Terminologia e corpora per il dialogo italo–tedesco
ISBN 978-88-548-8650-6
DOI 10.4399/97888548865067
pag. 151–188 (settembre 2015)

GEROM

Una proposta di banca dati lessicale e testuale

Claudio Fantinuoli

1. Introduzione

Il progetto GEROM[1] si propone di realizzare un nuovo strumento di consultazione lessicale, di tipo bilingue e dinamico, pensato per supportare l'analisi e la traduzione di testi legati ad ambiti quali la politica, la società e la cultura dei paesi di lingua tedesca e italiana. Esso mira a supportare l'attività di chiunque, a vario titolo, operi per migliorare il dialogo e la comunicazione, specialistica e non, tra i due paesi: in primo luogo gli interpreti e i traduttori chiamati ad operare in modo professionale nella traduzione di testi afferenti agli ambiti sopra indicati, ma anche tutti coloro che per varie ragioni si confrontano con testi in lingua tedesca e italiana e desiderano eseguire approfondite analisi di natura lessicale, sia in chiave sincronica che diacronica.

La piattaforma lessicale e testuale si prefigge di superare alcuni dei limiti che caratterizzano le banche terminologiche tradizionali, descritte nel capitolo precedente. Sfruttando i progressi compiuti negli ultimi anni in numerose discipline, la piattaforma ambisce a creare uno strumento di consultazione che aumenti l'affidabilità, il carico informativo, le tipologie di analisi eseguibili e, non da ultimo, la longevità dei dati raccolti in queste banche dati. In particolare, fra le principali sfide alle

1. www.gerom.eu.

quali il progetto cerca di rispondere, rientra il superamento della classica contrapposizione tra opera di consultazione lessicografica e terminografica, in termini di metodologia e di approccio adottato, che risulta in generale più descrittivo nei dizionari e tendenzialmente prescrittivo nelle banche dati terminologiche[2]. In passato questa contrapposizione ha provocato una certa reticenza, anche in ambito accademico, a concepire e realizzare strumenti linguistici capaci di integrare repertori lessicali considerati tra loro poco compatibili, perché legati a temi di natura profondamente diversa (si pensi alla terminologia tecnica di un determinato settore industriale in contrapposizione a temi di natura socio–politica riguardanti un determinato paese) o perché diversi sono gli obiettivi perseguiti (descrivere la realtà linguistica oppure normarla).

GEROM desidera conciliare questi diversi approcci, creando uno strumento bilingue in grado di raccogliere sia lessici di specialità, relativi ad esempio ai linguaggi economici o scientifici, sia lessici che abbracciano l'attualità politica e sociale, che in molti casi riguardano temi legati solo ad una delle due realtà linguistiche, quali la riforma del lavoro e del welfare, oppure intrisi di elementi emotivi, connotativi e ideologici, come l'immigrazione e il dibattito sul futuro dell'Unione Europea.

Per far fronte a queste esigenze è stato necessario progettare una piattaforma che abbandonasse, almeno in parte, le impostazioni classiche adottate da progetti analoghi, sia in termini di organizzazione delle schede, sia in termini di risorse utilizzate. Le schede tradizionalmente impiegate all'interno delle banche terminologiche, statiche per loro natura, presentano un van-

2. La lessicografia ha il compito di elaborare le regole e la metodologia per la raccolta, la gestione e la descrizione del patrimonio lessicale di una lingua o di una sua varietà. Tipicamente il prodotto finale del lavoro lessicografico è il dizionario. Essa basa la sua attività di raccolta e organizzazione dei repertori lessicali su un approccio semasiologico, che pone al centro delle sue indagini il segno linguistico di cui ne studia il significato. La terminografia si occupa invece dei lessici speciali e ha come oggetto i termini. Opera generalmente in un'ottica onomasiologica, si interroga cioè su come sono o dovrebbero essere denominati gli oggetti di indagine, partendo dunque dal concetto per arrivare al termine.

taggio evidente da un punto di vista didattico, poiché offrono un modello standard, quindi semplice da imparare e da operazionalizzare, tuttavia esse non sono in grado di rispondere adeguatamente né alla complessità dei lessici specialistici, come hanno evidenziato negli ultimi anni gli studi terminologici[3], né tantomeno alle peculiarità dei lessici legati all'attualità e alla loro traduzione, operazione quest'ultima mai completamente imparziale (v. le riflessioni contenute nel primo capitolo).

Fra le strategie adottate per far fronte a queste sfide, un ruolo particolare lo riveste l'abbandono della classica struttura della scheda terminologica, incentrata su un termine *vedette* e su una serie di sinonimi gerarchizzati, preferendo una struttura che, seppur orientata al significato, permetta di organizzare le parole con maggiore flessibilità (v. 3). L'integrazione di una serie di risorse in grado di estendere il carico informativo delle schede, adattandolo alle potenziali esigenze dell'utente finale, contribuisce inoltre al superamento di limiti quali la carenza informativa delle stesse e la loro inadeguatezza a descrivere il livello emotivo–valutativo del lessico, nonché i complessi rapporti tra lessemi nel tempo e nello spazio sociale (diastratia). Per ovviare, almeno in parte, a questi limiti è possibile offrire strumenti per eseguire un'approfondita analisi dei contesti d'uso della parola ricercata e dei suoi possibili traducenti, così come ci insegna la linguistica dei corpora, oppure osservare le frequenze di distribuzione della parola all'interno dei repertori testuali raccolti (in base alle fonti, agli autori, ecc.). Non da ultimo, le informazioni linguistiche ed extra–linguistiche disponibili sul web, come proposte traduttive, definizioni e immagini, possono contribuire a completare ed estendere l'offerta informativa registrata nelle schede.

Il presente capitolo intende fornire una panoramica delle soluzioni adottate all'interno del progetto GEROM. Partendo

3. Cfr. F. Bertaccini, M. Prandi, S. Rintuzzi, S. Togni, "Tra lessico naturale e lessici di specialità: la sinonimia". In: *Studi linguistici in onore di Roberto Gusmani*, a cura di R. Bombi, G. Cifoletti, F. Fusco, L. Innocente, V. Orioles, H. Marquardt, Edizioni dell'Orso, Alessandria 2006, pp. 171–192.

dalla descrizione della macrostruttura della piattaforma verranno analizzati il modello di scheda terminologica utilizzata, i corpora e le fonti esterne integrabili, cercando di descriverne l'utilità ai fini delle esigenze di analisi lessicale indicate nei nel primo e nel secondo capitolo del presente volume. I principi che governano la realizzazione dell'interfaccia utente e il workflow del suo funzionamento concludono infine il presente contributo.

2. Macrostruttura

La macrostruttura di GEROM si compone di tre parti, come illustra la figura 1: una banca dati contenente la raccolta dei repertori terminologici e i corpora; un sistema per il recupero, in tempo reale, di informazioni da fonti web e infine un'interfaccia grafica per l'interrogazione della banca dati e per la visualizzazione dei risultati.

Figura 1.

La banca dati rappresenta il cuore dell'architettura. Per ogni tema trattato all'interno del progetto, essa contiene un repertorio terminologico bilingue, organizzato in schede (v. 3), e un corpus comparabile associato a tale repertorio (v. 4). La banca

dati integra inoltre due grandi corpora generali per l'italiano e il tedesco[4]. La seguente figura ne schematizza la struttura.

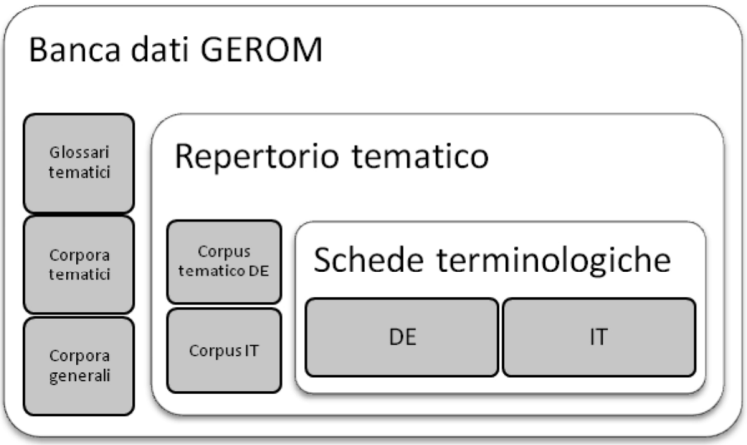

Figura 2.

La piattaforma si avvale di una struttura che prevede l'integrazione dei seguenti dati:

— Schede terminologiche
— Corpora monolingui comparabili sui temi affrontati nei singoli repertori terminologici
— Corpora generali monolingui
— Statistiche di distribuzione dei singoli lessemi
— Informazioni raccolte in tempo reale dal web

Queste risorse saranno oggetto di un'analisi dettagliata all'interno dei seguenti paragrafi.

4. M. Baroni, S. Bernardini, A. Ferraresi, E. Zanchetta, "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web–Crawled Corpora", *Language Resources and Evaluation*, vol. 43, n. 3 (2009), pp. 209–226.

