European Language Resource Coordination
Connecting Europe Facility

ELRC Workshop in Italia

ELRC in Italia

Simonetta Montemagni

Istituto di Linguistica Computazionale "A. Zampolli" (CNR)
Referente nazionale sul versante tecnologico di ELRC
(Technological Italian National Anchor)

ELRC Italia: chi siamo

Simonetta Montemagni

- Istituto di Linguistica Computazionale «A. Zampolli» - Consiglio Nazionale delle Ricerche
- Punto di riferimento nazionale sul versante tecnologico (Technological NAP)

Claudia Foti

- Ministero della Giustizia
- Punto di riferimento nazionale sul versante dei servizi pubblici (Public Services NAP)

European Language Resource Coordination
Connecting Europe Facility

ELRC-Italia@ILC-CNR Team

- Paola Baroni

- Sebastiana Cucurullo

- Claudia Soria

ELRC Italia: cosa cerchiamo

Tecnologie della lingua:

Come funzionano

- Paradigma dominante oggi rappresentato da sistemi basati su algoritmi di apprendimento automatico
  - algoritmo
  - **dati e risorse linguistiche**

ELRC Italia: cosa cerchiamo

Tecnologie della lingua

Come funzionano

- Paradigma dominante oggi rappresentato da sistemi basati su algoritmi di apprendimento automatico
  - algoritmo
  - **dati e risorse linguistiche**

- Per trattare nuovi domini e/o varietà d'uso della lingua
  - è necessario integrare l'evidenza su cui si basa il sistema con dati e risorse relativi al nuovo dominio / varietà di lingua

ELRC Italia: cosa cerchiamo

Tecnologie della lingua

Come funzionano

- Paradigma dominante oggi rappresentato da sistemi basati su algoritmi di apprendimento automatico
  - algoritmo
  - **dati e risorse linguistiche**

- Per trattare nuovi domini e/o varietà d'uso della lingua
  - è necessario integrare l'evidenza su cui si basa il sistema con dati e risorse relativi al nuovo dominio / varietà di lingua

ELRC 1 e 2: fasi

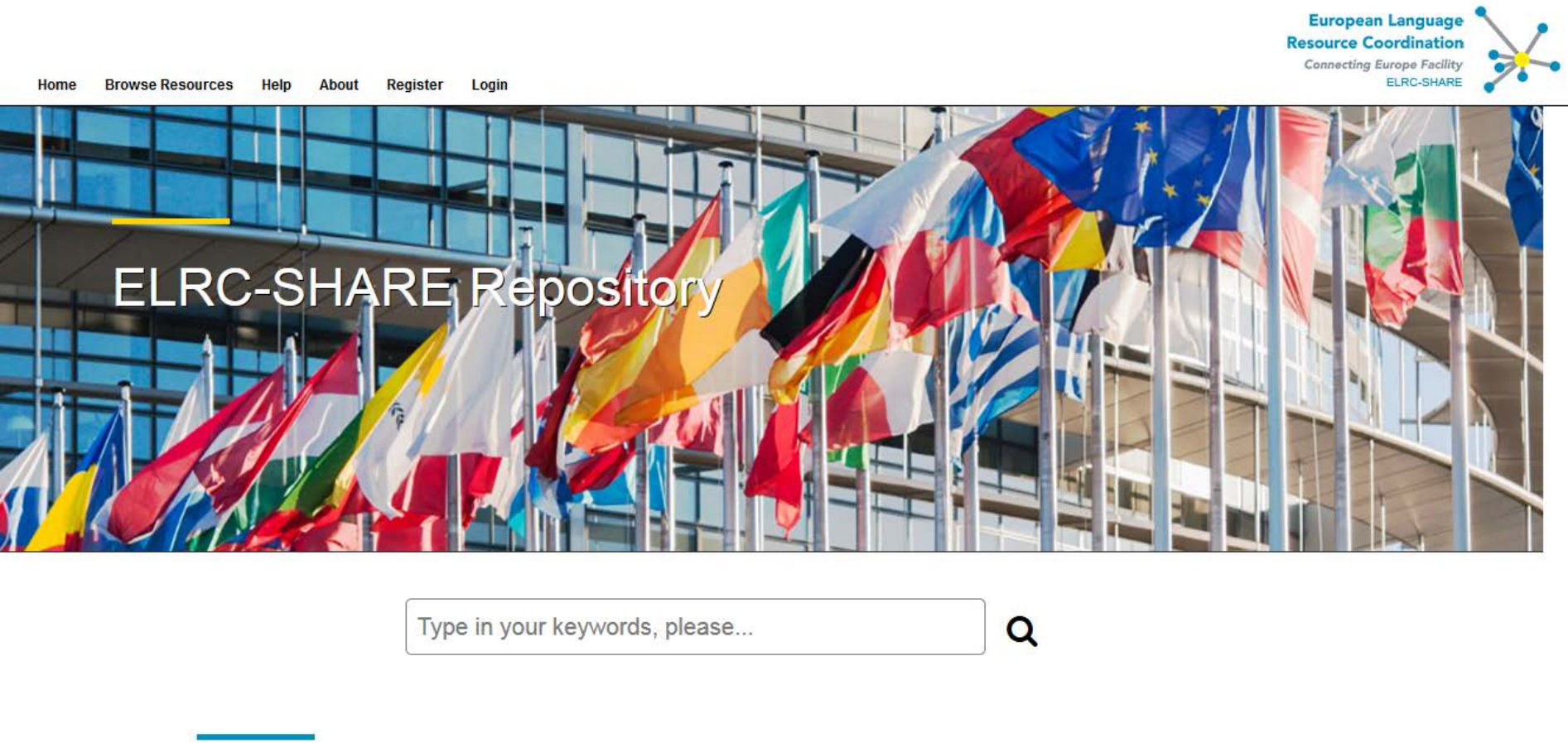
ELRC 1 (2015-2017)

