

Web-based term mining tra terminologie e memorie di traduzione

Bruno Ciola, Natascia Ralli

Abstract:

The management of terminological data and translation memories is extremely important for terminologists and translators. With the help of an example, BISTRO, the juridical terminological information system, we aim at showing how such a system is best exploited. BISTRO offers several tools and resources, such as terminology data bases, bilingual corpora, term extraction, text recognition, text alignment etc. for the translation and text analysis of legal and administrative documents. The purpose is both to enhance the use of consistent terminology and at the same time reduce the time spent in terminology research. The system can be accessed, updated and integrated online, making it thus an ideal working tool, also for external collaborators.

1. Il lavoro terminologico

Inteso come strumento di supporto alla traduzione o in generale per favorire la comunicazione tecnico-scientifica oppure per la standardizzazione e normazione linguistica, il lavoro terminologico (LT) rappresenta una fra le attività più dispendiose e talvolta anche difficili da pianificare in termini di tempo e risorse. Lo scopo del LT è di soddisfare le specifiche richieste nel minor tempo possibile, con l'impiego delle risorse adeguate allo scopo, garantendo la massima qualità. La qualità, tuttavia, dipende dalle esigenze del prodotto finale (banca dati terminologica, glossario, dizionario elettronico o cartaceo ecc.), degli utenti (traduttori, linguisti in generale, esperti del settore, ecc.) ed anche della sua funzione.

In generale si distingue fra il lavoro terminologico sistematico e quello ad-hoc. Nel LT sistematico vengono descritti i concetti di un ambito limitato e ben circoscritto, confrontandoli in un successivo momento con i concetti nelle altre lingue creando relazioni di equivalenza tra di essi. In questo caso il lavoro preparativo (selezione e acquisizione dei testi da utilizzare, preparazione del materiale, ricerche di approfondimento, consultazione dell'esperto ecc.) sarà utile nell'intero corso delle ricerche terminologiche. Diversa è la realtà nel LT ad-hoc, ove il linguista-terminologo è chiamato a risolvere un preciso problema linguistico cui dovrà trovare una soluzione nel minor tempo possibile.

Un primo passo verso la razionalizzazione e automazione del lavoro terminologico è rappresentato dai programmi per la gestione e il trattamento della terminologia. Diversamente dai dizionari elettronici, che spesso coprono solamente un ambito assai limitato e quindi peccano di dettagli e precisione, i primi permettono di gestire i dati terminologici in modo più flessibile sia in fase di elaborazione sia in fase di utilizzazione dei dati. Possono anche interagire con strumenti per l'estrazione (semi)automatica della terminologia, caratterizzati da un processo in cui il sistema è in grado di estrapolare da un dato testo i termini (sostantivo, gruppo di sostantivi, verbi, collocazioni, fraseologie ecc.) contenuti. Al lavoro automatico va in ogni caso affiancato l'intervento da parte del linguista-terminologo nel controllo e nella revisione; permette, tuttavia, di aiutare il lavoro soprattutto nelle procedure macchinose e monotone. In un successivo momento la lista di termini potrà essere confrontata con quelli già presenti nella banca dati per segnalarne i termini mancanti. Una simile procedura si può utilizzare anche per la creazione ex-novo di dizionari bilingui, ad esempio in presenza di testi bilingui allineate frase per frase. In questo caso, successivamente alla

scansione linguistica del testo di partenza, vengono proposti i probabili traduttori dei termini rilevati della lingua di arrivo. Anche in questo caso il lavoro automatico del computer viene affiancato dall'uomo che dovrà intervenire per fornire orientamento alla macchina.

Nel LT sono tuttavia necessarie ulteriori risorse, che servono a 1) comprendere il termine più adatto ad un'espressione nella lingua di partenza: essendo il terminologo solo di rado anche un esperto in materia, avrà bisogno di testi di riferimento descrittivi che gli forniscano informazioni aggiuntive ai termini che sta trattando; 2) rilevare eventuali termini sinonimi o varianti, un aspetto importante nel LT sistematico; 3) verificare o accertare il significato di un determinato termine ed il suo uso: perché una volta rilevato un possibile traduttore, sarà necessario accertarne l'equivalenza (parziale o totale) nei confronti del termine di partenza.