3. Schede terminologiche

La banca dati terminologica si compone di una serie di raccolte lessicali, organizzate per domini, su un'ampia gamma di temi che vanno dall'attualità al costume, dalla politica alla società, oltre a includere i classici domini specialistici in ambito tecnico, giuridico e scientifico. I singoli repertori sono formati da un numero variabile di schede terminologiche, definite dalla norma ISO come un "insieme strutturato di dati terminologici che si riferiscono ad un concetto"[5]. I campi che costituiscono le singole schede contengono dunque tutte le informazioni ritenute utili, ai fini della raccolta stessa, per descrivere, in entrambe le lingue, un concetto e tutte le realizzazioni linguistiche ad esso legato.

Fra le peculiarità del modello di scheda adottato ritroviamo l'abbandono del termine *vedette*, ovvero del termine principale della scheda. La scelta del termine *vedette*, tradizionalmente motivata dalla struttura gerarchica degli alberi concettuali attraverso i quali i termini vengono rappresentati in terminologia, non prende in considerazione la realtà stratificata e differenziata delle raccolte terminologiche, una realtà che nella pratica si rivela piuttosto lontana dall'ideale normativo di univocità evocato dalla scienza terminologica[6]. La strutturazione dei dati terminologici secondo il principio del termine *vedette* risulta infatti poco adatta ad accogliere le esigenze di natura descrittiva, tipica dell'approccio lessicografico, poiché incapace, ad esempio, di rappresentare la variabilità sociolinguistica dei lessemi, che all'interno di vere realizzazioni testuali si manifesta nell'utilizzo molto frequente della sinonimia parziale[7]. Per questo motivo, all'interno di GEROM la struttura con termine *vedette* lascia

5. ISO 1087, 6.1.3.

6. Cfr. H. Gerzymisch–Arbogast, *Termini im Kontext*, Narr, Tübingen 1996.

7. Cfr. M. Cerruti "Il concetto di variabile sociolinguistica a livello del lessico", *Studi italiani di linguistica teorica e applicata*, vol. 40, n. 2 (2011), pp. 211–231.

il posto ad una struttura basata sui synset, così come proposta da Bertaccini[8] e descritta nel dettaglio nel paragrafo successivo.

Una seconda caratteristica del modello di scheda adottato riguarda il numero di campi utilizzato e di conseguenza la quantità di informazioni registrate. Frutto di un'attenta analisi condotta sulle esigenze degli utenti della piattaforma, il numero di campi inseriti nel modello di scheda utilizzato è molto inferiore rispetto a quello teorizzato dagli studi classici sulla terminologia[9]. Esso è tuttavia in linea, se non addirittura superiore, a quello utilizzato in progetti analoghi, in primis quelli accademici. Come sottolineano Bertaccini e Lecci[10] infatti:

> A seconda del tipo di ricerca che affrontiamo, il modello delle schede deve essere coerente con il tipo di studio che stiamo realizzando e quindi con la tipologia di informazioni che stiamo capitalizzando.

Dovendosi orientare alle concrete esigenze dell'ampia ed eterogenea platea dei fruitori finali, nelle schede del progetto si è scelto di mantenere solo i campi principali, scartando le informazioni ritenute di importanza secondaria. Inoltre, la scelta di utilizzare un numero abbastanza ridotto di campi è riconducibile, in alcuni casi, all'ubiquità della registrazione manuale dovuta all'integrazione dinamica di numerose fonti di informazione esterne, come verrà descritto nel paragrafo 6.

Il modello della scheda adottato prevede dunque una serie di campi di base, generalmente riconosciuti come indispensabili per la compilazione di una raccolta terminologica[11], ovvero:

— termine e sue varianti

8. Cfr. F. Bertaccini, M. Prandi, S. Rintuzzi, S. Togni, *op. cit.*

9. Cfr. J.C. Sager, *Language Engineering and Translation. Consequences of Automation*, Benjamins, Amsterdam 1994.

10. F. Bertaccini, C. Lecci, "Conoscenze e competenze nell'attività terminologica e terminografica", *Terminologia, ricerca e formazione, Publifarum*, n. 9 (2009).

11. Cfr. J.C. Sager, *op. cit.*

— contesto
— fonte contesto
— categoria grammaticale
— definizione
— fonte definizione
— collocazioni
— note

Alcuni di questi campi, seppur ridondanti rispetto alle informazioni dinamiche che completano l'offerta informativa (v. 6), sono stati comunque mantenuti, da un lato proprio in virtù di una prassi ormai consolidata in ambito terminografico, dall'altro per sopperire ad eventuali limiti, imprecisioni o errori che potrebbero caratterizzare le informazioni generate dinamicamente.

Ai campi di cui sopra se ne aggiungono alcuni propriamente amministrativi:

— autore
— data di compilazione
— dominio

Oltre ai campi standard appena elencati, le schede prevedono una serie di campi specifici, di norma non presenti all'interno delle banche dati terminologiche tradizionali, selezionati per soddisfare le esigenze peculiari del progetto. Tali campi risultano caratterizzanti in presenza, ad esempio, di concetti culturalmente e/o ideologicamente connotati, legati cioè alla realtà culturale e linguistica di una sola delle due lingue oppure all'orientamento politico, culturale e ideologico di un determinato parlante o scrivente.

Fra questi campi rientrano:

— direzione traduttiva
— tipo di equivalenza
— frequenza d'uso

I seguenti paragrafi intendono fornire una breve panoramica dei campi che compongono le schede terminologiche e del loro significato all'interno dell'architettura complessiva.

3.1. *Termine e sue varianti: il synset*

Come introdotto in precedenza, l'abbandono del campo contenente il termine principale (*vedette*), attorno al quale normalmente ruota l'intera scheda terminologica, rappresenta un elemento di novità rispetto alle banche terminologiche tradizionali. Ad esso si sostituisce il concetto di "synset", ovvero l'idea di legare in un unico nodo semantico tutti i lessemi utilizzati per identificare un determinato referente (concetto). Questi gruppi o cluster lessicali sono accomunati dalla presenza di una qualche relazione di sinonimia che intercorre tra le parole che li costituiscono, senza la necessità di eleggere una di esse a termine preferito. Poiché sono posti sullo stesso piano concettuale, tutti i lessemi che fanno parte di un synset verranno utilizzati per indicizzare il motore di ricerca della banca dati, permettendo il recupero delle informazioni associate al termine ricercato e a tutte quelle presenti nel cluster (in entrambe le lingue). Dal punto di vista concettuale, con synset si intende:

> Un insieme di sinonimi, che possono essere descritti da un'unica definizione perché esprimono lo stesso senso. Aggregando i sinonimi attorno al loro significato, non si ha la ridondanza che si avrebbe associando ad ogni termine tutti i suoi sinonimi.[12]

All'interno del progetto GEROM, questa definizione viene adattata per integrare, nella trattazione monolingue del lessico, anche una prospettiva sociolinguistica, necessaria a nostro avviso per affrontare discipline i cui concetti non sono, almeno sul piano teorico, così ben definiti, neutrali e univoci, come negli ambiti tecnici e scientifici e possono essere espressi da parole

12. Cfr. F. Bertaccini, M. Prandi, S. Rintuzzi, S. Togni, *op. cit.*

cariche di componenti emotive e connotative[13] (v. il primo capitolo per una serie di esempi riferiti all'italiano e al tedesco). I synset non raccolgono dunque soltanto i sinonimi in senso stretto, ma anche i cosiddetti quasi–sinonimi. Si tratta di termini il cui livello di sinonimia è tale da designare, all'interno di un determinato testo o gruppo di testi, lo stesso concetto, ma insufficiente a giustificare tanto una vera sinonimia contestuale, ovvero un'ampia gamma di contesti in cui l'interscambiabilità è possibile, quanto una vera sinonimia cognitiva, con l'azzeramento delle differenze fra le parole in termini di significati emotivi o valutativi a fronte del medesimo significato cognitivo. Poiché i synset non riuniscono solo parole legate da un vero rapporto di sinonimia, ma anche unità lessicali il cui livello di sinonimia potrà essere, in molti casi, assai parziale e comunque limitato al solo ambito di indagine, si è scelto di abbandonare la denominazione di sinonimo o quasi–sinonimo a favore della più generica "variante", con la quale si intenderà dunque una delle unità lessicali raggruppate in un synset.

