in duplice formato, *.txt e *.csv.

[11] Per la lista completa si rinvia a https://universaldependencies.org/u/feat/index.html

```
,text,upos,xpos,lemma,feats
0,Museo,NOUN,S,museo,Gender=Masc|Number=Sing
1,archeologico,ADJ,A,archeologico,Gender=Masc|Number=Sing
2,nazionale,ADJ,A,nazionale,Number=Sing
3,di,ADP,E,di,
4,Campli,PROPN,SP,Campli,
5,Il,DET,RD,il,Definite=Def|Gender=Masc|Number=Sing|PronType=Art
6,museo,NOUN,S,museo,Gender=Masc|Number=Sing
7,ha,VERB,V,avere,Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
8,sede,NOUN,S,sede,Gender=Fem|Number=Sing
9,in,ADP,E,in,
10,alcuni,DET,DI,alcuno,Gender=Masc|Number=Plur|PronType=Ind
11,ambienti,NOUN,S,ambiente,Gender=Masc|Number=Plur
12,di,ADP,E,di,
13,l',DET,RD,il,Definite=Def|Number=Sing|PronType=Art
14,antico,ADJ,A,antico,Gender=Masc|Number=Sing
15,convento,NOUN,S,convento,Gender=Masc|Number=Sing
16,di,ADP,E,di,
17,San,PROPN,SP,San,
18,Francesco,PROPN,SP,Francesco,
19,fondato,VERB,V,fondare,Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part
20,verso,ADP,E,verso,
21,la,DET,RD,il,Definite=Def|Gender=Fem|Number=Sing|PronType=Art
22,fine,NOUN,S,fine,Gender=Fem|Number=Sing
23,di,ADP,E,di,
24,il,DET,RD,il,Definite=Def|Gender=Masc|Number=Sing|PronType=Art
25,XIII,ADJ,NO,XIII,NumType=Ord|Number=Sing
26,secolo,NOUN,S,secolo,Gender=Masc|Number=Sing
27,.,PUNCT,FS,.,
```

Figura 1. Annotazione morfo-sintattica (numero progressivo, token, UPOS, XPOS, lemma, *feature* morfo-sintattiche)

A ciascun elemento della risorsa (i.e. corpus, subcorpus, testo) è associato un file di metadati in formato xml (estensione: *.imdi), compilato seguendo le definizioni previste dallo standard internazionale IMDI – *ISLE Meta Data Initiative*. In particolare, i file di metadati contengono, in riferimento a ciascun testo della risorsa:

– ID, nome completo del sito censito e sua localizzazione geografica;
– URL da cui sono stati acquisiti i contenuti e responsabile del dato (es. Direzione Regionale Musei, Soprintendenza);
– data di scaricamento del testo;
– codifica dei caratteri adottata (*character encoding*);
– livello (*0-100*) e metodologia (*manuale*, *automatica*, *semi-automatica*) di validazione del dato;
– dimensione del file (in *byte*).

Per la creazione e la gestione dei metadati è stato utilizzato il software Arbil (Withers 2012)[12].