2. Information Retrieval

Nasce da qui l'esigenza dell'*information retrieval* che ha lo scopo di trovare informazioni su una specifica questione. L' *information retrieval* consiste nella formulazione di una richiesta (un termine, un gruppo di parole, fraseologie ecc.), la quale permette di trovare i documenti corrispondenti ai criteri di ricerca all'interno di un corpus¹, vale a dire un insieme di testi riguardante uno o diversi settori. Al giorno d'oggi, uno fra i maggiori corpora, nella loro più ampia accezione, è Internet, una risorsa che permette di ricercare velocemente e gratuitamente una notevole quantità di testi. Non sempre, tuttavia, è facile ottenere le informazioni, basandosi la ricerca esclusivamente su singole parole chiave "decontestualizzate", ovvero separate dal contesto e che possono avere significati diversi in contesti diversi. Sarà eventualmente possibile combinare diversi termini nella stessa ricerca per ulteriormente filtrare i dati provenienti dal corpus, ma per ottenere un risultato ancora più mirato sono necessari ulteriori sistemi. Nella ricerca in Internet ad es. la selezione si potrà fare in base ai tipi di siti da selezionare preventivamente (istituzionali <> privati, portali verticali, archivi di riviste). In altri casi si potranno sfruttare i risultati di una classificazione dei testi compresi nel corpus. Tramite la classificazione è possibile assegnare ad ogni testo diversi descrittori in relazione al contenuto (ad es. in ambito giuridico in base al tipo di diritto: diritto internazionale, diritto civile o penale, ecc.; oppure alla tipologia testuale: legge, sentenza, manuale, commento, rivista giuridica, ecc.). Sarà altresì possibile indicizzare i testi, una procedura particolarmente vantaggiosa nel caso di corpora molto ampi. Creando un indice, è possibile a) escludere eventuali parole "inutili" che appesantiscono la ricerca (cd. stop word, come ad esempio articoli, congiunzioni, ecc.); b) rilevare gli stessi termini o gruppi di termini che si ripetono solo una volta, inserendo un puntatore che rinvia alla posizione originale nel testo. Rispetto al testo integrale, tuttavia, l'indicizzazione comporta anche a degli svantaggi, permettendo di ricercare solamente nell'ambito della sistematicità così com'è stata impostata, con il rischio di non poter più accedere alle informazioni liberamente.

3. Acquisizione dei testi

Tuttavia, per poter effettuare una classificazione oppure un'indicizzazione dei testi, sarà necessario acquisirli preventivamente. Oltre ai testi disponibili in forma digitale (Cd-rom ecc.) che i relativi editori potranno eventualmente mettere a disposizione, bisognerà affrontare la problematica dei diritti d'autore quando si tratta di utilizzare testi reperiti da Internet. Si possono copiare i dati da Internet su un server locale, tuttavia, per non violare eventuali diritti di copyright, è possibile rappresentare i dati solamente come citazione, limitando quindi anche la trasformazione e l'adattamento degli stessi. Per ovviare a questo problema sarebbe possibile segnalare

¹ In questo contesto, per corpus bilingue si intendono anche le cd. *memorie di traduzione*, ovvero testi allineati a livello di frase nelle due lingue, di cui uno rappresenta la traduzione dell'altro.

esclusivamente il *link* della rispettiva pagina, ma pure con questa soluzione, vista la breve vita dei contenuti sul web, il risultato non sarebbe soddisfacente.