Adottando un principio di organizzazione dei dati orientato al concetto e non al termine, le schede terminologiche sono così in grado di registrare denominazioni utilizzate per designare lo stesso concetto, ma che dal punto di vista semantico non è possibile considerare come veri sinonimi. Si pensi all'esempio sulle varianti "migrante / immigrato / clandestino", tre termini utilizzati dalle testate giornalistiche italiane per identificare, all'interno del dibattito sull'immigrazione, le persone che sbarcano a Lampedusa. Di questi tre lessemi non è possibile affermare che si tratti di tre veri sinonimi, come la consultazione di un qualsiasi dizionario della lingua italiana potrà facilmente confermare. Eppure, all'interno del dominio di analisi le tre parole vengono utilizzate per identificare lo stesso referente. Come è facile immaginare,

13. Cfr. M. Cerruti, *op. cit.* e F.W. Riggs, "Descriptive terminology in the social sciences". In: *Handbook of terminology management: Basic aspects of terminology management* a cura di S.E. Wright, G. Budin, Benjamins, Amsterdam 1997, pp. 184–196.

la scelta lessicale non è imparziale. Un'indagine preliminare sulla distribuzione dei lessemi in due corpora, bilanciati e paragonabili, rappresentativi di due quotidiani italiani di diversa estrazione politica, "La Repubblica" e "Il Giornale", rivela come il termine in assoluto più utilizzato per indicare chi sbarca a Lampedusa è "migrante" nel primo e "immigrato" nel secondo. Oltre alla connotazione più negativa che assume la parola "immigrato" rispetto a "migrante" (v. 3.9 per il ruolo delle collocazioni nel determinare il tipo di associazione legata ad una parola) è interessante notare come "Il Giornale" faccia un uso del 519,50 % più frequente rispetto a "La Repubblica" della parola "clandestino", parola dal significato molto diverso, e soprattutto più negativo, rispetto alle prime due. Analizzando un campione casuale di frasi dove compare questa unità lessicale, è possibile notare come le frasi siano apparentemente neutrali dal punto di vista argomentativo, poiché si limitano a descrivere dei fatti di cronaca. Da quest'esempio appare evidente che i quotidiani facciano un uso, più o meno consapevole e mirato, di determinati lessemi perché in grado di suscitare associazioni, positive o negative (raramente neutrali), in linea con l'orientamento della testata e dei suoi giornalisti (v. anche esempi bilingui contenuti nel primo capitolo del presente volume).

La registrazione delle varianti avviene in schede terminologiche basate sui synset rappresentate schematicamente nella figura 3. Come si evince, a livello intralinguistico ogni variante è provvista di un set di campi che la specificano e che permettono di differenziarla dalle altre varianti appartenenti allo stesso synset. A livello interlinguistico, l'approccio basato sui synset permette di superare la struttura simmetrica tipica delle banche dati terminologiche tradizionali che, in misura diversa in base al dominio che rappresentano, assumono come dominante il punto di vista della lingua di partenza, mentre non rappresentano correttamente, o solo parzialmente, il punto di vista della lingua di arrivo. Per valorizzare e riflettere correttamente la prospettiva di ogni lingua e cultura, Hudon propone lo sviluppo di

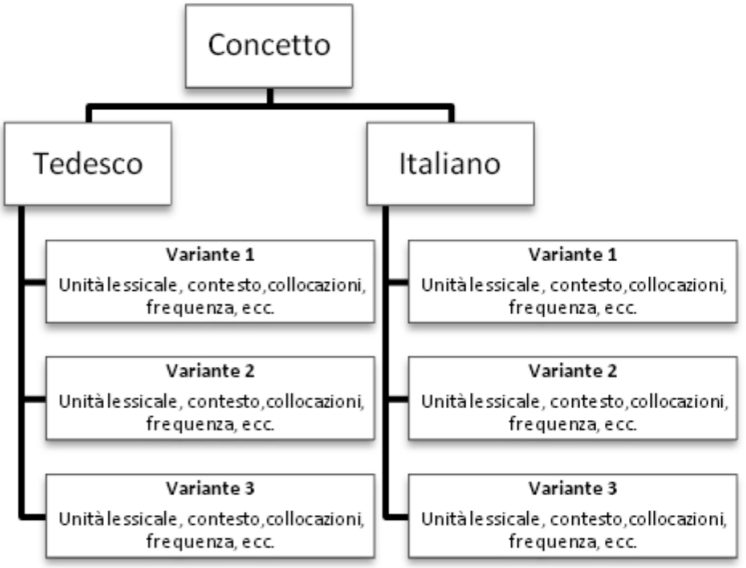

Figura 3.

strutture asimmetriche[14]. In queste strutture le diverse versioni linguistiche possono variare non solo nel numero di termini registrati, ma anche nella modalità con la quale vengono strutturati. Secondo questo principio, ogni lingua poggerebbe, a livello intralinguistico, su una struttura indipendente. I collegamenti fra le due lingue verrebbero poi costituiti fra queste strutture, garantendo così una rappresentazione più fedele delle diverse lingue e culture. Nella struttura a synset, questo ponte di collegamento è rappresentato dalla definizione del concetto registrato al suo interno. La definizione, per sua natura sempre dipendente dal dominio di cui essa farà parte, è priva, per quanto possibile, di ogni valutazione di tipo emotivo–valutativo e si concentra solo sulla rappresentazione metalinguistica del

14. M. Hudon, *Guide pratique pour l'élaboration d'un thésaurus documentaire*, ASTED, Montréal 2009.

suo referente. Della dimensione emotivo–valutativa, particolarmente interessante nell'esempio sopra riportato, se ne faranno invece carico le informazioni registrate nei campi della scheda, legati quindi alla singola realizzazione linguistica (variante), insieme a quelle recuperate dinamicamente dai corpora associati al repertorio terminologico (v. ad esempio il paragrafo 5 per le analisi statistiche all'interno dei corpora legati ai singoli repertori terminologici).

L'organizzazione delle parole in strutture basate sui synset può rappresentare una soluzione innovativa per integrare l'approccio onomasiologico, ad appannaggio quasi esclusivo della terminografia, con l'approccio semasiologico, tipico invece della lessicografia. Grazie all'elevata flessibilità della struttura appena introdotta, esso sembra inoltre prestarsi bene a gestire in un unico strumento di registrazione e consultazione lessicale termini afferenti a linguaggi specialistici, tendenzialmente monoreferenziali, condivisi e privi di connotazioni, con parole utilizzate in contesti meno speciali, carichi magari della dimensione emotivo–valutativa di cui sopra.

3.2. *Contesto*

Per ogni variante inserita nella scheda, il contesto rappresenta una porzione di testo in cui compare il termine in esame. Poiché si tratta di una concreta realizzazione linguistica del concetto registrato nel synset, il contesto permette di cogliere il significato e soprattutto le possibili connotazioni e le sfumature di significato del termine al quale fa riferimento. La selezione del contesto d'uso avviene a discrezione del compilatore della scheda, in base a valutazioni quali la rilevanza dell'esempio per il dominio trattato, la pertinenza rispetto al concetto rappresentato nella scheda e, qualora applicabili, la tipicità in termini di elementi valutativi e connotativi.

Sebbene la piattaforma supporti l'estrazione di un elevato numero di esempi contestualmente alla fase di consultazione della scheda, questo campo continua a mantenere una certa rile-

vanza nella struttura complessiva. L'esempio in esso contenuto, infatti, in virtù della sua oculata selezione, offre all'utente la possibilità di avvicinarsi in modo veloce e mirato al significato denotativo e connotativo che il termine assume all'interno del dominio di studio. Qualora l'utente fosse invece interessato a un'analisi più approfondita, potrà ricorrere ai numerosi esempi d'uso, sotto forma di concordanze, estratti direttamente dal corpus associato al glossario e dal corpus generale. I dati disponibili nel quadro statistico, come ad esempio la distribuzione del termine nel corpus generale e nelle sue subcomponenti (in base alla fonte, alla tipologia testuale, ecc.), completano l'offerta informativa riguardante la contestualizzazione del termine all'interno dei testi analizzati, come auspicato nel paragrafo 1.4 del presente volume.

3.3. Fonte contesto

Questo campo contiene la fonte del contesto d'uso selezionato e inserito nella scheda. Per una maggiore interattività della piattaforma, le indicazioni della fonte, qualora liberamente accessibile sul Web, prevedono l'esplicitazione del link ipertestuale per poter accedere direttamente al documento originale.

3.4. Categoria grammaticale

Il campo "Categoria grammaticale" fornisce indicazioni sulla classe grammaticale del termine, ovvero se questo è un sostantivo, un aggettivo o un verbo. Tale informazione risulta particolarmente utile per disambiguare casi di omonimia grammaticale, come ad esempio per la parola "cancerogeno", parola che, comunemente utilizzata come aggettivo, ricorre in alcuni contesti e tipologie testuali anche come sostantivo, solitamente al plurale, come illustra il seguente esempio:

> In maniera più o meno esplicita, la permanenza della spazzatura nelle strade e i roghi di immondizia sono stati associati all'inciden-

> za di alcuni tumori più alta rispetto alla media nazionale, come se l'esposizione a cancerogeni anche molto potenti per qualche settimana fosse sufficiente a far sviluppare istantaneamente la malattia. La realtà, come sempre, è assai più complessa.[15]

I sintagmi vengono ricondotti alla testa, quindi verbo per il sintagma verbale ("ingranare la marcia"), sostantivo per il sintagma nominale ("tutela ambientale"), aggettivo per il sintagma aggettivale ("blu marino"), ecc.