[12] https://archive.mpi.nl/forums/t/arbil-information-manuals-download/1045

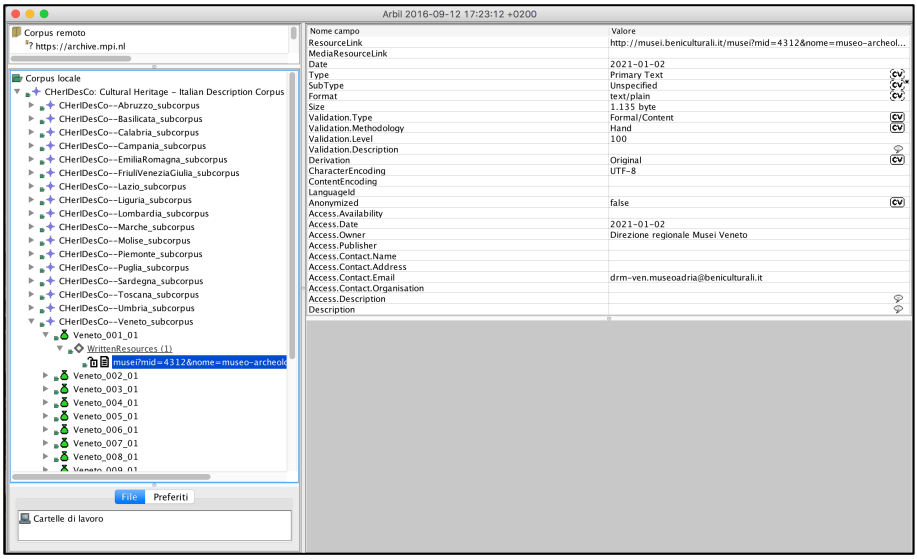

Figura 2. Creazione e gestione dei metadati con il software Arbil

3.3 Accesso

Il corpus è disponibile su due piattaforme: Github, nella pagina del gruppo Unior NLP dedicata alle risorse del progetto (https://github.com/unior-nlp-research-group/SMACH-corpora-CheriDesCO), e Sketch Engine (ske.li/cheridesco).[13]

La doppia distribuzione è volta ad assicurare la massima usabilità del corpus per tutte le tipologie di utenti. Nel primo caso, infatti, l'intera risorsa (testi, metadati e annotazione) è liberamente scaricabile e utilizzabile per applicazioni avanzate NLP (§4); nel secondo, l'utente finale può disporre di un'interfaccia intuitiva per compiere interrogazioni di natura lessicale sui testi (Kilgarriff *et al.* 2004, 2014): ad esempio estrazione della lista di frequenza, consultazione delle concordanze, *word sketch*, n-grammi, analisi delle *keyword* (Figura 3).

[13] Attualmente CHerIDesCo è condiviso nella piattaforma Sketch Engine, ma non è interrogabile come *open corpus*. Gli utenti che volessero consultare il corpus possono richiedere l'accesso agli autori.

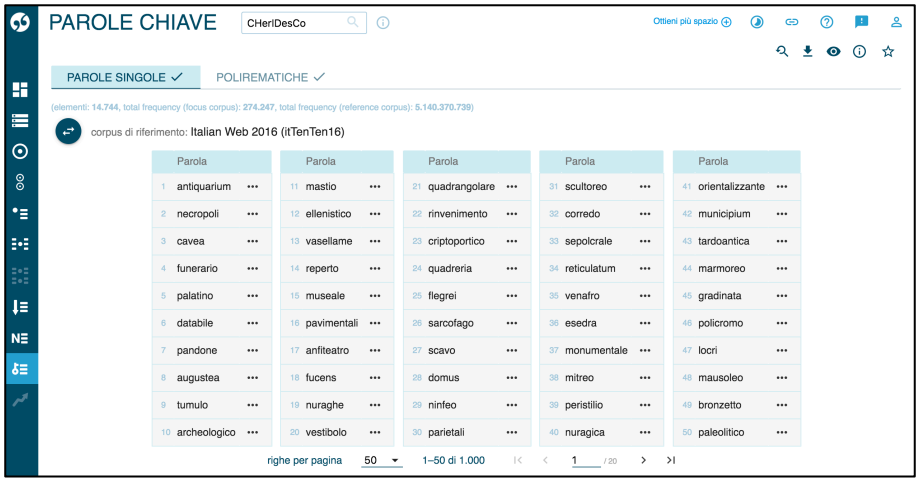

Figura 3. Analisi delle keyword con Sketch Engine

4. Il lessico del corpus CHerIDesCo: alcuni dati quantitativi

Per fornire al lettore un quadro più completo delle caratteristiche linguistiche di CHerIDesCo, nei paragrafi che seguono verranno presentati alcuni dati sulle proprietà del lessico della risorsa. Come termine di paragone verrà utilizzato CORIS (*CORpus di Italiano Scritto,* Rossini Favretti, Tamburini & De Santis 2002), corpus generale di riferimento per l'italiano scritto.[14]

4.1 Lista di frequenza

Le specificità del lessico di CHerIDesCo possono essere apprezzate, in prima battuta, a partire dall'analisi delle liste di frequenza estratte dal corpus. Come si può osservare nella Tabella 2, in cui sono riportate le liste delle venti parole "piene" (token e lemmi) più frequenti, il dominio artistico-culturale è dominante. Tale peculiarità appare ancora più evidente se la lista viene posta a confronto con un'estrazione simile condotta su un corpus generalista come CORIS (Tabella 3).