4. Consultazione dei dati

Un altro aspetto importante è quello legato all'*ergonomicità* del lavoro. Dover utilizzare diverse applicazioni per funzioni diverse (ad esempio un programma per la gestione della terminologia, un'interfaccia per accedere ai corpora monolingui, un'altra per consultare corpora bilingui, un'altra ancora per accedere ad un programma di estrazione terminologica ecc.), rappresenta un notevole svantaggio – sia in fase di preparazione dei dati (quindi nel lavoro del terminologo), sia in fase di consultazione da parte dell'utente. E' quindi auspicabile poter utilizzare un'interfaccia unica da cui accedere direttamente alle varie funzioni. Si verrebbe così a creare un sistema che permette di accedere, tramite un'unica interrogazione, alla banca dati terminologica, ai corpora bilingui, ai testi eventualmente raccolti da Internet ed anche direttamente ai testi presenti su Internet.

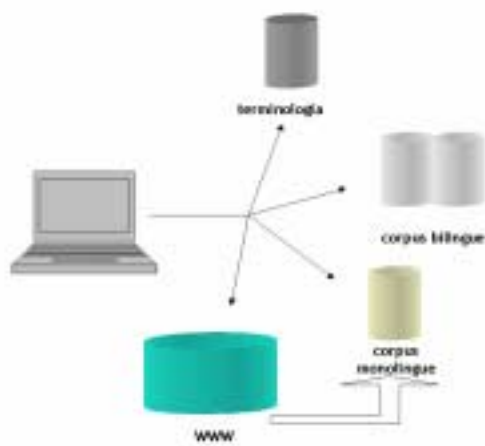


fig. 1: ricerca parallela nelle terminologie, nei corpora e in Internet

4. BISTRO

Un'applicazione pratica relativa alle problematiche affrontate nei paragrafi precedenti è rappresentata da BISTRO (<http://www.eurac.edu/Bistro>), il *Sistema Informativo della Terminologia Giuridica Bolzano*, che mette a disposizione dell'utente una serie di strumenti volti all'analisi terminologica e testuale, nonché alla traduzione di documenti giuridici ed amministrativi.

Sviluppato dall'area scientifica "Lingua e Diritto" dell'Accademia Europea di Bolzano, il sistema consta di:

- una banca dati contenente complessivamente ca. 16000 termini, elaborati in lingua italiana, tedesca e ladina, che vengono, di norma, visualizzati con le relative definizioni e contesti esemplificativi, fornendo in tal modo un valido aiuto sia al traduttore sia al giurista e rendendo il lavoro terminologico esauriente da ogni punto di vista;
- un corpus (CATEX) costituito da testi paralleli, di carattere giuridico, in lingua italiana e tedesca, contenente ad es. il Codice Civile, il Codice Penale, il Testo Unico delle Imposte sui Redditi, il Codice Processuale Civile. Tale corpus rappresenta un valido strumento di ricerca per il terminologo e per il traduttore nella ricerca di contesti, definizioni, collocazioni, ecc.

BISTRO utilizza Internet quale piattaforma di lavoro. Tale supporto consente di ampliare ed aggiornare la banca dati in tempo reale, nonché di accedere in qualsiasi momento al sistema, senza bisogno di alcuna registrazione o licenza commerciale.

BISTRO si potrebbe definire come un “insieme di strumenti linguistici integrati in un’unica interfaccia”, quali:

- il modulo di gestione terminologica
- il modulo di gestione del corpus
- il modulo di annotazione testi
- il modulo di estrazione dei termini.

Ricerca termini

Il modulo di gestione terminologica è rappresentato dalla funzione **ricerca termini** attraverso cui è possibile consultare la banca dati dell’area “Lingua e Diritto”. Tale opzione viene attivata tramite una finestra in cui l’utente deve digitare il termine, di carattere giuridico, che desidera visualizzare. Specificato il termine, il sistema visualizza una *hitlist* (fig. 2), nella quale compaiono tutti i *match* reperiti da BISTRO all’interno del suo database terminologico. Facendo click sul logo “Bistro”, vicino al termine desiderato, si accede ad una scheda terminologica (fig. 2) in cui l’utente trova una serie di informazioni relative al termine designato, quali settore di appartenenza, definizione, contesto, termine corrispondente in lingua tedesca, che può variare a seconda del sistema legale (tedesco, austriaco, svizzero), eventuale commento.