3.5. *Definizione*

La definizione è un enunciato che descrive linguisticamente il concetto registrato nella scheda terminologica permettendo di distinguerlo dagli altri. Essa viene tipicamente estratta da enciclopedie, dizionari o documenti ufficiali. In alternativa può essere redatta direttamente dal compilatore della scheda consultando, ogni qualvolta si riveli necessario, uno specialista della disciplina oppure può essere estratta dai testi in cui gli stessi termini sono contenuti[16].

La definizione, priva per quanto possibile di ogni valutazione di tipo emotivo–valutativo, si focalizza solo sulla rappresentazione metalinguistica del suo referente. Il ruolo che la definizione gioca all'interno della scheda è centrale, poiché essa risulta indispensabile per giustificare il raggruppamento in un unico synset di lessemi che non sono legati da un vero rapporto di sinonimia, forte ed esplicito, ma che tuttavia all'interno del dominio di cui esse fanno parte rimandano allo stesso referente.

15. Fonte: *Rifiuti e tumori, un legame tutto da chiarire*, "Corriere della Sera" (18/01/2008).

16. Per un approfondimento sui diversi tipi di definizione, ad esempio linguistica, ontologica o terminologica, si rimanda a M.T. Cabré, *Terminology. Theory, methods and applications*, Benjamins, Amsterdam 1999.

3.6. *Fonte definizione*

Questo campo contiene la fonte, preferibilmente ufficiale laddove disponibile, della definizione del referente registrato nella scheda.

3.7. *Direzione traduttiva*

In questo campo viene registrata, ogni qualvolta ritenuto indispensabile, la direzione linguistica con cui la scheda è stata elaborata. Quando all'interno di una banca dati terminologica non si registrano solo lessemi di specialità, ma anche parole culturalmente connotate, è indispensabile identificare in modo chiaro e univoco il sistema di riferimento dal quale esse provengono. Nel caso di un lavoro basato sui corpora sarà necessario segnalare all'utente i corpora dai quali è stata estratta la parola ricercata e i corpora dai quali provengono i traducenti proposti, in altre parole in quale lingua è stata eseguita l'estrazione terminologica di partenza e in quale lingua quella per l'identificazione degli equivalenti linguistici.

Come avviene nella maggior parte delle banche dati terminologiche, la parola cercata dall'utente finale viene ricercata all'interno di tutti i glossari tematici che compongono la banca dati. Affinché possa produrre risultati utili, scongiurando possibili fonti di errore, la ricerca di un lessema non specializzato dovrebbe però essere effettuata solo fra quei termini identificati come appartenenti, al momento della compilazione della scheda, al sistema concettuale della lingua di partenza. Tuttavia, in una struttura dei dati orientata al concetto, dove la reversibilità della lingua di arrivo e di quella di partenza sono la norma, la ricerca verrà eseguita anche fra quei termini che rientravano, sempre al momento della compilazione, nel sistema concettuale della lingua di arrivo. Senza un opportuno dispositivo di segnalazione della direzione traduttiva, l'utente rischierebbe di ottenere proposte traduttive che in realtà sono legate solo alla lingua di partenza (quella in cui sta effettuando la ricerca)

e non alla lingua di arrivo[17]. Il termine italiano "Forconi", ad esempio, nella sua accezione di movimento di protesta nato in seguito alla recente crisi economica italiana, ha fra i traducenti tedeschi, estratti da un corpus tedesco contenente testi giornalistici riferiti all'attualità italiana, "Mistgabel–Bewegung"[18] e "Heugabel–Bewegung"[19]. In questo esempio indicare la direzione di lettura dell'equivalenza risulta indispensabile per segnalare all'utente che si tratta di concetti legati ad una realtà storico–sociale italiana e che i termini tedeschi sono dei traducenti di concetti riferiti ad una realtà non esistente in Germania, ottenuti nella fattispecie attraverso un calco. Senza un indicatore di direzionalità di questo tipo, una ricerca svolta in tedesco per il termine "Mistgabel" produrrebbe come traducente italiano "I Forconi". Questo risultato potrebbe indurre l'utente finale a pensare di trovarsi di fronte ad un termine relativo alla realtà tedesca. Scegliendo i traducenti italiani registrati nella scheda, l'utente rischierebbe dunque di operare una scelta traduttiva potenzialmente errata.

A differenza del classico approccio terminologico, che non prevede una particolare direzione di lettura delle schede, in un approccio che tende a conciliare aspetti lessicografici e terminografici[20], la direzione di lettura delle equivalenze non può dunque essere sempre reversibile. Essa dipenderà infatti da una serie di fattori: dal tema trattato, se questo è ad esempio riferito solo ad una delle due realtà linguistico–culturali; dalle caratteristiche dei testi raccolti nei corpora (ad es. testi tedeschi sul welfare tedesco vs. testi italiani sul welfare della Germania oppure testi italiani sul welfare dell'Italia) e, in termini più generali, dall'impostazione data al lavoro terminografico.

17. Questo è l'elemento caratterizzante delle opere lessicografiche, come i dizionari bilingui, per i quali non è possibile invertire l'ordine di lettura delle entrate.

18. Fonte: Deutschlandfunk

19. Fonte: Die Zeit

20. Per un maggior approfondimento, ad esempio, sul livello di descrittività o normatività delle raccolte terminografiche, vedi R. Arntz, F. Meyer, H. Picht, *Einführung in die Terminologiearbeit*, Olms, Hildesheim 2004, p. 193.

All'interno delle schede terminologiche un ruolo fondamentale nel disambiguare i sistemi concettuali ai quali i termini fanno riferimento viene tradizionalmente svolto dalla definizione. In molti casi questo tipo di disambiguazione può avvenire anche in base alla natura stessa del termine (legato in modo evidente solamente alla lingua di partenza, ovvero a quella nella quale si sta effettuando la ricerca) o in base alle equivalenze proposte dal terminologo (preferenza per parafrasi, spiegazioni, ecc., quindi tendenza all'esplicitazione e all'adattamento del concetto estraneo alla cultura della lingua di arrivo). Tuttavia, non è possibile garantire che quest'opera di disambiguazione avvenga sempre in modo implicito; si pensi ad esempio al caso in cui l'utente non è in grado di riconoscere a quale sistema concettuale il termine appartenga, per mancanza di conoscenze pregresse oppure a causa dell'invecchiamento di certi termini o locuzioni. In tutti questi casi l'esplicitazione della direzione traduttiva può rivelarsi particolarmente utile ed evitare traduzioni errate o poco appropriate.

3.8. *Tipo di traducente*

Insieme all'indicazione di direzionalità, il dato "Tipo di traducente" riveste un ruolo importante nel consentire all'utente finale di compiere una scelta ponderata tra le varie realizzazioni linguistiche del concetto registrato nella scheda. Utile in primis a chi è chiamato a ricercare un'equivalenza linguistica[21] da utilizzare nel processo traduttivo, questa informazione favorisce l'individuazione del traducente più idoneo alla situazione comunicativa e alla tipologia testuale affrontata. In particolare, laddove ritenuto utile, questo campo registrerà se l'equivalente traduttivo è un calco, un prestito, un neologismo o una parafrasi

21. Per una trattazione dettagliata dell'equivalenza nelle lingue speciali vedi R. Arntz, F. Mayer, H. Picht, *op. cit.*

Ad esempio, per rendere in italiano il concetto tedesco del "Mini–Job", una forma di contratto a tempo ridotto prevista dalla riforma del lavoro del 2003 (la cosiddetta riforma Hartz), fra i possibili traducenti riscontrabili nel corpus italiano dedicato al mercato del lavoro sono attestati una *spiegazione,* "part–time a regime ridotto", e un *prestito*, "Minijob". Interessante notare come quest'ultimo, già introdotto in Italia dieci anni prima per parlare della riforma tedesca, è ritornato nuovamente in auge nel dibattito politico, questa volta per riferirsi a fatti squisitamente italiani, ovvero alle riforme del mercato del lavoro proposte dal governo Renzi a fine 2014. Quest'ultima considerazione, inoltre, anch'essa utile per evitare la selezione di varianti che assumono nel corso del tempo nuovi usi, viene registrata nell'unico campo generico previsto nel modello di scheda chiamato "Commento", nel quale il compilatore può inserire valutazioni di varia natura sul lessema e il suo utilizzo.

