Progetto PRIN 2008:

Portale per l'Accesso alle Risorse Linguistiche per l'Italiano

Cristina Bosco (Unità 1, Università di Torino)

Deliverable D6.1:

Scelta della piattaforma software e progetto dell'organizzazione interna del portale (Documento)

Il presente documento si colloca nell'ambito del Work Package 6 del progetto, che concerne lo sviluppo e la manutenzione del portale.
Partendo da un'attenta analisi dei portali esistenti per il Natural Language Processing (NLP) di altre lingue, l'attività nell'abito del WP6 ha riguardato in modo particolare la strutturazione del portale di accesso alle risorse ed agli strumenti per la lingua italiana e la scelta ed introduzione dei contenuti nel portale stesso.
Il documento è organizzato nelle seguenti sezioni:

1. Analisi dei portali per il NLP (di lingue diverse dall'italiano)

2. Descrizione e motivazioni della struttura del portale e delle scelte software

3. Ricerca ed introduzione delle informazioni nel portale

1. Analisi dei portali per il NLP

È stata condotta una ricerca sul web finalizzata a trovare i portali attualmente esistenti per il NLP. Tra tutti questi portali ne sono stati selezionati 6, in quanto maggiormente sviluppati e ricchi di contenuti. Questi 6 sono diventati oggetto dell'analisi che segue, in cui viene evidenziata in modo schematico la struttura di ognuno dei portali.

Oficina de Espanol en la Sociedad de la Informacion (OESI)
http://oesi.cervantes.es/ (solo in spagnolo)

1. Cosa è OESI:
   - obiettivo
   - funzioni
   - attività

2. Tecnologie linguistiche (descrizione testuale di):
   - tecnologie del parlato
   - tecnologie dello scritto
   - risorse linguistiche
   - ricadute socio-economiche
   - benefici

3. Dati dell'ingegneria linguistica in Spagna:
   - progetti
   - gruppi
   - organizzazioni

   (le 3 voci danno l'accesso ad una maschera di ricerca dove nella query, oltre a nome e data si può scegliere l'area di ricerca e, per le organizzazioni, il tipo)

4. Notizie: poche in evidenza + motore di ricerca

5. Agenda: motore di ricerca per eventi

6. Infoteca: motore di ricerca per pubblicazioni

7. Lista dei progetti a cui OESI partecipa

8. Risorse utili: form per contatti con OESI

French Language Technology Portal
http://www.technolangue.net/ (solo in francese)

1. Tecnologie della lingua**:**
   - Introduzione al dominio (descrizione)
   - Limiti e problemi
   - Cifre chiave (descrizione del mercato delle TL)
   - Studi di casi e visite di imprese

2. Panorama:
   - Attori del dominio (lista di enti, progetti + form per iscrizione)
   - Formazione e mestieri (descrizione tipi di lavoro che si fa nelle TL)
   - Tesi (in corso e disponibili)
   - Risorse e strumenti (accesso a repository esterni di risorse)
   - Iniziative nazionali ed europee (nello sviluppo di altri portali simili)

3. Azione Technolangue:
   - Presentazione
   - Temi
   - Progetti

4. Norme e standard:
   - Posta in gioco
   - Standard

5. Spazio per notizie: (3 per ogni tipo delle seguenti)
   - Attualità
   - Sviluppi
   - Approfondimenti
   - Manifestazioni
   - Ultima ora

Croatian language technologies
http://www.hnk.ffzg.hr/jthj/default_english.htm (in croato e in inglese)

1. Istituzioni ed associazioni
2. Progetti
3. Corpora (divisi per lingua; non treebank (che stanno sotto Strumenti))
4. Dizionari
5. Strumenti
6. Parlato (progetti relativi a)
7. Conferenze
8. Glossario (parallelo in inglese, francese, tedesco, spagnolo, italiano, danese, finnico, croato)
9. Miscellanea (siti rilevanti per TL croate)

Swedish centre for documentation and information on Language Technology (SLATE)
http://sprakteknologi.se/ (in svedese e tedesco)

1. Benvenuto:
   - Definizione di TL
   - Sistemi da provare
   - Definizione del sito
   - Contatti

2. Attori delle TL svedesi:
   - Organizzazioni
   - Enti privati
   - Progetti
   - Persone (link a ELSENET e Linguist List)

3. Risorse:
   - Prodotti
   - Risorse tecniche e sistemi di ricerca e sviluppo
   - Risorse linguistiche

4. Didattica (divisa in pre e post laurea)

5. Notizie (e offerte di lavoro)

6. Documenti:
   - Politici
   - Report generali
   - Archivi di documenti svedesi
   - Archivi internazionali

7. Link:
   - Terminologia
   - TL dei pasei nordici
   - TL internazionali
   - Newsletter
   - Mailing list
   - Liste di conferenze
   - Enti/iniziative per la valutazione
   - Programmi di ricerca e finanziamenti
   - Liste di offerte di lavoro
   - Archivi
   - Riviste