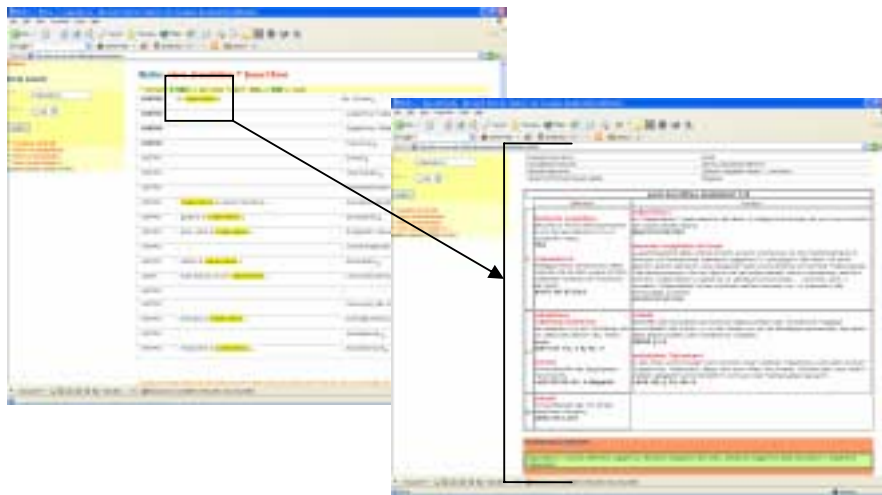


fig. 2: funzione “ricerca termini”

Ricerca nel corpus

Il modulo di gestione del corpus si esplica attraverso la funzione **ricerca nel corpus**, valido strumento di aiuto che permette sia la ricerca di contesti e definizioni sia la ricerca di collocazioni, fraseologismi, ricorrenze, ecc.

In BISTRO il corpus è costituito da:

- CATEX (cfr. punto 4)
- siti Internet di stampo giuridico

Il filtraggio dei dati contenuti nel corpus avviene sulla base di parametri (obbligatori e non) che vengono stabiliti in precedenza dall'utente. Quello che si ottiene è una ricerca selettiva, ben circoscritta. La selezione avviene indicando

- il tipo di lingua (italiano, tedesco, ladino)
- il tipo di sistema legale (italiano, austriaco, tedesco federale, svizzero, europeo, internazionale)
- il tipo di diritto (diritto penale, amministrativo, civile, legislazione sociale, stradale ecc.)
- l'ordinamento giuridico (internazionale, europeo, statale, regionale, comunale)
- tipologia testuale (legge, regolamento, altro).

Una volta inserito il termine, che si desidera cercare, BISTRO apre una finestra contenente una *hitlist* (fig. 3) di link interni ed esterni. I link interni si riferiscono ai *match* contenuti nel CATEX, mentre quelli esterni rimandano a siti Internet di carattere legale, contenenti il termine desiderato. Dal *match* indicato nel CATEX è poi possibile accedere al testo di consultazione (fig. 3), ed eventualmente alla sua versione tedesca.

