Costituzione di un corpus giuridico parallelo italiano-arabo

Fathi Fawi
Dipartimento di Studi linguistici e culturali comparati
Università Ca' Foscari – 30123 Venezia
Email: fathi_fawi@yahoo.com

Abstract

Italiano. I corpora paralleli rappresentano un'importanza assoluta per tante applicazioni della linguistica computazionale, come la traduzione automatica, l'estrazione delle terminologie, o la disambiguazione semantica, ecc. In questo lavoro presentiamo il nostro tentativo di creare un corpus giuridico parallelo italiano-arabo allineato a livello di frase e annotato a livello morfosintattico.

English. *Parallel corpora are an important resource for many applications of computational linguistics, such as machine translation, terminology extraction, semantic disambiguation, etc. In this paper we present our attempt to build an Italian-Arabic parallel corpus in the legal domain, aligned at the sentence level and tagged at the POS level.*

1 Introduzione

Con il crescente sviluppo delle tecnologie informatiche che consentono di raccogliere, gestire ed esplorare enormi quantità di dati linguistici, l'interesse alla creazione di corpora linguistici è cresciuto recentemente in una maniera esponenziale. È indubbio che oggigiorno l'enorme disponibilità dei dati sul web ha agevolato significativamente la costituzione e la distribuzione dei corpora linguistici sia i corpora monolingui che quelli multilingui. In effetti, i corpora costituiscono una risorsa essenziale per il campo linguistico soprattutto per le analisi contrastive tra due o più lingue, per la didattica delle lingue straniere e per gli studi lessicografici e di traduzione. Nell'ambito della linguistica computazionale i corpora linguistici, e in particolare quelli paralleli, acquistano un'importanza assoluta, soprattutto per applicazioni come la traduzione automatica, l'estrazione di terminologie o la disambiguazione semantica.

Tuttavia, non tutte le lingue prendono ugualmente parte a corpora paralleli bilingui o multilingui. In effetti, l'arabo è una lingua che presenta una limitata partecipazione a corpora paralleli, soprattutto a quelli specialistici. È un fenomeno che si può considerare come un possibile effetto della modesta disponibilità sul web di testi paralleli in lingua araba e in altre lingue, nonché della complessità del sistema morfologico arabo.

In questo contributo cerchiamo di esporre la nostra esperienza con la creazione di un corpus giuridico parallelo italiano-arabo specializzato nel diritto internazionale. È un corpus allineato a livello di frase e annotato morfosintatticamente. Una versione del corpus bilingue allineato a livello di frase sarà disponibile gratuitamente per la comunità scientifica al sito del Laboratorio di Linguistica Computazionale dell'Università di Ca' Foscari, Venezia[1].

2 Stato dell'arte

A nostra conoscenza, fino al tempo di questo lavoro non esiste un corpus parallelo italiano-arabo nel dominio giuridico. Nell'ambito del progetto *L'arabo per la 488* (Picchi et al., 1999) è stato creato un corpus parallelo italiano-arabo di testi generici: si tratta di progetto finalizzato allo sviluppo di strumenti e risorse tanto per la lingua italiana quanto per la lingua araba, con particolare cura per l'aspetto contrastivo. Se invece guardiamo allo stato dell'arte delle nostre due lingue come partecipi insieme ad altre lingue di corpora paralleli, troviamo che l'italiano prende parte a risorse testuali multilingue in misura maggiore rispetto all'arabo.

Dei corpora paralleli in italiano e altre lingue ricordiamo *Bononia Legal Corpus* (Rossini Favretti et al., 2007), che è un corpus inglese-italiano di testi giuridici paralleli e comparabili,

1 http://project.cgm.unive.it

sviluppato presso l'università di Bologna. Il progetto è costituito in due fasi: nella prima fase si è costruito un corpus pilota, costituito da corpora paralleli in inglese e in italiano; mentre nella fase successiva vengono aggiunti corpora comparabili nelle due lingue riguardanti testi nell'ambito legislativo, giudiziario e amministrativo per analizzare le caratteristiche linguistiche dei due sistemi legali. Inoltre, nell'ambito del progetto CATEX (*Computer Assisted Terminology Extraction*) presso l'Accademia Europea di Bolzano è stato realizzato un corpus giuridico parallelo italiano-tedesco (Gamper, 1998). Questo corpus comprende una raccolta di leggi italiane con la relativa traduzione in tedesco con una dimensione di quasi 5 milioni di tokens, ed è allineato a livello di frase.