- Identificazione potenziali fornitori di dati (*stakeholders*)
- Primo Workshop ELRC in Italia (15 marzo 2016)
- Questionario sulle risorse linguistiche esistenti per la specializzazione del sistema di traduzione automatica della Commissione Europea
- Selezione, raccolta e formattazione delle risorse linguistiche scelte

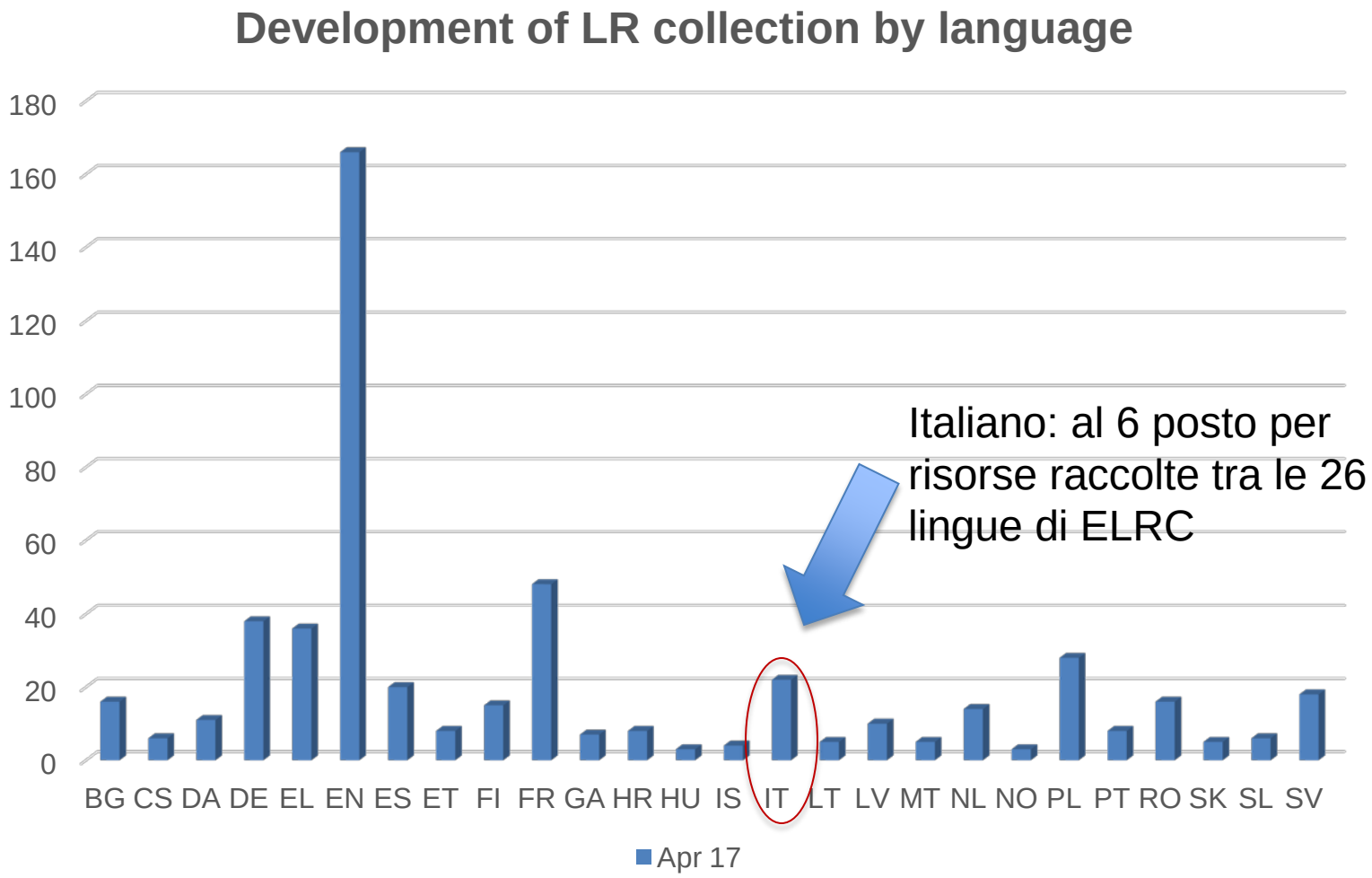
ELRC 2 (2017-2019)