Finnish Language Technology Documentation Centre in Finland (FiLT)
https://kitwiki.csc.fi/twiki/bin/view/FiLT/FiLTWikiEn (in finlandese, svedese e inglese)

1. Informazione e conoscenza:
    - Didattica (link a corsi nei paesi finnici e in quelli del nord Europa)
    - Articoli (selezione di paper su linguaggi finnici)
    - Tecnologie (descrizione statica e link a LTW)
    - Terminologia

2. Attori e gruppi:
    - Organizzazioni
    - Enti privati
    - Progetti
    - Persone

3. Sistemi e risorse:
    - Prodotti
    - Materiali
    - Sistemi di ricerca

4. Comunicazioni ed eventi:
    - Liste di news su TL nei paesei nordici
    - Altre fonti di informazione

5. Lingua Sami (lista di esperti della lingua con loro contatti)

6. Istruzioni per aggiungere dati (possibilità di contribuire con modalità tipi wikipedia (TWiki) sempre controllata via password)

7. Motori di ricerca

8. Lista di progetti internazionali

Language Technology World (LTW)
http://www.lt-world.org/ (in inglese)

1. Motore di ricerca (interno) su TL

2. Informazioni generali:
   - Tecnologia del linguaggio:
     - Definizione
     - Organizzazioni
     - Fonti di informazione
     - Pubblicazioni
   - Su LTW (info sul servizio, comitati e collaboratori, ringraziamenti)
   - Advisory board internazionale (lista dei membri con dettagli)
   - LTW – archivio notizie (organizzate in ordine temporale)

3. Link esterni (a motori di ricerca)

4. Notizie brevi

5. Altre notizie

6. Calls

7. Informazione e conoscenza
   - Fonti di informazione (servizi di informazione, archivi di pubblicazioni, servizi locali di informazione, iniziative di standardizzazione, materiale didattico, fonti rilevanti di informazione)
   - Tecnologie (ogni voce della lista contiene una definizione e una lista: persone, organizzazioni, risorse, eventi, pubblicazioni)
   - Abbreviazioni (lista di acronimi)

8. Attori e gruppi (liste strutturate)
   - Persone
   - Progetti (europei e nazionali)
   - Organizzazione

9. Sistemi e risorse (liste strutturate)
- Sistemi di ricerca e sviluppo
- Prodotti
- Archivi (siti che ospitano risorse e sistemi per TL)

10. Comunicazioni e diritti (IPR)
- Notizie
- Conferenze
- Brevetti (rilasciati dal European Patent Office su TL

2. Descrizione e motivazioni della struttura del portale e delle scelte software

Lo scopo primario del progetto PARLI è di rendere disponibili le risorse linguistiche e gli strumenti per il NLP italiano alla comunità di ricerca. In accordo con tale scopo, si è previsto lo sviluppo di un sito per l'accesso alle risorse ed agli strumenti per il NLP in lingua italiana.
Il sito è stato inoltre pensato come punto di riferimento per raccogliere in un unico luogo virtuale e diffondere le informazioni relative al progetto stesso.

Grazie all'analisi dei portali esistenti per altre lingue (vedi sez. 1 di questo documento), ed alla discussione con il gruppo di progetto, si è deciso di organizzare il sito nelle seguenti sezioni:

- Portale PARLI:
  rappresenta il punto di accesso da siti esterni e contiene informazioni sul portale; è organizzato nelle seguenti sezioni:
  - Risorse, una lista delle risorse esistenti strutturata in base al tipo delle risorse; cliccando sul nome di ognuna delle risorse che compare nella lista viene aperta una scheda che contiene la descrizione della risorsa, gli autori, la licenza, il link al sito web, i contatti a cui rivolgersi per avere la risorsa o informazioni maggiori su di essa
  - Strumenti, una lista degli strumenti esistenti strutturata in base al tipo degli strumenti; analogamente a quanto accade per le risorse (di cui al punto precedente), cliccando sui nomi degli strumenti presenti nella listasi apre una scheda con la descrizione, gli autori, la licenza, il link al sito web, i contatti
  - Link, due liste, di cui una di iniziative relative al NLP italiano, ed una di strumenti multilingui o sviluppati per lingue diverse dall'italiano, ma di cui si conosce l'applicazione a tale lingua, strutturata in base al tipo di strumento; per ogni strumento viene proposto il link ad una o più pubblicazioni in cui lo strumento è stato originariamente presentato, ed il link ad una o più pubblicazioni che descrivono specificamente l'applicazione dello strumento alla lingua italiana
  - Segnalazioni, un form per consentire agli utenti di proporre l'introduzione di nuove risorse/strumenti nel portale e contattare gli organizzatori

- Progetto PARLI:
  rappresenta il punto di accesso alle informazioni sul progetto ed è organizzato nelle seguenti sezioni:
  - Membri, una lista dei proponenti del progetto e di altri ricercatori che hanno offerto la loro collaborazione
  - Attività, una descrizione sintetica delle attività svolte o in corso nell'ambito del progetto