Per quanto concerne, invece, i corpora paralleli in arabo e altre lingue si rammenta EAPCOUNT (Hammouda, 2010), che è un corpus parallelo inglese-arabo con 341 testi delle Nazioni Unite allineati a livello di paragrafo. Inoltre, si menziona il corpus creato presso il laboratorio di linguistica computazionale dell'università autonoma di Madrid (Samy et al., 2006). Si tratta di un corpus parallelo multilingue (inglese-spagnolo- arabo) che contiene una collezione dei documenti delle Nazioni Unite, allineati a livello di frase e annotati morfosintatticamente.

3 Progettazione del corpus

Come dominio tematico del corpus abbiamo scelto il diritto internazionale e in particolare i diritti umani nel mondo. La scelta di questo genere testuale ha le seguenti motivazioni:

- Il linguaggio giuridico è uno dei linguaggi settoriali che presentano molte peculiarità sui diversi livelli di analisi linguistica, il che rende indifferibilmente necessario fornire e sviluppare corpora di testi giuridici;
- Per quanto riguarda la lingua araba, la maggior parte dei corpora giuridici disponibili sul web riguarda il codice di famiglia dei paesi arabi, che, ispirato ai principi della Shariah Islamica, contiene tante terminologie islamiche che non hanno corrispondenti in italiano. Per il problema dell'intraducibilità dei termini giuridici islamici tra l'arabo e l'italiano, abbiamo pensato quindi al diritto internazionale, dove risulta limitata l'influenza della dimensione religiosa dei termini;
- L'accuratezza della traduzione dei testi paralleli è un fattore essenziale soprattutto trattandosi di terminologie giuridiche, e nei documenti dell'Organizzazione delle Nazioni Unite (ONU) abbiamo trovato un livello di traduzione tanto accurato, visto il carattere ufficiale dei documenti.

4 Descrizione del corpus

I documenti del corpus sono dell'ONU. Si tratta di una grande raccolta di accordi, convenzioni, protocolli internazionali sempre nell'ambito del diritto internazionale in generale e dei diritti umani in particolare. La lingua originale dei documenti del corpus parallelo è l'inglese e sia i testi italiani che i testi arabi sono una traduzione dall'inglese. I testi del corpus si dividono in due categorie: la prima comprende un insieme di convenzioni e accordi internazionali nell'ambito dei diritti umani nel mondo, mentre la seconda contiene le convenzioni dell'Organizzazione Internazionale del Lavoro (ILO). In totale il corpus comprende all'incirca 1,1 milione di parole. Tabella 1 indica i dettagli del corpus.

language	n.parole	n.frasi	lunghezza media delle frasi	type/token ratio
Italiano	545682	18675	30	0.028
Arabo	615947	18391	39	0.068

Tabella 1. Dati statistici del corpus

5 Costituzione e preparazione del corpus

Per i testi del corpus il web rappresenta la fonte principale sia per i testi arabi che per quelli italiani. Il risultato di questa fase è un insieme di documenti in formato PDF in entrambe le lingue. Il formato PDF non consente, tuttavia, un trattamento automatico dei testi, quindi bisogna convertire i testi nel formato "Plain text format" che è adeguato a qualsiasi trattamento computazionale del corpus, e poi salvare i testi in UNICODE che è adeguato nel nostro caso dato che i sistemi di scrittura delle due lingue di interesse sono diversi.

Il processo della conversione non è, tuttavia, banale come sembra, soprattutto per la lingua araba. Fra le notevoli osservazioni individuate durante la conversione dei testi arabi ricordiamo: la perdita di alcuni caratteri, lo scambio tra certi caratteri (soprattutto tra "ا" e "ل"), l'inversione della direzione di scrittura (soprattutto i numeri), la perdita del formato del testo originale, ecc. Tutto questo richiede un grande sforzo per rimuovere ogni forma di "rumore" e restituire la

normalità dei testi. Nel caso dei testi italiani gli errori derivati dalla conversione riguardano maggiormente il cambiamento del formato del testo originale.