- Estensione della rete di *stakeholders*
- Secondo Workshop ELRC in Italia (27 settembre 2018)
- Identificazione, selezione e raccolta di ulteriori risorse linguistiche, con particolare attenzione a servizi pubblici digitali

ELRC Italia: cosa abbiamo raccolto

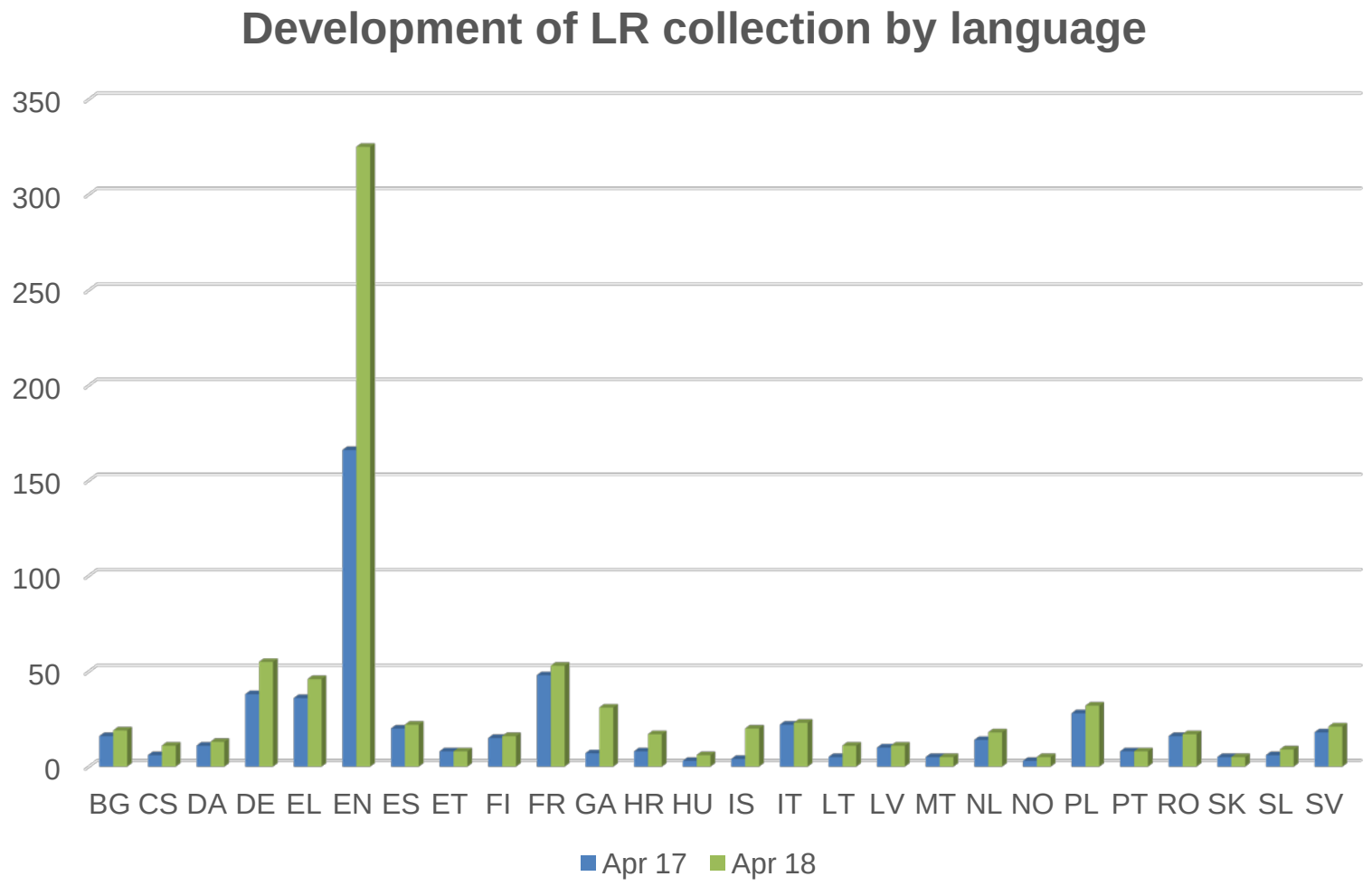


ELRC Data: Progress by language

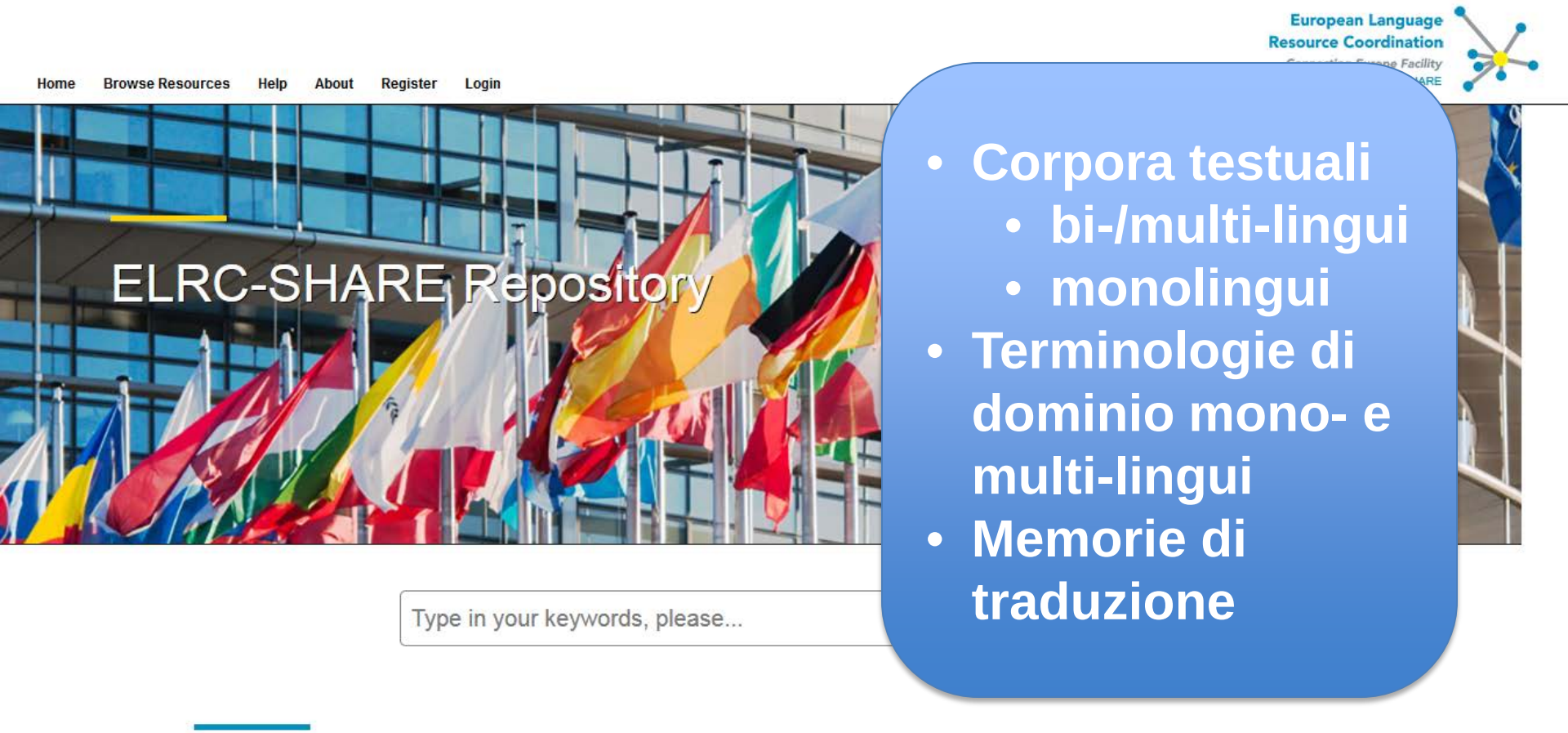


European Language Resource Coordination
Connecting Europe Facility

ELRC Data: Progress by language



ELRC Italia: cosa abbiamo raccolto



- Corpora testuali
  - bi-/multi-lingui
  - monolingui
- Terminologie di dominio mono- e multi-lingui
- Memorie di traduzione

ELRC Italia: cosa abbiamo raccolto
Parametri di variazione

- TIPO DI RISORSA
  - corpus testuale, lessico, memoria di traduzione, terminologia
- TIPOLOGIA DI TESTO
  - leggi, atti amministrativi, manuali, siti web, ecc.
- LINGUA/E
  - Monolingue, bilingue o multilingue
- ALLINEATO/NO
  - testi con traduzione: la traduzione è allineata, per es. a livello di frase, paragrafo, ecc.
- ANNOTATO LINGUISTICAMENTE/NO
  - arricchito con annotazioni di tipo linguistico e/o strutturale
- FORMATO
  - Formato di rappresentazione della risorsa, ad es. .xml, .doc, .txt, .pdf, tmx, ecc.
- DIMENSIONI
  - dimensioni della risorsa, ad es. numero di parole/caratteri;, translation units, kB/Mb, ecc.
- LICENZA D'USO