3.9. *Collocazioni*

Questo campo contiene un elenco delle principali collocazioni individuate per il termine di ricerca e per tutte le varianti registrate nel synset corrispondente. Benché le collocazioni possano essere individuate dall'utente finale attraverso l'ordinamento alfabetico delle KWiC oppure in maniera automatica mediante calcolo statistico, la registrazione manuale e quindi supervisionata delle principali collocazioni costituisce un valore aggiunto per la scheda, perché la arricchisce di una serie di informazioni di elevata affidabilità particolarmente utili per l'utente. Questo valore può essere sintetizzato ricordando come la letteratura specializzata consideri le collocazioni indispensabili per la formazione del significato della parola e come, di conseguenza, la loro esamina possa contribuire a migliorarne non solo la comprensione a livello semantico, ma anche a livello

emotivo–valutativo[22]. Le collocazioni, infatti, come sottolineano ad esempio Gabrielatos e Baker, sono fortemente indicative delle idee e dei concetti più comunemente associati ad una determinata parola[23]. Se ad esempio le parole "immigrato" e "clandestino" compaiono frequentemente insieme, per effetto del fenomeno chiamato *priming* tenderemo a pensare ad un concetto anche senza la necessità che l'altro sia effettivamente presente in una determinata realizzazione testuale. Nel nostro esempio, ciò si traduce in un'associazione negativa che il lettore attribuirà alla parola "immigrato" anche non in presenza dell'aggettivo "clandestino"[24]. È dunque ipotizzabile pensare che nell'esempio riportato al paragrafo 3.1 la scelta lessicale "migrante" utilizzata con una frequenza elevata da un certo tipo di orientamento giornalistico sia proprio volta a "neutralizzare" la connotazione negativa associata alla parola "immigrato".

Oltre a contribuire in misura decisiva all'analisi di una determinata parola, avere a disposizione una serie di collocazioni può rivelarsi particolarmente utile in fase produttiva, cioè quando si è chiamati a scrivere o tradurre testi su un determinato tema. Le collocazioni infatti costituiscono in tutti gli ambiti, anche in quelli meno tecnici e scientifici, un elemento linguistico fondamentale per un parlante/scrivente intenzionato a produrre testi che siano da un lato idiomatici e dall'altro capaci di soddisfare le aspettative linguistiche del suo ascoltatore/lettore[25].

22. Cfr. J.R. Nattinger, J.S. DeCarrico *Lexical Phrases and Language Teaching*, Oxford University Press, Oxford 1992 e J. Sinclair *Corpus, Concordance, Collocation*, Oxford University Press, Oxford 1991, pp. 115–116.

23. C. Gabrielatos, P. Baker "Fleeing, Sneaking, Flooding. A Corpus Analysis of Discursive Constructions of Refugees and Asylum Seekers in the UK Press, 1996–2005", *Journal of English Linguistics*, vol. 36, n. 1 (2008), pp. 5–38.

24. M. Stubbs "Lexical Density: A Technique and Some Findings". In: *Talking about Text. Studies Presented to David Brazil on his Retirement*, a cura di M. Coulthard, English Language Research, Birmingham 1986, pp. 27–42.

25. C. Fantinuoli, "Einbindung von Korpora im Übersetzungsunterricht als Schlüssel zur Professionalisierung". In: *Projekte und Projektionen in der translatorischen Kompetenzentwicklung*, a cura di S. Hansen-Schirra, D. Kiraly, Lang, Frankfurt/M. 2013, pp. 173–188.

4. Selezione dei termini

La scelta dei termini da inserire nelle schede terminologiche avviene sulla base di un approccio di tipo quantitativo e qualitativo. Una prima lista di termini in entrambe le lingue si ottiene grazie alla funzione di estrazione terminologica e di conteggio delle frequenze offerte dal software TranslatorBank[26]. Il modulo di estrazione implementato è di tipo ibrido, integra cioè un approccio linguistico, con l'identificazione della struttura morfologica dei termini sulla base delle classi grammaticali più ricorrenti, e un approccio statistico, basato sul confronto delle distribuzioni di frequenza tra il corpus in esame e uno generale di grandi dimensioni. Dopo aver eseguito la tokenizzazione e l'arricchimento morfosintattico del corpus, il software estrae le combinazioni di classi grammaticali impostate, ad esempio sostantivo+preposizione+sostantivo per l'italiano, e la loro frequenza normalizzata[27]. Questa frequenza viene poi confronta con quella dello stesso termine all'interno del corpus generale. L'obiettivo di questo confronto è quello di identificare i termini che, ricorrendo proporzionalmente con una frequenza maggiore all'interno del corpus specializzato rispetto a quello generale, possano essere considerati "tipici" del primo[28].

Il workflow seguito nella fase di compilazione delle schede prevede, per ogni lingua, le seguenti fasi:

26. C. Fantinuoli, "Design and Development of a Freeware Text Analysis Tool for Translation Tasks". In: *Proceeding of the IV International Conference on Corpus Use and Learning to Translate*, Alicante (in stampa).

27. In statistica, e in particolare nella linguistica dei corpora, con *normalizzazione* si intende l'esigenza di riportare le grandezze riferite a due gruppi di diversa dimensione ad un valore che possa essere paragonato. Se ad esempio si volessero confrontare le occorrenze di un determinato lessema in un corpus A composto da 1.000.000 token e in un corpus B da 5.000.000 token, il valore di frequenza ottenuto nel corpus di maggiori dimensioni dovrebbe essere ridotto del *fattore di normalizzazione* pari al rapporto tra corpus A e corpus B.

28. Per un maggior approfondimento sugli algoritmi utilizzati, vedi C. Fantinuoli "Computerlinguistik in der Dolmetschpraxis unter besonderer Berücksichtigung der Korpusanalyse", *Translation: Corpora, Computation, Cognition. Special Issue on Parallel Corpora: Annotation, Exploitation, Evaluation*, vol. 1, n. 1 (2011), pp. 45–74.

— estrazione automatica dei lessemi
— analisi manuale dei lessemi estratti automaticamente per eliminare quelli mal formati o poco interessanti
— integrazione manuale della lista con quei lessemi che, anche se non presenti nell'elenco originario, sono considerati importanti per il dominio di studio
— raggruppamento dei lessemi in synset
— estrazione delle collocazioni per tutte le varianti individuate.

Le strutture lessicografiche ottenute per le due lingue vengono poi collegate tra di loro creando un ponte interlinguistico mediato dal concetto e quindi basato sulla definizione elaborata per ogni singolo synset (v. 3.5).

5. Corpora

La piattaforma integra al suo interno una serie di corpora specializzati. I singoli corpora, ciascuno dei quali dedicato a un tema specifico, sono composti da due subcorpora, uno italiano e uno tedesco, caratterizzati da analoghi principi di design. Considerate le esigenze specifiche del tema di cui vogliono essere rappresentativi, i due subcorpora presentano una composizione analoga in termini di dimensioni (numero di testi/parole), tipologia testuale e fonti.

All'interno della piattaforma, i corpora svolgono una duplice funzione: da un lato fungono da base testuale per l'individuazione degli elementi lessicali tipici del dominio di interesse nella lingua di partenza e nella lingua d'arrivo (v. 3); dall'altro servono come base di dati per estendere in modo dinamico il livello di informatività delle schede terminologiche. Sia per il termine ricercato che per i traducenti verranno infatti estratte dai corpora informazioni quali il contesto d'uso, visualizzabili nella classica forma delle KWiC, e la distribuzione all'interno delle varie componenti dei suborpora (v. 5).

5.1. *Corpus Design*

Come spiegato in precedenza, a ogni raccolta terminologica è associato un corpus, bilingue e comparabile, con testi inerenti al tema trattato. Per garantire un elevato grado di comparabilità, le due componenti del corpus vengono raccolte seguendo la medesima procedura, ovvero utilizzando gli stessi principi di design in termini di tipologia testuale, di fonti e di dimensioni. Poiché il progetto GEROM ambisce a coprire un'ampia e variegata gamma di domini, tutte le variabili che incidono sulla composizione finale del corpus vengono stabilite di volta in volta, prestando attenzione a diversi fattori, quali la quantità di testi disponibili su un determinato tema e l'orientamento che si desidera dare al lavoro (un tema potrà essere ad esempio affrontato concentrandosi sul trattamento che gli è stato riservato dai mass media, dalle fonti istituzionali, dal settore economico–industriale, ecc.). In linea generale, l'obiettivo è quello di creare un corpus che rappresenti il più ampio spettro possibile di produzione testuale dedicata al tema in esame, includendo quindi numerose varietà di testi, sia in termini di tipologia testuale che di fonte. Solo partendo da un corpus che sia rappresentativo e bilanciato sarà infatti possibile compilare schede terminologiche capaci di descrivere in modo dettagliato e fedele la realtà lessicale di un determinato tema, senza penalizzare nessuna varietà. La scheda si potrà così arricchire di numerose realizzazioni linguistiche di un determinato concetto e diverrà rappresentativa della situazione linguistica che il lavoro terminografico vuole descrivere.

La selezione delle fonti avviene in modo tale da mantenere un equilibrio tra le varie componenti che formano il corpus, sia a livello monolingue che bilingue. In presenza di temi di natura tecnico–specialistica, la selezione dei testi avviene soprattutto sulla base di considerazioni di tipo qualitativo e di affidabilità, e i subcorpora comprenderanno testi tipici del dominino in esame: fonti legislative, manualistica, materiale divulgativo, ecc. A differenza dei temi di natura più specializzata, e laddove il

tema trattato lo richieda, ad esempio quando si affronta un tema socio–politico in grado di polarizzare l'opinione pubblica, le fonti, in primis quelle di carattere giornalistico, dovranno essere selezionate non solo in virtù di considerazioni legate alla tipologia testuale e all'affidabilità, ma anche in modo tale da garantire un equilibrio tra i vari orientamenti politico–ideologici esistenti. Anche in questo caso i subcorpora di entrambe le lingue dovranno essere costruiti in modo speculare.