[14] I dati qui riportati di riferiscono alla versione del corpus CORIS pubblicata nel 2017 (150Mw, aggiornata al 2016), consultabile all'URL: http://corpora.dslo.unibo.it/coris_ita.html

Tabella 2. Lista di frequenza del corpus CHerIDesCo: primi venti token (a sinistra) e lemmi (a destra)

rank	token	f. assoluta	f. relativa	lemma	f. assoluta	f. relativa
1	secolo	907	0,00306	secolo	1007	0,00339
2	più	845	0,00285	archeologico	846	0,00285
3	museo	628	0,00212	più	845	0,00285
4	area	585	0,00197	museo	720	0,00243
5	età	568	0,00191	antico	654	0,00220
6	città	544	0,00183	area	648	0,00218
7	anche	517	0,00174	età	568	0,00191
8	parte	502	0,00169	primo	559	0,00188
9	a.c.	471	0,00159	edificio	548	0,00185
10	archeologico	438	0,00148	parte	544	0,00183
11	edificio	396	0,00133	città	544	0,00183
12	piano	390	0,00131	romano	542	0,00183
13	complesso	350	0,00118	anche	537	0,00181
14	romana	337	0,00114	grande	509	0,00171
15	antica	320	0,00108	piano	473	0,00159
16	interno	318	0,00107	a.c.	471	0,00159
17	grande	314	0,00106	sala	437	0,00147
18	chiesa	306	0,00103	complesso	418	0,00141
19	fino	298	0,00100	interno	401	0,00135
20	sala	288	0,00097	opera	384	0,00129

Tabella 3. Lista di frequenza del corpus CORIS: primi venti parole “piene” (token)

rank	token	f. assoluta	f. relativa
1	essere	3462454	0,02297
2	avere	1526082	0,01012
3	fare	548788	0,00364
4	suo	547717	0,00363
5	quello	537476	0,00356
6	questo	529291	0,00351
7	potere	487273	0,00323
8	tutto	424861	0,00281
9	altro	351502	0,00233
10	dire	327375	0,00217
11	dovere	277821	0,00184
12	anno	274738	0,00182
13	solo	225860	0,00149
14	due	199188	0,00132
15	andare	196213	0,00130
16	venire	193793	0,00128
17	primo	189128	0,00125
18	molto	181296	0,00120
19	stare	172761	0,00114
20	volere	170072	0,00112

Sotto tale profilo il corpus appare omogeneo anche nelle sue articolazioni interne. A questo proposito, si riportano le prime dieci parole contenuto più frequenti in ciascun subcorpus (Tabella 4).

Tabella 4. CHerIDesCo, le dieci parole contenuto più frequenti nei subcorpora (in parentesi sono riportate le frequenze assolute)