6 Trattamento del corpus

Fino al passo precedente, lo stato del corpus è grezzo, cioè senza nessuna annotazione linguistica utile per esplorare ed interrogare il corpus in modo migliore. L'importanza dei corpora annotati consiste non solo nella possibilità di esplorare ed estrarre informazioni dal testo, ma anche nel fornire "training e valutazione di algoritmi specifici in sistemi automatici." (Zotti, 2013).

Il trattamento automatico del nostro corpus comprende le seguenti fasi:

6.1 Segmentazione

La segmentazione dei testi è stata effettuata nelle due lingue a livello di frase. Per segmentare i testi abbiamo utilizzato un algoritmo nel pacchetto NLTK basato sulla punteggiatura (".", "?", "!"). Tuttavia, non mancano gli errori anche in questa fase; soprattutto per la mancanza dell'uso delle lettere maiuscole in arabo.

Vista la natura giuridica dei testi, si sono registrate alcune peculiarità riguardanti i confini di frase nei testi del corpus. In questo caso il segno della fine frase non è solo il punto finale come è il caso dei testi generali, ma i segni ":", ";" si possono considerare anche confine di frase, soprattutto quando iniziano una lista di clausole o commi. Il risultato di questa fase è un testo segmentato a livello di una sola frase per riga.

6.2 Tokenizzazione

Tokenizzare un testo significa ridurlo nelle sue unità ortografiche minime, dette tokens, che sono unità di base per ogni successivo livello di trattamento automatico. La complessità di questo compito dipende maggiormente dal tipo di lingua umana in trattamento nonché dal suo sistema di scrittura.

Nell'ambito del trattamento automatico della lingua araba riconoscere l'unità ortografica di base delle parole arabe appare un compito particolarmente complicato per effetto della complessità della morfologia araba, basata su un sistema flessionale e pronominale molto ricco (Habash, 2010). Ne consegue che per disambiguare al meglio le unità lessicali di un testo arabo ogni sistema di tokenizzazione necessita di un analizzatore morfologico. Per tokenizzare i testi arabi del corpus abbiamo utilizzato il sistema MADA+TOKAN[2] (Habash et al., 2009) che nel nostro caso ha avuto un'accuratezza all'incirca 98%. Nel caso dei documenti italiani si è utilizzato il tokenizzatore disponibile al sito di ItaliaNLP Lab[3].

6.3 Allineamento

Per il processo di allineamento si intende rendere due testi, o due unità testuali (nel nostro caso due frasi) allineati l'uno di fronte all'altro. Questa fase si configura come un processo essenziale lavorando sui corpora paralleli. L'allineamento viene effettuato normalmente da appositi programmi che si servono di metodi statistici e linguistici per mettere in corrispondenza due unità di testo l'una è traduzione dell'altra. Nel caso dei metodi statistici si utilizzano i calcoli probabilistici della lunghezza delle unità (frasi, parole, caratteri) dei due testi paralleli per stabilire una adeguata equivalenza tra i due testi in esame. Inoltre, il metodo statistico si può arricchire di repertori lessicali derivati da dizionari o corrispondenze traduttive prestabilite. Non c'è dubbio che l'utilizzo del metodo ibrido appare più conveniente soprattutto quando si tratta di lingue che hanno sistemi di scrittura tanto diversi tra loro, come per es. le lingue del nostro corpus.

Per allineare i nostri testi, abbiamo utilizzato *LogiTerm* che fa parte di *Terminotix*[4]. Questo programma segmenta e allinea automaticamente due testi creando il risultato in formati diversi (HTML, XML, TMX). L'accuratezza dell'allineamento nel nostro caso è all'incirca 95%, quindi non mancava un intervento manuale per correggere alcuni errori dovuti in generale alle caratteristiche linguistiche delle due lingue in questione. La maggior parte degli errori individuati durante l'allineamento riguarda la lunghezza della frase araba. Come si può osservare dal numero totale delle frasi nella Tabella 1, la lingua araba tende a congiungere le frasi, quindi non è raro di trovare un livello di allineamento 2 a 1. Dopo la verifica manuale dei risultati di questa fase, i testi allineati sono salvati in due formati XML e TMX.