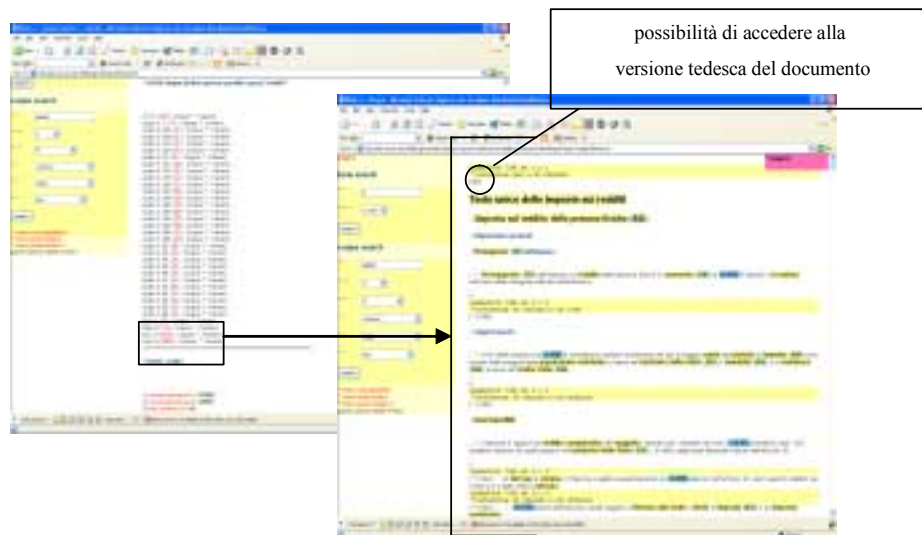


fig. 3: ricerca nel CATEX

All'interno del modulo di gestione del corpus è, inoltre, possibile richiamare la funzione KWIC (Key Words in Context), ossia "parole nel contesto", una sorta di *concordancer* di norma utilizzato per creare glossari o dizionari, ma che può trovare anche una sua applicazione pratica in ambito traduttivo, o di apprendimento stesso della lingua, in quanto permette di individuare le caratteristiche lessicali e le collocazioni tipiche di un dato termine.

Come si accede a tale funzione? Una volta inserito il termine e selezionato il testo di consultazione (all'interno del CATEX), si apre una finestra contenente tutte le occorrenze del termine, evidenziato al centro della pagina e visualizzato con alcune parti del contesto in cui viene usato (fig. 4). Dal segmento è poi possibile accedere alla versione italiana e tedesca del testo di consultazione.

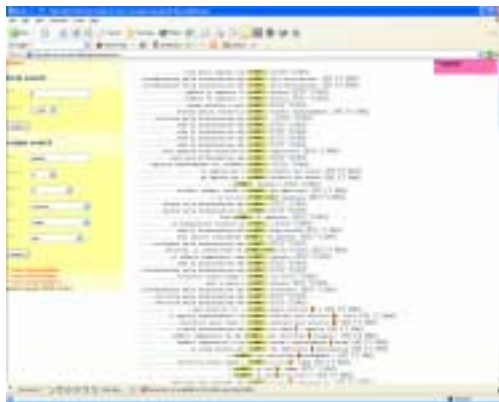


fig. 4: funzione KWIC

Annotazione testi

Il modulo di **annotazione testi** è possibile effettuare una ricerca terminologica all'interno di un testo o di una pagina Web, selezionati dall'utente. L'utilizzo di tale funzione permette di ottenere le seguenti informazioni:

- disponibilità di definizioni, contesti o traduzioni nella banca dati di BISTRO
- accertamento della presenza di un termine all'interno della banca dati attraverso l'evidenziazione del medesimo con colori diversi
- settore di appartenenza del termine (diritto amministrativo, penale, ecc.) con relativa traduzione.

L'annotazione del testo può essere eseguita in due modi: il primo consiste nell'inserire l'URL del testo, che si vuole "esaminare", nella maschera di BISTRO; il secondo nell'inserire il documento in questione (*.doc., rtf, o *.txt) all'interno della maschera di BISTRO tramite la comune funzione di "copia - incolla". Una volta effettuata tale operazione, il testo viene visualizzato e i termini, presenti nella banca dati, evidenziati. Dai singoli termini si può poi accedere alle relative schede terminologiche.

Tale funzione consente dunque al traduttore di confrontare i termini della lingua di partenza, con quelli già presenti nella banca dati in modo da individuare immediatamente i termini che necessitano di una traduzione o di prendere, invece, direttamente visione dei possibili traduttori nella lingua di arrivo.