Subcorpus	**Lemmi più frequenti**
Abruzzo_subcorpus	('chiesa', 82), ('più', 71), ('secolo', 68), ('primo', 58), ('anche', 53), ('archeologico', 48), ('antico', 46), ('museo', 43), ('parte', 42), ('anno', 41);
Basilicata_subcorpus	('tela', 39), ('olio', 37), ('secolo', 36), ('cm', 35), ('archeologico', 34), ('x', 31), ('a.c.', 27), ('area', 24), ('romano', 20), ('museo', 19);
Calabria_subcorpus	('secolo', 96), ('archeologico', 76), ('area', 69), ('a.c.', 64), ('più', 63), ('museo', 60), ('età', 56), ('provenire', 52), ('città', 44), ('romano', 44);
Campania_subcorpus	('archeologico', 153), ('secolo', 147), ('età', 142), ('antico', 135), ('edificio', 130), ('più', 121), ('area', 116), ('parte', 111), ('città', 111), ('sala', 105);
EmiliaRomagna_subcorpus	('secolo', 51), ('più', 50), ('antico', 45), ('museo', 41), ('sala', 38), ('primo', 37), ('piano', 37), ('chiesa', 35), ('castello', 34), ('parte', 32);
FriuliVeneziaGiulia_subcorpus	('secolo', 24), ('museo', 24), ('romano', 23), ('più', 20), ('archeologico', 18), ('anche', 18), ('città', 14), ('edificio', 14), ('fase', 13), ('struttura', 13);
Lazio_subcorpus	('secolo', 215), ('più', 188), ('antico', 165), ('archeologico', 165), ('area', 126), ('museo', 115), ('grande', 113), ('primo', 113), ('anche', 109), ('parte', 108);
Liguria_subcorpus	('secolo', 23), ('museo', 20), ('palazzo', 20), ('archeologico', 13), ('galleria', 12), ('città', 11), ('più', 11), ('fortezza', 11), ('edificio', 10), ('anche', 10);
Lombardia_subcorpus	('più', 36), ('età', 35), ('secolo', 33), ('archeologico', 29), ('romano', 28), ('museo', 28), ('grande', 26), ('area', 21), ('pubblico', 20), ('parco', 20);
Marche_subcorpus	('romano', 54), ('archeologico', 49), ('area', 45), ('età', 36), ('a.c.', 32), ('città', 27), ('sec.', 26), ('resto', 23), ('piano', 23), ('antico', 22);
Molise_subcorpus	('castello', 49), ('sala', 40), ('secolo', 35), ('piano', 28), ('età', 27), ('cavallo', 27), ('anche', 26), ('primo', 23), ('conservare', 22), ('museo', 21);
Piemonte_subcorpus	('secolo', 38), ('castello', 36), ('giardino', 31), ('grande', 24), ('museo', 19), ('area', 18), ('archeologico', 18), ('antico', 18), ('architetto', 17), ('intervento', 17);
Puglia_subcorpus	('secolo', 57), ('archeologico', 48), ('museo', 29), ('più', 29), ('città', 27), ('anche', 27), ('a.c.', 26), ('piano', 25), ('età', 23), ('edificio', 22);
Sardegna_subcorpus	('più', 76), ('tomba', 69), ('secolo', 60), ('area', 57), ('archeologico', 57), ('struttura', 52), ('costituire', 48), ('villaggio', 47), ('nuragico', 47), ('a.c.', 47);
Toscana_subcorpus	('museo', 131), ('più', 75), ('opera', 71), ('collezione', 69), ('piano', 65), ('primo', 60), ('parte', 60), ('secolo', 59), ('giardino', 58), ('edificio', 57);
Umbria_subcorpus	('secolo', 37), ('romano', 26), ('teatro', 23), ('museo', 21), ('più', 20), ('archeologico', 20), ('città', 19), ('necropoli', 17), ('reperto', 16), ('edificio', 15);
Veneto_subcorpus	('museo', 33), ('archeologico', 28), ('più', 24), ('nazionale', 20), ('piano', 19), ('percorso', 17), ('area', 15), ('antico', 15), ('città', 14), ('sala', 14);

4.2 Incidenza delle parti del discorso

La natura descrittiva dei testi che compongono la risorsa ha un impatto sulla distribuzione delle parti del discorso. Come si può apprezzare osservando i dati presentati in Tabella 5 e Figura 4, nel corpus CHerIDesCo i sostantivi hanno un'incidenza più alta se confrontati con il corpus di riferimento:[15] costituiscono infatti il 27,42% del totale (di cui 20,88% nomi comuni e 6,54% nomi propri) rispetto al 24,45% (19,13% + 5,32%) del corpus CORIS. Al contrario, la classe lessicale dei verbi ha minor rilevanza dal punto di vista quantitativo (6,5% vs. 9,04%).
Tali dati sono perfettamente in linea con le caratteristiche lessicali dei testi descrittivi illustrate dalla letteratura scientifica.

Tabella 5. UPOS

CHerIDesCo			CORIS		
UPOS	f. assoluta	f. relativa	UPOS	f. assoluta	f. relativa
NOUN	61993	0,20882	NOUN	28834111	0,19129
ADP	52992	0,17850	PUNCT	24328824	0,16140
DET	48125	0,16211	ADP	21643534	0,14359
PUNCT	31037	0,10455	ADJ	14087404	0,09346
ADJ	28718	0,09674	VERB	13631027	0,09043
PROPN	19402	0,06535	DET	11863779	0,07871
VERB	19284	0,06496	PROPN	8022634	0,05322
CCONJ	8498	0,02863	ADV	6990871	0,04638
ADV	7320	0,02466	PRON	6477549	0,04297
PRON	6746	0,02272	CCONJ	5433133	0,03604
AUX	6133	0,02066	AUX	5068773	0,03363
NUM	5815	0,01959	NUM	2077842	0,01379
SCONJ	669	0,00225	SCONJ	1857440	0,01232
X	96	0,00032	X	371386	0,00246
SYM	23	0,00008	INTJ	43775	0,00029
INTJ	20	0,00007			

[15] Per assicurare un confronto tra le due risorse il *tag-set* del corpus CORIS, basato sulle linee guida EAGLES (http://corpora.dslo.unibo.it/TCORIS/EAGLES-like_POSTagset.pdf), è stato convertito nelle classi previste da UPOS.