Nella realizzazione del repertorio terminologico sul tema "Sbarchi a Lampedusa", ad esempio, si è scelto di creare un corpus contenente soltanto testi giornalistici, vista l'elevata copertura data a questo tema da parte dei mass media. La selezione delle fonti è stata operata con l'obiettivo di raggiungere un elevato grado di rappresentatività del panorama giornalistico delle due lingue. In un tema che polarizza a tal punto l'opinione pubblica, ambire ad un'elevata rappresentatività significa anche ottenere un corpus che sia bilanciato dal punto di vista degli orientamenti politico–ideologici in esso rappresentati, in modo tale da poter riflettere le differenze lessicali che questi comportano[29]. Nella fattispecie, per quanto riguarda il subcorpus italiano si è scelto di inserire i testi provenienti da quattro testate giornalistiche: il "Corriere della Sera", "La Repubblica", "Il Giornale" e "Il Fatto Quotidiano", rappresentativi di linee editoriali più conservative, progressiste o alternative. Per ottenere un subcorpus tedesco comparabile a quello italiano, oltre a selezionare testi aventi lo stesso tema, sono state scelte delle testate tedesche il più possibile "speculari" a quelle italiane in termini di diffusione e di orientamento editoriale: "Frankfurter Allgemeine Zeitung", "Süddeutsche Zeitung", "Die Welt" e "Taz". Sebbene il confronto statistico possa avvenire anche tra corpora di diverse dimensioni provvedendo alla normalizzazione statistica dei valori, l'obiettivo in fase di costruzione dei

29. I concetti di rappresentatività e bilanciamento sono assai controversi nell'ambito della linguistica dei corpora anche in virtù dell'impossibilità nel raggiungerli. Durante la fase di design, l'obiettivo è quello di operare delle scelte, di per sé arbitrarie e quindi sempre sindacabili, che ambiscano ad avvicinarsi a questi ideali.

corpora è quello di mantenere un certo equilibrio anche in termini di dimensioni. Nell'esempio sopra riportato, il subcorpus italiano consta di 200 testi per un totale di 129.556 token, suddivisi in modo omogeneo tra le varie testate, mentre quello tedesco consta di 200 testi per un totale di 140.103 token.

5.2. *Creazione semiautomatica dei corpora*

La selezione dei testi che costituiscono i corpora specializzati comparabili può avvenire in modalità completamente manuale oppure in modalità semiautomatica. A differenza della raccolta manuale, più precisa in termini qualitativi, ma evidentemente più dispendiosa in termini di tempo, la modalità semiautomatica permette di raccogliere ingenti quantità di materiale testuale in poco tempo e, grazie a sistemi interattivi di controllo, pertinenti al tema trattato. Essa si basa su una prassi ormai consolidata sia nell'ambito dell'attività traduttiva professionale[30] sia nell'ambito della lessicografia multilingue[31], ovvero sulla creazione di corpora specializzati attraverso l'identificazione automatica sul Web e il download di testi inerenti il dominio di studio (v. 6).

All'interno del progetto le attività di raccolta semiautomatica dei testi avvengono mediante CorpusCreator, un programma gratuito per la creazione di corpora da Internet[32]. Il principio di funzionamento di questo software è simile a quello implementato da BootCat, uno dei primi programmi ideati per la creazione di corpora generali e specializzati dal Web[33]. Con

30. F. Zanettin, "Corpora in translation practice". In: *Language Resources for Translation Work and Research*, a cura di E. Yuste-Rodrigo, LREC Workshop Proceedings, Las Palmas 2002, pp. 10–14.

31. A. Ferraresi, S. Bernardini, G. Picci, M. Baroni, "Web Corpora for Bilingual Lexicography: A Pilot Study of English/French Collocation Extraction and Translation". In: *Using Corpora in Contrastive and Translation Studies*, a cura di R. Xiao, Cambridge Scholars, Newcastle 2010, pp. 337–359.

32. C. Fantinuoli, *op. cit.*, (in stampa).

33. M. Baroni, S. Bernardini, "BootCaT: Bootstrapping corpora and terms from the web". In: *Proceedings of the 4th Language Resources and Evaluation Conference*

esso condivide alcune caratteristiche di base, come il principio di selezione dei testi da scaricare mediante una serie di parole chiave e l'utilizzo di un motore di ricerca per la selezione delle pagine rilevanti. A differenza di quest'ultimo, CorpusCreator si contraddistingue per l'elevata semplicità d'uso. Con un'interfaccia e un principio di funzionamento simile a quello di un comune motore di ricerca, il software è particolarmente adatto anche ad un pubblico non esperto. Inoltre supporta la ricerca e l'elaborazione sia di materiale in formato HTML che PDF, i due principali formati usati per la disseminazione delle informazioni sul web. Alcune funzionalità, come la presenza di un filtro in grado di eliminare dalle pagine HTML i cosiddetti *boilerplate*[34] oppure le routine per ripulire e normalizzare i testi raccolti (correzione di accenti, sillabazione, ecc.), permettono di ottenere corpora sufficientemente grandi e puliti senza la necessità di intervenire con un'ulteriore fase di post–editing manuale.

I testi scaricati e preparati sono salvati in formato XML. Oltre al testo vero e proprio, questo formato permette di memorizzare anche meta–informazioni relative ai singoli testi come la fonte, la tipologia testuale e l'autore. Tali informazioni, inserite dal software in modo automatico o con procedura guidata, serviranno da un lato per fornire indicazioni sulla fonte delle KWiC visualizzate, permettendo così all'utente di accedere direttamente al testo originale, e dall'altro fungeranno come base per il computo delle statistiche di cui si parlerà nel paragrafo successivo.

Come spiegato in precedenza, la procedura di creazione semiautomatica del corpus parte da una serie di parole chiave selezionate dall'utente in virtù della loro tipicità per il dominio preso in esame. Per ridurre il problema della distorsione[35] del

(LREC), Lisbona 2004, pp. 1313–1316.

34. Con *boilerplate* si intendono quelle parti testuali di una pagina web non interessanti o addirittura controproducenti per la creazione di un corpus, come ad esempio i menu o i pulsanti.

35. Con distorsione (in inglese *bias*) si intende la tendenza di un corpus, e quindi

campione di testi raccolti, le parole chiave devono essere "neutre" e non "marcate", non devono ad esempio presentare alcun tipo di connotazione che potrebbe influenzare la selezione dei testi e quindi tutte le analisi compiute a valle, dall'estrazione dei termini tipici del dominio all'identificazione dei loro traducenti. Nel progetto dedicato all'immigrazione e in particolare agli sbarchi di Lampedusa (v. 3.1), le parole chiave scelte per individuare i testi sono state "Lampedusa" e "Immigrazione" per l'italiano e "Lampedusa" e "Einwanderung" per il tedesco. In questo modo si è potuto garantire da un lato la selezione di testi inerenti esclusivamente al tema in esame, dall'altro la scelta di parole relativamente neutre ha favorito la selezione di testi che non avessero un orientamento ideologico ben preciso, cosa che invece sarebbe accaduta se si fosse scelto come termine di ricerca "clandestino". In questo caso i testi avrebbero sì avuto a che fare con il tema oggetto dello studio, ma il corpus sarebbe stato distorto a favore di testi con un approccio critico rispetto al tema.

Poiché l'obiettivo è quello di ottenere due subcorpora che siano comparabili, i tratti caratteristici delle parole chiave scelte per la ricerca dei testi dovranno essere molto simili in entrambe le lingue. Il numero di testi e token di ogni subcorpus dovrà inoltre essere sufficientemente elevato per garantire un buon livello di rappresentatività del corpus. Se attualmente le dimensioni vengono stabilite a priori sulla base di considerazioni quali la disponibilità di materiale o la composizione del corpus, in futuro si auspica l'implementazione di un metodo di calcolo della rappresentatività[36].

dei risultati ottenibili da esso, ad essere influenzato dalla modalità e dai principi secondo i quali è stato creato. Poiché tutti i dati raccolti sono già distorti in un modo o nell'altro, così come per i parametri di rappresentatività e bilanciamento, l'obiettivo non è l'assenza di possibili distorsioni, ma la loro riduzione e il loro controllo.

36. G.C. Pastor, M. Seghiri, "Specialized Corpora for Translators: A Quantitative Method to Determine Representativeness", *Translation Journal*, vol. 11, n. 3 (2007).