Italian

Search

Filter by:

24 Language Resources (Page 1 of 2)

Europea
Resource C
Connecting

Help About Register Login

- German (15)
- French (14)
- Spanish; Castilian (10)
- Polish (5)
- Finnish (4)
- Modern Greek (1453-) (4)
- Swedish (4)
- Hungarian (3)
- Romanian; Moldavian; Moldovan (3)
- Bulgarian (2)
- Czech (2)
- Dutch; Flemish (2)
- Latvian (2)
- Croatian (1)
- Danish (1)
- Estonian (1)
- Lithuanian (1)
- Portuguese (1)

Audioguide for the Military History Museum in Vienna 0 137

German | Italian

Open Under-PSI

Audioguide for the Military History Museum in Vienna (Processed) 0 26

German | Italian

Open Under-PSI

Austrian Armed Forces Military Dictionaries 5 70

English | French | German | Hungarian | Italian

Open Under-PSI

CATEX (German-Italian parallel corpus of legal and administrative texts) 0 49

German | Italian

Non-standard/ Other Licence/ Terms

Informazioni associate a ogni risorsa in ELRC-SHARE

CATEX (German-Italian parallel corpus of legal and administrative texts)

Italian-German legal body established under the project
CATEX - Computer Assisted Terminology Extraction (hereinafter "CATEX");
CATEX is an Italian-German parallel corpus of ca. 5 million words that consists of the following Italian legal texts:

- Read issued by the Autonomous Province of Bolzano - South Tyrol;
- Civil Code;
- Read complementary to the Code Civil ice;
- Code of Civil Procedure;
- Code of Criminal Procedure;
- Order of Notaries;
- Cod ice of Failure and other bankruptcy procedures;
- Consolidated Income Tax Act;
- Administrative process;

mail: marcello.soffritti@unibo.it
type file: Corpus Encoding Standard (CES)
size: 2573095 Tokens. Read Less

Back

Distribution

Availability: Available

Licences

Non-standard/ Other Licence/ Terms

Distribution Details

Contact Person

Simonetta Montemagni

text

Bilingual text corpus

Languages

Italian (it)

German (de)