Estrazione termini

Attraverso il modulo **estrazione termini** BISTRO consente di estrapolare da pagine Web o da documenti in formato *.doc., *.rtf, o *.txt, selezionati dall'utente, potenziali "termini-candidati" per la ricerca terminologica, ricerca che può essere effettuata all'interno di un corpus sia monolingue sia bilingue. La selezione semi-automatica dei termini viene eseguita basandosi su un'analisi di tipo statistico e linguistico. Vengono dunque presi in considerazione i seguenti criteri:

- la frequenza, la quale individua le unità che maggiormente ricorrono all'interno del testo o, più in generale, del corpus selezionato;
- *stop-words*, elementi che vengono scartati automaticamente dal sistema, come congiunzioni, articoli, preposizioni ecc.;

- “termini-modello”, criterio che utilizza termini già presenti nella banca dati: ad esempio se nel database è contenuto il termine “colpa”, il sistema può riconoscere come potenziali candidati “colposo” o “colpevolezza”.

L’**estrazione termini** permette al linguista-terminologo di confrontare i termini estrapolati con quelli presenti nella banca dati, onde evitare doppie entrate, nonché di aiutarlo nell’individuazione dei termini mancanti.

Conclusioni

Nella presente relazione si è voluto sottolineare l’importanza dell’aspetto legato all’*ergonomicità* del lavoro in ambito linguistico. Il poter utilizzare un unico strumento che racchiuda in sé applicazioni di gestione terminologica, di gestione del corpus, di estrazione terminologica ecc. rappresenta di fatto un vantaggio sia dal punto di vista dei tempi di lavoro sia, conseguentemente, dei costi.

Il prototipo BISTRO, presentato durante la relazione, cerca di andare incontro a quelle che sono le esigenze sia del traduttore sia del terminologo. In un ambito prettamente specifico, come quello giuridico, tale strumento si rivela di grande ausilio in quanto da un lato offre delle spiegazioni di carattere giuridico legate al termine o concetto, che si rivelano fondamentali per la comprensione del significato, dall’altro propone dei possibili traduttori nella lingua di arrivo, oltre ad offrire una serie di informazioni di carattere linguistico, quali collocazioni, fraseologismi ecc..

Applicato per il momento solo in Alto Adige in cui, data la realtà locale caratterizzata dalla parificazione della lingua italiana e tedesca, è necessario tradurre in entrambe le lingue ogni sorta di documento, BISTRO costituisce un prototipo da poter applicare anche ad altre lingue ed altri ambiti di carattere specifico.

Bibliografia

- Aston, G. (2001). "Learning with corpora: an overview" in G. Aston (ed.) *Learning with corpora*. Houston: Athelstan, 7-45.
- Kugler, M., K. Ahmad & G. Thurmair (Eds.) (1991). *Translator's Workbench. Tools and Terminology and Text Processing*.
- Martoglia, R.. EXTRA (2001). *Progetto e Sviluppo di un Ambiente per Traduzioni Multilingua Assistite*. Università degli studi di Modena e Reggio Emilia, Facoltà di Ingegneria (Tesi di Laurea AA 2000-2001)
- Melby, A. (1992). "The translator workstation". In J. Newton. (1992). 147-165.
- Streiter, O. & Zielinski, D. & Ties, I. & Voltmer, L. (2002). „Similarity-based Term Extraction for Minority Languages: A case- study on Ladin". In: *Proceedings Soziolinguistica y Language Planning*, Ortisei, Italy.
- Streiter, O. & Knapp, J. & Voltmer, L. (2003). "A browser-like repository for open learning resources". In: *ED-Media, World Conference on Educational Multimedia, Hypermedia & Telecommunications*, Honolulu, Hawaii.
- Streiter, O. & Voltmer, L. (2002). "Document Classification for Corpus-based Legal Terminology". In: *The 8th International Conference of the International Academy of Linguistic Law*, Iași, Romania
- Wright, S. E. and Budin, G. (1997). *Handbook of Terminology Management*, vol. 1, Amsterdam, Philadelphia: John Benjamins Publishing Company.