6. Quadri statistici

I quadri statistici offrono all'utente una serie di informazioni integrative sulle varianti lessicali contenute nella scheda selezionata. Grazie alle informazioni di natura statistica estraibili dai corpora, in prima linea quelle di distribuzione del termine all'interno delle componenti che possono costituire un corpus (diverse fonti, generi testuali, ecc.), l'utente è in grado di eseguire un'analisi più particolareggiata del termine esaminato, estesa ad esempio alle implicazioni ideologiche e alle eventuali dimensioni emotivo–valutative associate a determinate realizzazioni linguistiche. Queste informazioni, oltre a consentire analisi lessicali particolarmente approfondite, consentiranno, qualora ci si trovi di fronte al compito di dover tradurre testi orali o scritti, di operare una scelta del traducente più consapevole e motivata. In particolare, qualora le schede presentino più varianti per esprimere un determinato concetto, le informazioni contenute nel quadro statistico potranno essere decisive per spiegare le eventuali differenze fra le varianti lessicali registrate nella scheda.

Per ogni variante registrata, i valori computati sono:

— Frequenza all'interno del corpus specializzato, ovvero quello tematico associato al progetto terminologico e creato contestualmente alla stesura delle schede
— Frequenza all'interno delle sottocomponenti del corpus, ad esempio in base al genere testuale (Lex/Press/Info/Fiction/Handbook), alla fonte, allo stato (originale/tradotto) o all'autore (uomo/donna e età)
— Frequenza all'interno del corpus di monitoraggio[37] associato al progetto terminologico, creato ad intervalli regolari, ad esempio ogni 5 anni, con la stessa metodologia e

37. Con corpus di monitoraggio si intende qui un corpus creato al momento dell'interrogazione della banca dati da parte dell'utente finale con gli stessi parametri e la stessa procedura utilizzata per la creazione del corpus di progetto.

gli stessi parametri utilizzati per il corpus primario
— Frequenza all'interno del corpus generale
— Frequenza sul web contestuale alla stesura della scheda
— Frequenza sul web contestuale alla consultazione della scheda

I dati sulla distribuzione consentono di avanzare una molteplicità di considerazioni, sia di tipo sincronico che di tipo diacronico. Dal punto di vista sincronico, le frequenze di distribuzione delle varianti lessicali, sia a livello monolingue che bilingue, permettono di contestualizzare l'uso di quella variante, in base a variabili quali l'autore, la tipologia testuale e la fonte, così da evidenziarne, ad esempio, connotazioni di tipo ideologico–culturale o emotivo–valutativo (prevalenza di una realizzazione linguistica in testate giornalistiche conservatrici piuttosto che progressiste).

Dal punto di vista diacronico, invece, poiché le schede terminologiche sono soggette ad un invecchiamento fisiologico, in particolare se vertono su temi legati all'attualità, il confronto tra la ricorrenza del termine al momento della stesura del progetto e al momento della sua interrogazione può fornire indicazioni sul grado di vitalità del termine ricercato. Questo confronto avviene confrontando la frequenza all'interno del corpus associato al progetto con quella del corpus di monitoraggio. Le differenze nelle frequenze dei termini riscontrate in questi due corpora comparabili permetterà di individuare una tendenza sulla diffusione del termine in due momenti diversi.

Per integrare questo metodo si ricorre all'interrogazione del web mediante un comune motore di ricerca. Anche in questo caso la prima interrogazione avviene al momento di realizzazione del repertorio terminologico e la seconda al momento della sua interrogazione[38]. L'obiettivo è sempre quello

38. Per altri studi sul lessico che utilizzano i risultati di motori di ricerca si veda ad es. T. Chklovski, P. Pantel, "VERBOCEAN: Mining the Web for Fine–Grained Semantic Verb Relations". In: *Proceedings of Conference on Empirical Methods in Natural*

di evidenziare eventuali variazioni della ricorrenza del termine in due finestre temporali diverse. Sebbene questo confronto presenti degli evidenti limiti dovuti da un lato all'affidabilità dei risultati offerti dai motori di ricerca (scarsa trasparenza dei metodi utilizzati per il computo dei risultati, cambiamenti di tali metodi nel corso degli anni, ecc.) e dall'altro all'impossibilità, o quantomeno alla difficoltà, di circoscrivere il conteggio a un dominio o a una tipologia testuale particolare (si pensi anche al semplice problema degli omografi). L'indice calcolato con questo metodo, combinato con quello ottenuto con il confronto diretto di due corpora comparabili descritto poc'anzi, può offre una semplice indicazione dello stato di diffusione del termine (e quindi di vitalità) in due diversi archi temporali. Dai primi test informali eseguiti si è potuto verificare che tale procedura, nonostante i già citati limiti, consente di ottenere un'indicazione utile per segnalare se la realizzazione linguistica registrata nella scheda sia potenzialmente invecchiata.

7. Il web come fonte di informazioni

Uno degli obiettivi principali del World Wide Web è quello di permettere agli utenti di accedere a una vasta gamma di informazioni e di condividerle. Negli ultimi decenni, il web è diventato un contenitore mondiale e multilingue di conoscenze, un vero e proprio «universal repository of human knowledge and culture, which has allowed unprecedented sharing of ideas and information in a scale never seen before»[39]. Nel web sono disponibili oggi milioni di documenti in tutte le principali lingue del mondo e virtualmente su ogni tema. Questa sua caratteristica lo ha reso particolarmente interessante, se non addirittura rivoluzionario, per il settore della traduzione, che

Language Processing, Barcellona 2004, pp. 33–40.

39. R. Baeza–Yates, B. Ribeiro–Neto, *Modern Information Retrieval: the concepts and technology behind search*, Longman Publishing, Boston 2011.

da sempre è dipeso dall'accesso a informazioni per colmare il gap linguistico ed extra–linguistico che normalmente esiste tra il produttore del testo originale e il traduttore. Anche in virtù dell'accessibilità e semplicità d'uso che lo caratterizza, la nascita del World Wide Web e l'evoluzione di Internet hanno influenzato così profondamente la traduzione da portare Zanettin a definirlo "the most familiar and user friendly environment for translators"[40].

I traduttori utilizzano la grande quantità di informazioni disponibili sul web come fonte per approfondire il tema trattato nel testo che sono chiamati a tradurre, sia da un punto di vista nozionistico che linguistico–terminologico. I motori di ricerca vengono in questa pratica comunemente utilizzati come un *corpus surrogate* per confermare o scartare delle ipotesi traduttive, ad esempio ricercando l'uso di un termine o di un'espressione e verificandone la sua esistenza o la sua frequenza d'uso[41]. Tale forma di sfruttamento del web, benché rappresenti la modalità di uso più diffusa tra i traduttori, presenta tuttavia diversi limiti, come l'impossibilità, o quantomeno la difficoltà, di eseguire valutazioni quantitative, di restringere le ricerche solo al dominio o ai generi testuali pertinenti oppure di eseguire analisi linguisticamente utili. Per ovviare a questi limiti, i motori di ricerca possono essere utilizzati anche come *corpus shop*, ovvero come punti di accesso per recuperare documenti relativi al tema da tradurre, spesso chiamati impropriamente testi paralleli[42], e per costruire i cosiddetti *disposable corpora*[43], corpora creati

40. F. Zanettin, *op. cit.*, 2002, p. 12.

41. M. Baroni, S. Bernardini, S. Evert "A WaCky Introduction Wacky!". In: *Working papers on the Web as Corpus*, a cura di M. Baroni, S. Evert, GEDIT, Bologna 2006, pp. 9–40.

42. Ovvero "target language texts on the same subject matter, belonging to the same genre as the source text", C. Nord, *Text Analysis in Translation. Theory, Methodology, and Didactic Application of a Model for Translation–oriented Text Analysis*, Rodopi, Amsterdam 2005, p 171.

43. K. Varantola, "Translators and Disposable Corpora". In: *Corpora in Translation Education*, a cura di F. Zanettin, S. Bernardini, D. Stewart, St. Jerome, Manchester 2003, pp. 55–70.

ad–hoc per un particolare progetto traduttivo. Ed è proprio quest'ultimo modo di sfruttare il web, in grado di unire i suoi vantaggi, prima descritti, alle possibilità offerte dalla linguistica dei corpora, quello adottato all'interno del progetto GEROM (v. 4.2).

Oltre alla grande mole di testi che permette di consultare, il web è particolarmente utile per i traduttori anche per il numero di risorse linguistiche precompilate, liberamente disponibili e aggiornate, come glossari, tesauri ed enciclopedie che il traduttore può avere a portata di mano. Nonostante alcuni limiti di queste banche dati, ad esempio il mancato controllo di qualità dei traducenti proposti, soprattutto se si tratta di piattaforme gestite con il contributo di volontari (principio di Wikipedia), questo tipo di risorse è in grado di fornire ai traduttori ottimi spunti con i quali avvicinarsi alla risoluzione del problema traduttivo affrontato. Messe in rete da organizzazioni internazionali, oppure frutto del lavoro collaborativo degli utenti del web, molte di queste fonti sono gratuitamente accessibili e i loro dati, se impiegati per scopi non commerciali, riutilizzabili anche in altre applicazioni. È dunque plausibile pensare che, proprio in virtù dell'abbondanza e dell'attualità delle informazioni disponibili, il web si presti ad integrare, in modo dinamico e contestuale all'utilizzo della banca dati, le informazioni registrate staticamente nelle schede terminologiche.