Linguality

Linguality type: Bilingual

Text Format

XML

Size

2,573,095 Tokens

Character encoding

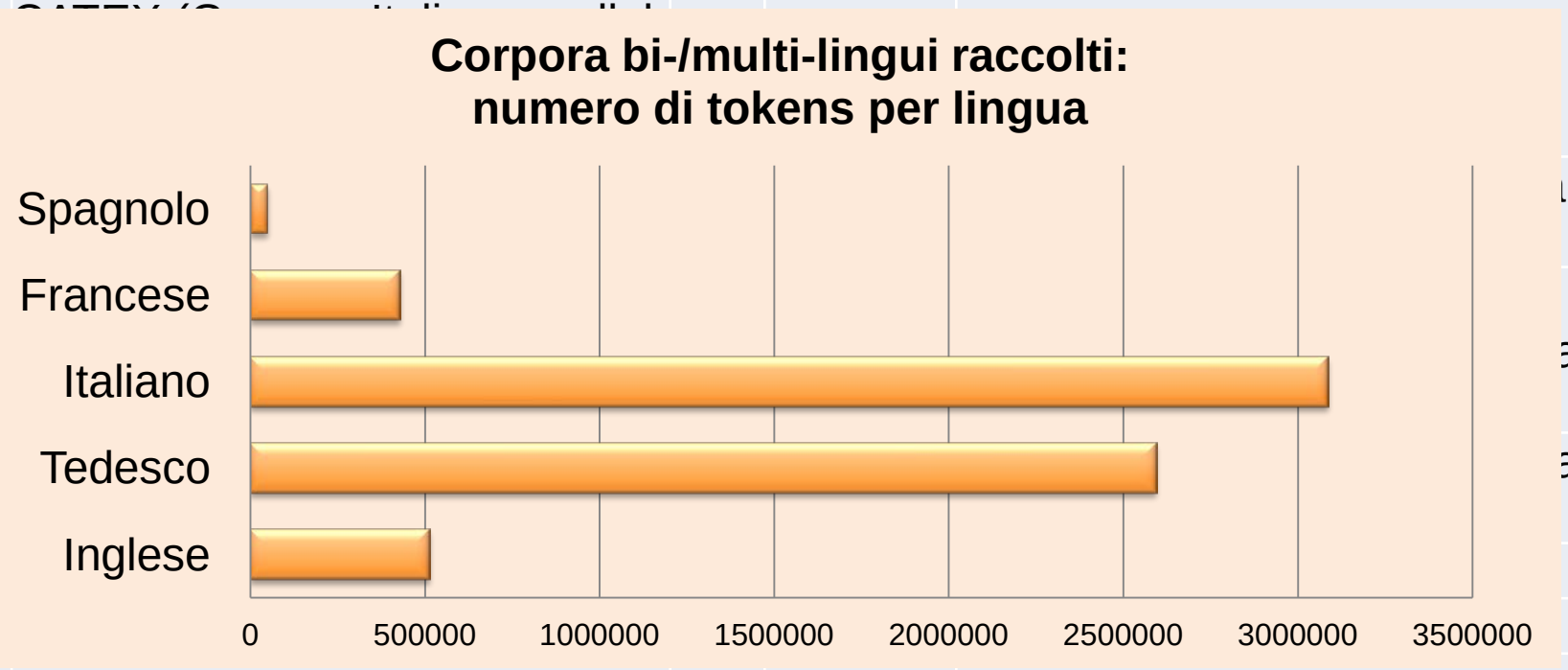
UTF-8

ELRC Italia: corpora testuali raccolti

269	Corpora of legal text	XML	Biling	English | Italian
333	CATEX (German-Italian parallel corpus of legal and administrative texts)	XML	Biling	German | Italian
265	PAeSI : Public Administration and Foreign Immigrants	XML	Multiling	English | French | Italian | Spanish; Castilian
266	CHARTER OF VALUES OF CITIZENSHIP AND INTEGRATION	XML	Multiling	English | French | German | Italian | Spanish; Castilian
267	INFORMATION FOR VICTIMS OF A CRIME	XML	Multiling	English | French | German | Italian | Spanish; Castilian
268	Corpus EPTIC	XML	Multiling	English | French | Italian
270	Legal Texts	XML	Multiling	English | French | Italian
292	PaWaC - Public Administration Web as Corpus	TXT	Monoling	Italian

ELRC Italia: corpora testuali raccolti

269	Corpora of legal text	XML	Biling	English | Italian
333				
265				nish;
266				alian |
267				alian |
268				
270				
292	PaWaC - Public Administration Web as Corpus	TXT	Monoling	Italian