- Obiettivi, una descrizione degli obiettivi che il progetto si propone di raggiungere
- Documenti, link a tutti i documenti prodotti nell'ambito di PARLI, dalla proposta ai deliverable, alle pubblicazioni

Tra le informazioni che il Portale offre rispetto alle risorse ed agli strumenti, si è pianificato di introdurre anche una sorta di etichettatura che ne evidenzi l'interoperabilità. Infatti, tra le principali ricadute del progetto PARLI si auspica lo sviluppo di strumenti e risorse che siano tra loro interoperabili[1]. Questo aspetto non è stato per ora sviluppato per due motivi: da un lato sembra che per ora siano stati dedicati sforzi molto limitati all'integrazione di risorse, dall'altro probabilmente il lavoro fatto non sempre è stato pubblicato. Il monitorare l'interoperabilità delle risorse e strumenti esistenti, ed in parallelo di lavorare su risorse e strumenti che assumano come principio l'interoperabilità o per renderle maggiormente interoperabili, rimane comunque uno degli obiettivi irrinunciabili per il futuro.
Per consentire l'accesso ad informazioni che il portale non riporta esplicitamente sono stati introdotti dei link appositi, in particolare per le informazioni relative a:

- Chi si occupa di NLP in Italia, link a AI*IA (strutturata in organizzazioni, compagnie e progetti)
- Dove si studia NLP in Italia, link a AI*IA (strutturata in pre e post laurea).

Parimenti nel sito dell'AI*IA compare il link al sito PARLI per quanto riguarda le risorse e gli strumenti per il NLP italiano.

Tutte le sezioni del sisto sono state sviluppate sia in italiano sia in inglese e il passaggio alla lingua alternativa rispetto a quella visualizzata in un dato momento dall'utente è possibile tramite un pulsante presente su ogni pagina del portale e del sito del progetto. La documentazione è invece presente in una delle due lingue di riferimento, ma per ogni documento (quando non ovvio) si segnala la lingua in cui è redatto.

Per quanto riguarda le scelte software, il sito è stato sviluppato in linguaggio HTML utilizzando dei CSS che ne definiscono gli aspetti grafici ed alcuni script Java che consentono la visualizzazione completa o compatta delle schede relative a strumenti e risorse, e il passaggio tra i vari tab con cui sono strutturate le diverse sezioni sia del portale sia del sito del progetto.
Anche se inizialmente presi in considerazione, a seguito della discussione con il gruppo, sono stati esclusi strumenti per lo sviluppo collaborativo del sito, stile Wikipedia, e ci si è orientati verso uno sviluppo maggiormente controllato dai responsabili stessi del progetto. L'interazione con gli utenti viene comunque garantita dalla presenza dei form per le Segnalzioni.

[1] Si considerano interoperabili risorse e strumenti che siano tra loro compatibili in termini di formato di annotazione, e/o che mettano a disposizione gli strumenti necessari a renderli compatibili.

Sono stati parimenti esclusi strumenti come Google Sites, che forniscono le utilità necessarie alla costruzione di siti web, a favore di scelte grafiche e funzionali maggiormente autonome.

3. Ricerca ed introduzione delle informazioni nel portale

La popolazione del portale è stata condotta in due passi: creazione delle liste di risorse e strumenti, e compilazione delle schede ad essi relative.
Per quanto riguarda il primo passo, è stata fondamentale la consultazione del sito Evalita. Ognuna delle pubblicazioni fatte in relazione ai vari task della campagna di valutazione per il NLP per l'italiano (e rese disponibili sul sito) contiene infatti la descrizione di almeno uno strumento e una risorsa a cui quest'ultimo è stato applicato durante la campagna stessa. In molti casi il materiale presentato nella pubblicazione va oltre la descrizione di una singola risorsa e strumento, e contiene anche informazioni su esperimenti aggiuntivi con altri strumenti e risorse rispetto a quanto richiesto per partecipare al task a cui la pubblicazione si riferisce. Questo ha consentito di stilare una lista preliminare di risorse e strumenti, successivamente ampliata grazie a ulteriori ricerche sul web.
Anche per quanto riguarda il secondo passo, cioè la compilazione delle schede, il supporto del sito Evalita è stato importante, in quanto in alcuni casi, le informazioni presenti nel sito hanno consentito anche di compilare (parte del)la scheda da introdurre nel portale. Per completare il contenuto di tutte le schede si è condotta una ricerca sul web.

Le risorse elencate e schedate nel portale sono attualmente 22, mentre gli strumenti sono 48 nella sezione Strumenti e 40 nella sezione Link. Questi numeri sono destinati a crescere sia grazie alle segnalazioni degli utenti sia grazie alla raccolta dei dati relativi a Evalita 2011. Allo stato attuale sono stati infatti presi in considerazione tutti i dati delle prime due edizioni di Evalita (2007 e 2009), mentre è in corso la raccolta di dati relativi a Evalita 2011, dato che la campagna di valutazione si è appena conclusa.