2 We used version 3.2 of MADA+TOKAN
3 http://www.italianlp.it/
4 http://www.terminotix.com/index.asp?lang=en

<prop type="ltattr-match">1-1</prop>
<prop type="ltattr-id">17</prop>
<tuv xml:lang="it">
<seg>Ogni persona ha diritto al godimento dei diritti e delle libertà riconosciuti e garantiti nella presente Carta senza alcuna distinzione, in particolare senza distinzione di razza, sesso, etnia, colore, lingua, religione, opinione politica o qualsiasi altra opinione, di origine nazionale o sociale, di fortuna, di nascita o di qualsiasi altra situazione.</seg>
</tuv>
<tuv xml:lang="ar">
<seg>يتمتع كل شخص بالحقوق والحريات المعترف بها والمكفولة فى هذا الميثاق دون تمييز خاصة إذا كان قائما على العنصر أو العرق أو اللون أو الجنس أو اللغة أو الدين أو الرأى السياسى أو أى رأي آخر، أو المنشأ الوطني أو الاجتماعي أو الثروة أو المولد أو أى وضع آخر.</seg>
</tuv>
</tu>
<tu>

Tabella 2. Estratto del corpus allineato in TMX

<seg match="1-1" id="17">
<src>Ogni persona ha diritto al godimento dei diritti e delle libertà riconosciuti e garantiti nella presente Carta senza alcuna distinzione, in particolare senza distinzione di razza, sesso, etnia, colore, lingua, religione, opinione politica o qualsiasi altra opinione, di origine nazionale o sociale, di fortuna, di nascita o di qualsiasi altra situazione.</src>
<tgt>يتمتع كل شخص بالحقوق والحريات المعترف بها والمكفولة فى هذا الميثاق دون تمييز خاصة إذا كان قائما على العنصر أو العرق أو اللون أو الجنس أو اللغة أو الدين أو الرأى السياسى أو أى رأي آخر، أو المنشأ الوطني أو الاجتماعي أو الثروة أو المولد أو أى وضع آخر.</tgt>
</seg>

Tabella 3. Estratto del corpus allineato in XML

6.4 Annotazione del corpus

Per l'annotazione o l'etichettatura linguistica di un corpus si intende associare alle porzioni del testo informazioni linguistiche in forma di etichetta (tag o mark-up), sia per rendere esplicito il contenuto del testo sia per ottenerne una conoscenza approfondita. Il tipo di annotazione più conosciuto è quello morfosintattico o il cosiddetto POS (part-of-speech tagging), che consiste nell'attribuire ad ogni parola nel testo la sua categoria grammaticale. Il POS tagging possiede un'importanza rilevante nel trattamento automatico del linguaggio, in quanto rappresenta il primo passo nell'annotazione automatica dei testi, quindi gli errori riscontrabili durante questa fase potrebbero incidere sulle successive analisi.

Per taggare i testi arabi del nostro corpus, abbiamo utilizzato il pacchetto Amira 2.1 (Diab, 2009). Amira è un sistema di POS tagging basato sull'apprendimento supervisionato che utilizza le macchine a vettori di supporto (SVM). Questo sistema comprende tre moduli per il trattamento automatico della lingua araba: tokenizzazione, POS tagging, e base-phrase chunked. Nel nostro caso il sistema PoS Tagging di Amira raggiunge un'accuratezza all'incirca 94%.

Per i testi italiani si è usato Felice-POS-Tagger (Dell'Orletta, 2009). Felice-POS-Tagger è una combinazione di sei tagger, con tre algoritmi diversi. Ognuno dei tre algoritmi viene utilizzato per costruire un *left-to-right* (LR) tagger e un *right-to-left* (RL) tagger. L'accuratezza del Felice-POS-Tagger nel taggare i testi del nostro corpus è all'incirca 97%.