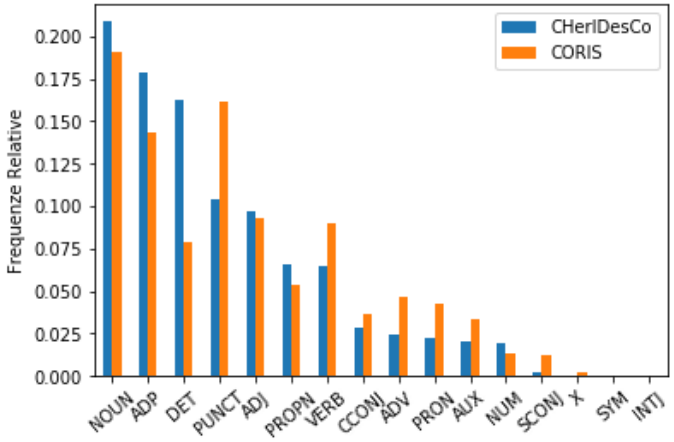

Figura 4. Incidenza relativa delle Parti del Discorso (UPOS) nei due corpora

5. Note conclusive

CHerIDesCo è un corpus specialistico monolingue italiano creato e pensato per le possibili applicazioni del NLP, sia *corpus-driven* che *corpus-based* (Tognini-Bonelli 2001), nell'ambito specifico dei BBCC. In particolare, per quanto riguarda la prima categoria di applicazioni, CHerIDesCo offre la possibilità, analogamente ad altri lavori affini (si veda ad esempio Billero & Nicolás Martinez 2017; Aresti & Lanini 2020), di costruire lemmari e dizionari. È inoltre possibile utilizzare il corpus, grazie alla peculiarità della sua ricchezza lessicale e in considerazione soprattutto della diversità delle istituzioni culturali rappresentate, per condurre alcuni tra i *task* più comuni nel NLP e nelle *Digital Humanities* come, ad esempio, il *Term Extraction* (TE) ed il *Named Entity Recognition* (NER) (van Hooland *et al*. 2015). In quest'ultimo senso è senz'altro auspicabile il ricorso a tecniche di collegamento delle Entità Nominate (NEL – *Named Entity Linking*) a determinate basi di conoscenza (KB – *Knowledge Base*) preesistenti (es. DBpedia, Lehmann *et al.* 2015), al fine soprattutto di provvedere alla disambiguazione delle entità medesime tramite il conferimento di identità univoche. Più in generale, un siffatto impiego del corpus può rivelarsi particolarmente utile nella creazione automatica di appositi glossari specialistici[16], per i

[16] Chiaramente, il contributo di CherIDESco può esser volto anche all'integrazione di glossari specialistici esistenti.

quali le risorse linguistiche disponibili in italiano sono spesso non sufficienti, come nel caso specifico del dominio dei beni archeologici, o addirittura assenti.

Infine, si segnala il potenziale utilizzo del corpus ai fini dell'analisi e del *modeling* dei *topic* (Blei, Ng & Jordan 2003). Tale strumento si rivela particolarmente prezioso proprio in relazione alle caratteristiche essenziali di CHerIDesCo: un corpus di descrizioni testuali delle diverse istituzioni culturali statali italiane (musei, parchi archeologici, pinacoteche, ecc.). L'analisi dei *topic* consentirebbe infatti, se applicata al dominio in oggetto, di analizzare l'offerta culturale pubblica a partire dal grado di differenziazione, qualora esistente, dei diversi contesti istituzionali dominanti, offrendo anche la possibilità di scorgere innovative configurazioni degli stessi sulla base di caratteristiche salienti comuni e di tratti distintivi irriducibili: a titolo meramente esemplificativo, la distinzione tra istituzioni a carattere prevalentemente archeologico e quelle deputate alla conservazione di manufatti del patrimonio storico-artistico nazionale come musei, pinacoteche e biblioteche.