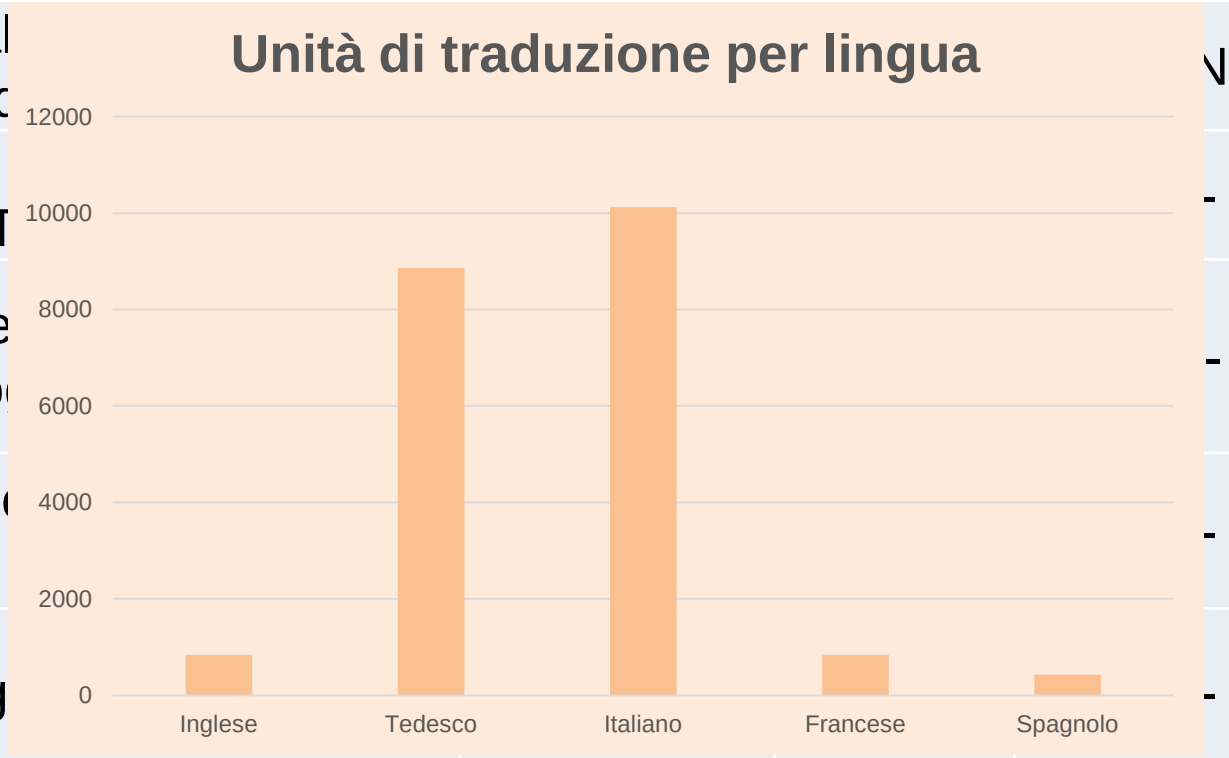


ELRC Italia: risorse terminologiche e memorie di traduzione raccolte

TYPE/ TOPIC	LINGUALITY	FORMAT	LANG. (ISO 639-3)	Entries
Legal terminology on children protection	multilingual	TMX	ITA-ENG-FRA	837
EU Justice Court Terminology	multilingual	TMX	ITA - DEU	514
Terminology of demographic services	multilingual	TMX	ITA - SPA	539
Terminology of cadastral services	multilingual	TMX	ITA - DEU	656
Legal terminology	multilingual	TMX	ITA - DEU	7692
			Total	10238

ELRC Italia: risorse terminologiche e memorie di traduzione raccolte

TYPE/ TOPIC	LINGUALITY	FORMAT	LANG. (ISO 639-3)	Entries
Legal chilc			NG-FRA	837
EU T			- DEU	514
Te demo			- SPA	539
Termin			- DEU	656
Leg			- DEU	7692
			Total	10238



European Language Resource Coordination
Connecting Europe Facility

ELRC SHARE: altre risorse con italiano

Letter of rights for persons arrested on the basis of a European Arrest Warrant	Corpus	PDF	Multilingual	Bulgarian | Dutch; Flemish | English | French | German | Italian | Latvian | Modern Greek (1453-) | Polish | Romanian; Moldavian; Moldovan	2713	Words
Austrian Armed Forces Military Dictionaries	Lexical conceptual resource	MS-Excel xlsx | MS-Word docx | PDF	Multilingual	French | English | Italian | Hungarian | German	27000	Entries
Cyprus at a glance	Corpus	PDF	Multilingual	English | French | Italian | Modern Greek (1453-) | Spanish; Castilian	40289	Words
Spanish-Italian website parallel corpus	Corpus	TMX	Bilingual	Italian | Spanish; Castilian	3319	Translation Units
Health Multilingual Terminologies	Lexical conceptual resource	XML	Multilingual	English | French | German | Italian | Spanish; Castilian	15908	Terms
Audioguide for the Military History Museum in Vienna	Corpus	MS-Word doc	Bilingual	German | Italian	19111	Words
Parallel texts from Swedish Work environment Authority	Corpus	PDF	Multilingual	Bulgarian | Czech | Hungarian | Italian | Latvian | Lithuanian | English | Estonian | Finnish | French | German | Modern Greek (1453-) | Polish | Romanian; Moldavian; Moldovan | Spanish; Castilian | Swedish	22	Files
Austrian Museum Websites (de-it)	Corpus	TMX	Bilingual	German | Italian	1037	Translation Units
Multilingual Public Procurement Terminology	Lexical conceptual resource	XML	Multilingual	Danish | English | Finnish | French | German | Italian | Modern Greek (1453-) | Polish | Portuguese | Spanish; Castilian | Swedish	1408	Terms
Audioguide for the Military History Museum in Vienna (Processed)	Corpus	TMX	Bilingual	Italian | German	867	Translation Units
Parallel texts from Swedish Work environment Authority (Processed)	Corpus	TMX	Multilingual	Swedish | Spanish; Castilian | Romanian; Moldavian; Moldovan | Polish | Modern Greek (1453-) | Lithuanian | Latvian | Italian | Hungarian | German | French | Finnish | Estonian | English | Czech | Bulgarian	18519	Translation Units
Parallel texts from Swedish Social Security Authority (Processed)	Corpus	TMX	Multilingual	Swedish | Spanish; Castilian | Romanian; Moldavian; Moldovan | Polish | Italian | German | French | Finnish | English | Croatian	9139	Translation Units
Parallel texts from Swedish Social Security Authority	Corpus	MS-Word docx | PDF	Multilingual	Croatian | English | Finnish | French | German | Italian | Polish | Romanian; Moldavian; Moldovan | Spanish; Castilian | Swedish	24	Files
Letter of rights for persons arrested on the basis of a European Arrest Warrant (Processed)	Corpus	TMX	Multilingual	Romanian; Moldavian; Moldovan | Polish | Modern Greek (1453-) | Latvian | Italian | German | French | English | Dutch; Flemish | Bulgarian	854	Translation Units
Cyprus at a glance (Processed)	Corpus	TMX	Multilingual	Spanish; Castilian | Modern Greek (1453-) | Italian | French | English	3733	Translation Units
ParaCrawl Parallel Corpora Release 1.1	Corpus	N/A	Multilingual	Spanish; Castilian | Romanian; Moldavian; Moldovan | Portuguese | Polish | Latvian | Italian | German | French | Finnish | English | Dutch; Flemish | Czech		
EUIPO - IP case law Italian-English	Corpus	N/A	Bilingual	Italian | English		
EUIPO - list of goods and services English, Spanish, French and German	Corpus	N/A	Multilingual	Spanish; Castilian | Italian | German | French | English		
DGT Corpus- Italian to Irish	Corpus	N/A	Bilingual	Italian | Irish		