Le/RD organizzazioni/S dei/EA lavoratori/S e/CC dei/EA datori/S di/E lavoro/S hanno/V il/RD diritto/S di/E elaborare/V i/RD propri/AP statuti/S e/CC regolamenti/S amministrativi/A ,/FF di/E eleggere/V liberamente/B i/RD propri/AP rappresentanti/S ,/FF di/E organizzare/V la/RD propria/AP gestione/S e/CC la/RD propria/AP attività/S ,/FF e/CC di/E formulare/V il/RD proprio/AP programma/S di/E azione/S ./FS

Tabella 3. Estratto del corpus italiano annotato a livello PoS Tagging

ل/RP منظمات/NNS_FP العمال/DET_NN و/CC ل/IN منظمات/NNS_FP أصحاب/NN العمل/DET_NN الحق/DET_NN في/IN وضع/NN دساتيرها/NN_PRP_FS3 و/CC لوائحها/NN_PRP_FS3 الإدارية/PUNC، DET_JJ_FS و/CC في/IN انتخاب/NN ممثليها/NNS_MP_PRP_FS3 ب/IN حرية/NN_FS كاملة/PUNC، JJ_FS و/CC في/IN تنظيم/NN إدارتها/NN_FS_PRP_FS3 و/CC نشاطها/PUNC، NN_PRP_FS3 و/CC في/IN إعداد/NN برامج/NN عملها/PUNC./ NN_PRP_FS3

Tabella 4. Estratto del corpus arabo annotato a livello PoS Tagging

7 Conclusione

In questo lavoro abbiamo cercato di dare una descrizione del nostro progetto di creare un corpus parallelo italiano-arabo nel campo del diritto internazionale. La costruzione di tale corpus risponde allo scopo generale di fornire risorse linguistiche utili alle applicazioni della linguistica computazionale, soprattutto considerando la mancanza visibile dei corpora paralleli italiano-arabo di testi specialistici. Il trattamento computazionale del corpus è arrivato fino al PoS tagging, estendibile nel futuro ad altri livelli di annotazione e di arricchimento. Nel futuro intendiamo estendere questo corpus in due sensi: verticale e orizzontale. L'estensione orizzontale riguarda l'aggiunta di altri testi giuridici, mentre quella verticale ha a che fare con il trattamento automatico del corpus a livelli più avanzati.

- Bibliografia

Dell'Orletta F. 2009. Ensemble system for Part-od-Speech tagging. In *Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian.* Reggio Emilia, Italy.

Delmonte R. 2007. *VEST* - Venice Symbolic Tagger. In *Intelligenza Artificiale*, Anno IV, N° 2, pp. 26-27.

Diab, M. 2009. Second generation AMIRA tools for Arabic processing: Fast and robust tokenization, POS tagging, and base phrase chunking. In *2nd International Conference on Arabic Language Resources and Tools*, Cairo, Egypt

Gamper, J. 1998. CATEX– A Project Proposal. In Academia, 14, 10-12

Hammouda S. 2010. Small Parallel Corpora in an English-Arabic Translation Classroom: No Need to Reinvent the Wheel in the Era of Globalization. In Shiyab, S., Rose, M., House, J., Duval J.,(eds.), *Globalization and Aspects of Translation,* UK: Cambridge Scholars Publishing

Lenci A., Montemagni S., Pirrelli V. 2012. *Testo e computer: elementi di linguistica computazionale*, Carocci editore, Roma

Rossini Favretti R., Tamburini F., Martelli E. 2007. Words from Bononia Legal Corpus. In *Text Corpora and Multilingual Lexicography* (W.Teubert ed.), John Benjamins

Samy, D., Moreno-Sandoval, A., Guirao, J.M., Alfonseca, E. 2006. Building a Multilingual Parallel Corpus Arabic-Spanish-English. In *Proceedings of International Conference on Language Resources and Evaluation LREC-06*, Genoa, Italy

Zotti, P. 2013. Costruire un corpus parallelo Giapponese-Italiano. Metodologie di compilazione e applicazioni. In Casari, M., Scrolavezza, P. (eds), *Giappone, storie plurali,* I libri di Emil-Odoya Edizioni. Bologna

Habash, N., Rambow, O., Roth, R. 2009. MADA+TOKAN: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization. In Choukri, K., Maegaard, B., editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*. The MEDAR Consortium, April.

Habash, N. 2010. Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers.

Picchi E. , Sassolini E. , Nahli O. , Cucurullo S. 1999. Risorse monolingui e multilingui. Corpus bilingue italiano-arabo. In *Linguistica computazionale*, XVIII/XIX, Pisa.