Ringraziamenti

Il progetto SMACH – *Semantic Multilingual Access to Cultural Heritage* è finanziato nell'ambito Strategia Nazionale di Specializzazione Intelligente (SNSI) mediante il Programma Operativo Nazionale "Ricerca e Innovazione" 2014-2020 (Asse I *Capitale Umano*, Azione I.2 *Attrazione e Mobilità dei Ricercatori*). Gli autori ringraziano Fabio Tamburini per i dati del corpus CORIS e l'Unior NLP group (in particolare la referente scientifica Johanna Monti) per il prezioso supporto. Sono inoltre grati ai due anonimi revisori per i loro preziosi commenti sul contenuto di una precedente versione del lavoro.

Riferimenti bibliografici

Aloia, N., Concordia, C. & Meghini, C. 2011. Europeana v1.0. In M. Agosti, F. Esposito, C. Meghini & N. Orio (eds), *Digital Libraries and Archives. 7th Italian Research Conference, IRCDL 2011, Pisa, Italy, January 20-21, 2011. Revised Papers* (Communications in Computer and Information Science, Vol. 249). Berlin-Heidelberg: Springer-Verlag, 127-129.

Aresti, A. & Lanini, L. 2020. *Corpus LBC Italiano*. Firenze: Firenze University Press.

Baroni, M. & Ueyama, M. 2006. Building general- and special-purpose corpora by Web crawling. In *Proceedings of the 13th NIJL international symposium, language corpora: Their compilation and application*, 31-40.

Bertinetto, P.M. & Ossola, C. 1982. *Insegnare stanca. Esercizi e proposte per l'insegnamento dell'italiano*. Bologna: il Mulino.

Billero, R., & Nicolás Martínez, M.C. 2017. Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. *CHIMERA Romance Corpora and Linguistic Studies* 4(2): 203-216.

Blei, D.M., Ng, A.Y. & Jordan, M.I. 2003. Latent dirichlet allocation. *The Journal of machine Learning research* 3: 993-1022.

Chiarcos, C., McCrae, J., Cimiano, P. & Fellbaum, C. 2013.Towards open data for linguistics: Linguistic Linked Data. In A. Oltramari, P. Vossen, L. Qin & E. Hovy (eds), *New Trends of Research in Ontologies and Lexical Resources*. *Ideas, Projects, Systems*. Heidelberg-New York-Dordrecht-London: Springer, 7-25.

Chiarcos, C., Nordhoff, S. & Hellmann, S. (eds) 2012. *Linked Data in Linguistics*. *Representing and Connecting Language Data and Language Metadata*. Heidelberg-New York-Dordrecht-London: Springer.

de Marneffe, M., Dozat, T., Silveira, N., Haverinen, K., Ginter, F., Nivre, J., & Manning, CD. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds), *Proceedings of LREC 2014: 9th International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), 4585–4592.

de Marneffe, M., MacCartney, B. & Manning C.D. 2006. Generating typed dependency parses from phrase structure parses. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (eds), *Proceedings of LREC 2006: 5th International conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), 449-454.

de Marneffe, M. & Manning, C.D. 2008. The Stanford typed dependencies representation. In J. Bos, E. Briscoe, A. Cahill, J. Carroll, S. Clark, A. Copestake, D. Flickinger, J. van Genabith, J. Hockenmaier, A. Joshi, R. Kaplan, T. Holloway King, S. Kuebler, D. Lin, J. Tore Lønning, C. Manning, Y. Miyao, J. Nivre, S. Oepen, K. Sagae, N. Xue & Y. Zhang (eds), *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation*. Coling 2008 Organizing Committee, 1-8.

Doerr, M. 2003. The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata. *AI Magazine* 24(3):75-92.

Doerr, M. 2009. Ontologies for Cultural Heritage. In S. Staab & R. Studer (eds.) *Handbook on Ontologies*. Berlin-Heidelberg: Springer.

Doerr, M. & Stead, S. 2011. Harmonized models for the Digital World: CIDOC CRM, FRBROO, CRMDig and Europeana EDM. In *Tutorial. 15th International Conference on Theory and Practice of Digital Libraries – TPDL*. Abstract.