ELRC Italia: risorse raccolte

- Corpora bi-/multi-lingui: 5.471.657 tokens
- Corpus monolingue annotato linguisticamente: 25.218.385 tokens
- Unità di traduzione: 10.238

- La raccolta continua
  - risorse identificate ma non ancora trasferite
  - **le vostre risorse** …

Stakeholders coinvolti

Istituzioni coinvolte

- Presidenza del Consiglio
- Parlamento
- Ministeri (Interni, Giustizia, …)
- Regioni e Province

Fornitori di dati ad oggi

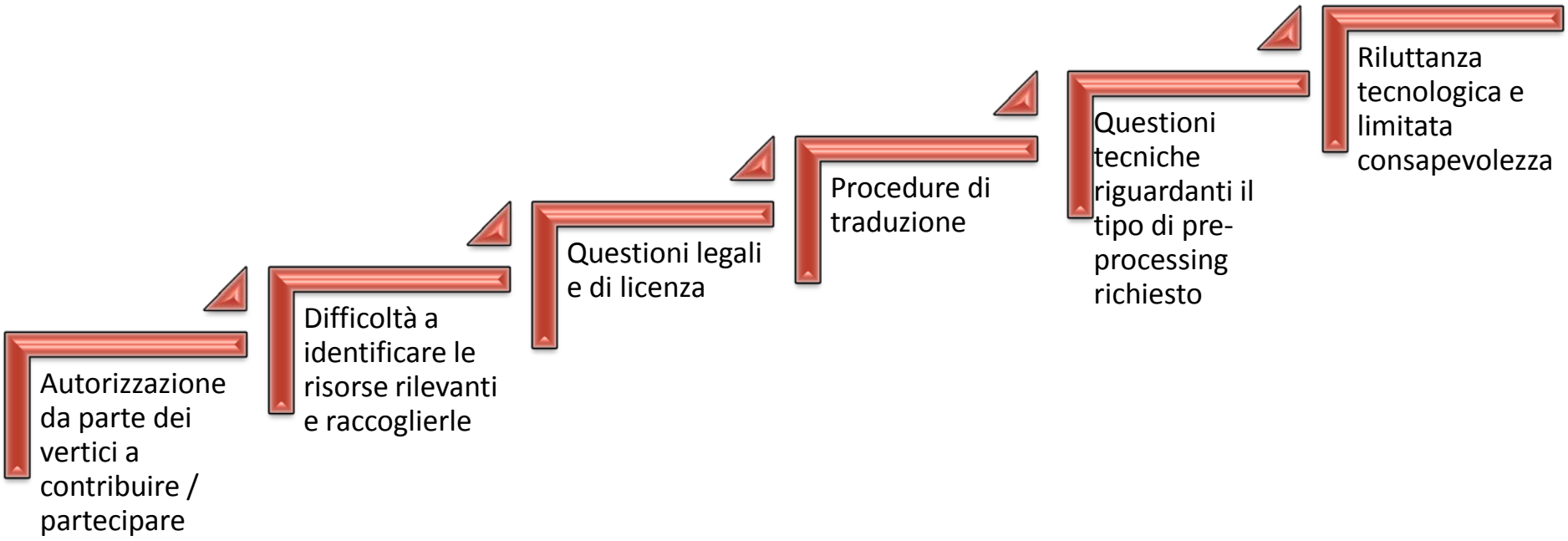
- Ministero degli Interni
- Ministero della Giustizia
- Provincia di Bolzano
- Prefettura di Firenze
- AgID - Agenzia per l'Italia Digitale
- Università di Bologna e di Pisa
- Institute for Specialised Communication and Multilingualism, EURAC Research

Potenziali fornitori di dati

- Tutte le istituzioni da voi rappresentate oggi

Ostacoli incontrati

Ostacoli principali



Risultati principali

- Quantità, qualità e varietà delle risorse raccolte
- Ampia e variegata rete di fornitori di dati, in via di espansione
- Superamento degli ostacoli incontrati

European Language Resource Coordination
Connecting Europe Facility

Grazie per l'attenzione!

Email: info@lr-coordination.eu
Website: www.lr-coordination.eu

ELRC Italia
Email: elrc.italy@ilc.cnr.it