Ehrmann, M., Turchi, M. & Steinberger, R. 2011. Building a multilingual Named Entity Annotated corpus using annotation projection. In R. Mitkov & G. Angelova (eds), *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Stroudsburg (PA): Association for Computational Linguistics, 118-124.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography* 1: 7-36.

Kilgarriff, A., Rychlý, P., Smrz, P. & Tugwell, D. 2004. The Sketch Engine. In G. Williams & S. Vessier (ed), *Proceedings of the 11th EURALEX International Congress*. Lorient, (France): Université de Bretagne-Sud, Faculté des lettres et des sciences humaines, 105-115.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P.N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S. & Bizer, C. 2015. DBpedia – A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web* 6(2): 167-195.

Loos, E.E., Anderson, S., Day, D.H., Jordan, P.C. & Wingate, J.D. 2003. *Glossary of linguistic terms*, SIL International.

Manzotti, E. 2009. La descrizione. Un profilo linguistico e concettuale. *Nuova secondaria* 4: 19-40.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O. Lenci, A., Zampolli, A., Fanciulli, F., Massetani, M., Raffaelli, R., Basili, R., Pazienza, M.T., Saracino, D., Zanzotto, F., Mana, N., Pianesi, F. & Delmonte, R. 2003. Building and the Italian Syntactic–Semantic Treebank. In A. Abeillé (ed), *Treebanks. Building and Using Parsed Corpora*. Dordrecht: Kluwer, 189-210.

Montemagni, S. & Simi, M. 2007. *The Italian dependency annotated corpus developed for the CoNLL–2007 Shared Task*. Technical report. Pisa: ILC–CNR.

Mortara Garavelli, B. 1988. *Manuale di retorica*. Milano: Bompiani.

Padó, S. & Lapata, M. 2009. Cross-lingual Annotation Projection of Semantic Roles. *Journal of Artificial Intelligence Research* 36:307-340.

Petrov, S., Das, D. & McDonald, R. 2012. *A universal part-of-speech tagset*. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds.), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), 2089–2096.

Qi, P., Dozat, T., Zhang, Y. & Manning, C.D. 2018. Universal Dependency Parsing from Scratch. In D. Zeman & J. Hajič (eds), *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Stroudsburg (PA): Association for Computational Linguistics, 160-170.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning C.D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In D. Jurafsky (ed), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg (PA): Association for Computational Linguistics, 101-108.

Roggia, C.E. 2011. Testi descrittivi. In: R. Simone (ed), *Enciclopedia dell'Italiano*. Roma: Istituto dell'Enciclopedia Italiana Treccani.

Rossini Favretti, R., Tamburini F., De Santis, C. 2002. CORIS/CODIS: A corpus of written Italian based on a defined and a dynamic model. In: A. Wilson, P. Rayson & T. McEnery (eds), *A Rainbow of Corpora: Corpus Linguistics and the Languages of the World*. Munich: Lincom-Europa, 27-38.

Spohr, D., Hollink, L. & Cimiano, P. 2011. A Machine Learning Approach to Multilingual and Cross-Lingual Ontology Matching. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy & E. Blomqvist (eds), *The Semantic Web – ISWC 2011* (Lecture Notes in Computer Science, vol 7031). Berlin-Heidelberg: Springer, 665-680.

Tiedemann, J. 2014. Rediscovering Annotation Projection for Cross-Lingual Parser Induction. In J. Tsujii & J. Hajic (eds), *Proceedings of COLING 2014, the 25th International*

Conference on Computational Linguistics: Technical Papers. Dublin City University & Association for Computational Linguistics, 1854-1864.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing.

UNESCO 1972. *Convention concerning the protection of the world cultural and natural heritage*. https://whc.unesco.org/en/conventiontext/ (ultimo accesso: 12 febbraio 2021).

van Hooland, S., De Wilde, M., Verborgh, R., Steiner T. & Van de Walle, R., 2015. Exploring entity recognition and disambiguation for cultural heritage collections. *Digital Scholarship in the Humanities* 30(2): 262-279.

Withers, P. 2012. Metadata management with Arbil. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (eds), *Proceedings of LREC 2012: 8th International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), 72-75.

Zeman, D. 2008. Reusable Tagset Conversion Using Tagset Drivers. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (eds), *Proceedings of LREC 2008: 6th International Conference on Language Resources and Evaluation*. Paris: European Language Resources Association (ELRA), 213-218.