Fra le informazioni potenzialmente integrabili è possibile annoverare le seguenti:

— Definizioni
— Proposte traduttive
— Immagini
— Key Words in Context estratte dal web
— Frequenze web dei lessemi

Con un'opportuna selezione delle fonti, ad esempio Wikipe-

dia[44] per le definizioni e Wiktionary[45] per i possibili traducenti, si potrebbe da un lato estendere in modo automatico il carico informativo delle schede terminologiche, dall'altro assicurare anche la presenza di informazioni più attuali rispetto a quanto registrato nelle schede stesse.

I limiti di un processo non controllato come questo devono tuttavia essere ben chiari agli utenti finali: nonostante una scelta accurata delle fonti risulta impossibile, proprio in virtù del processo di recupero non supervisionato, garantire l'elevata affidabilità delle informazioni visualizzate a cui un progetto come GEROM vuole puntare. Questo problema viene aggravato da fenomeni quali la sinonimia, l'omografia, l'impossibilità di restringere il dominio di interesse e la tipologia testuale che potrebbero produrre risultati non pertinenti alla scheda in esame. La consapevolezza dei limiti di questo approccio rende dunque necessario implementare un'interfaccia grafica che separi, in modo chiaro e preciso, le informazioni sicure e affidabili, perché frutto di elaborazione e controllo umano, da quelle potenzialmente errate perché generate dinamicamente.

8. Interfaccia grafica e workflow

Dopo aver illustrato nei precedenti paragrafi le funzionalità della piattaforma e il tipo di informazioni che si desidera visualizzare, si passerà ora a descrivere l'interfaccia grafica implementata e i principi di funzionamento.

Cosa rende un'interfaccia uomo–macchina efficiente? Shneiderman afferma:

> Well designed, effective computer systems generate positive feelings of success, competence, mastery, and clarity in the user community. When an interactive system is well–designed, the interface almost

44. www.wikipedia.org.
45. www.wiktionary.org.

> disappears, enabling users to concentrate on their work, exploration, or pleasure.[46]

Per raggiungere questo obiettivo, la piattaforma organizza le informazioni con un sistema modulare composto da quadri tematici visualizzabili in base al profilo selezionato. La struttura, identica e speculare per le due lingue, organizza le informazioni nei riquadri schematizzati nella seguente figura.

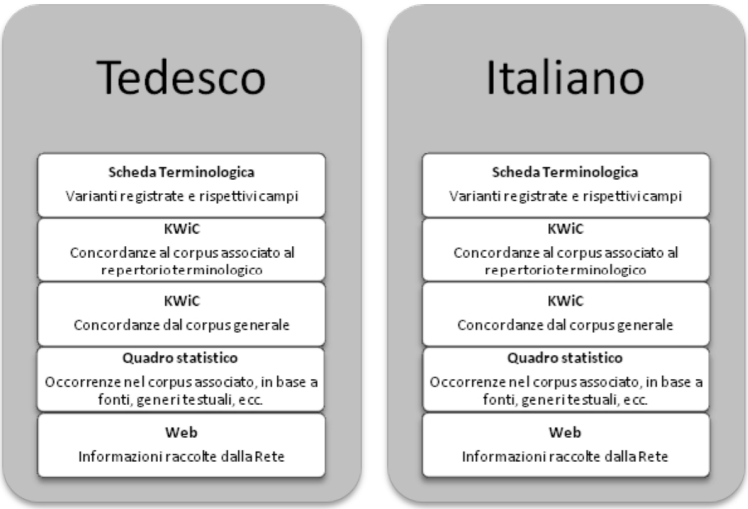

Figura 4.

Il primo riquadro è quello riservato alle informazioni registrate all'interno della scheda terminologica. Si tratta del riquadro principale, poiché contiene le informazioni frutto di un attento e supervisionato lavoro terminografico e quindi caratterizzate da un elevato grado di affidabilità. Al suo interno trovano spazio informazioni quali le realizzazioni linguistiche del concetto, le definizioni e il contesto (v. 3).

46. B. Shneiderman, C. Plaisant, *Designing the User Interface: Strategies for Effective Human–computer Interaction*, Longman Publishing, Boston 1997, p. 10.

Il secondo e il terzo riquadro contengono le concordanze del termine ricercato e delle sue eventuali varianti, estratte sia dal corpus specializzato associato al repertorio terminologico, quindi strettamente rilevanti per il campo di indagine, sia dal corpus generale, utile per analizzare il lessema indipendentemente da un contesto specifico. L'uso di due riquadri separati permette di separare le KWiC in base ai corpora dai quali sono state prelevate. Per ogni concordanza possono essere inoltre visualizzate le meta–informazioni registrate all'interno del corpus, come fonte, autore e tipologia testuale, e, qualora disponibile, l'URL dal quale è stato scaricato il documento originale, così da permettere all'utente di accedervi direttamente. Poiché gli URL registrati sono inevitabilmente destinati a diventare obsoleti con il passare del tempo e poiché i documenti originali per motivi di copyright[47] non possono essere memorizzati e resi disponibili nella loro interezza all'interno della piattaforma, questo riquadro prevede la possibilità di visualizzare il contesto immediato della concordanza, salvato direttamente all'interno del corpus, limitandolo tuttavia a poche frasi prima e dopo la stessa. Alle informazioni statistiche è invece dedicato il terzo riquadro. Il quarto e ultimo contiene infine le informazioni raccolte dal web, come definizioni, proposte traduttive o immagini.

Fra i principi fondamentali che governano il design di un'interfaccia utenti, Shneiderman sottolinea la necessità di fornire visualizzazioni alternative per gli utenti esperti e per i non esperti[48]. Le interfacce semplici sono facili da apprendere e immediatamente accessibili, ma sono in genere meno efficienti e flessibili; quelle avanzate, dal canto loro, permettono agli utenti esperti di ottenere risultati più approfonditi e di avere un maggiore controllo sul comportamento dell'interfaccia, cosa

47. Per maggiori informazioni sul tema del copyright nell'ambito della linguistica dei corpora cfr. F. Zanettin *Translation–Driven Corpora*, St. Jerome, Manchester 2012, pp. 52–55.

48. B. Shneiderman, C. Plaisant, *op. cit.*

che avviene tuttavia a scapito dell'immediatezza e della semplicità d'uso, soprattutto se l'interfaccia viene utilizzata solo con frequenza sporadica. In linea con questo principio, la soluzione adottata all'interno del progetto GEROM è quella di offrire all'utente non esperto un'interfaccia semplice, il cui utilizzo possa essere appreso velocemente e che fornisca solo le funzionalità di base, ad esempio la semplice ricerca del termine desiderato e la visualizzazione dei contenuti della scheda terminologica, in primis dei traducenti. L'utente esperto avrà invece a sua disposizione maggiori opzioni di ricerca, ad esempio la ricerca per parole complete, flesse, ecc., e la visualizzazione di più riquadri, in base alle proprie esigenze specifiche.

Il workflow per la ricerca prevede la digitazione della parola ricercata dall'utente nell'apposito campo di ricerca. Per rendere l'interrogazione un processo dinamico ed evitare che vengano immessi termini non presenti nella banca dati, l'applicazione esegue per ogni carattere digitato un confronto in tempo reale con il database terminologico e, in presenza di corrispondenze, le visualizza in un menu a tendina al di sotto del campo di immissione. Da questo elenco, sempre più breve e preciso man mano che l'utente immetterà altre lettere della parola ricercata, l'utente seleziona la parola desiderata. La ricerca della parola digitata viene eseguita all'interno di tutti i synset, e più precisamente in tutti i campi "Variante" della lingua di partenza. Qualora nella banca dati vi siano più corrispondenze per la parola immessa dall'utente, ad esempio nel caso di parole presenti in più raccolte terminologiche, il sistema prevede la disambiguazione del dominio d'interesse per il quale si desidera visualizzare le informazioni. Individuata la scheda interessata, per ogni singola realizzazione linguistica (variante) presente al suo interno vengono estratte le relative KWiC, sia dal corpus specializzato associato alla scheda sia da quello generale. Per estrarre le frasi contenenti la parola immessa indipendentemente dalle variazioni con le quali la parola può presentarsi (singolare/plurale, caso, ecc.) all'interno dei testi, la ricerca delle concordanze avviene sulla base del lemma ed è *case insen-*

sitive. In presenza di più varianti ortografiche o morfologiche della stessa parola, il numero di KWiC visualizzate nel riquadro apposito viene distribuito in modo proporzionale fra tutte le varianti, ovvero ponderandolo sulla base delle frequenze di ricorrenza all'interno del corpus, così da rappresentare nel modo più fedele possibile la realtà linguistica di quella parola. Per migliorare il grado di rappresentatività delle concordanze, le KWiC visualizzate vengono diversificate in base alle fonti da cui provengono, anche qui in modo proporzionale alla loro distribuzione. Insieme alle KWiC, per ogni variante registrata nella scheda vengono inoltre computate le frequenze all'interno dei vari corpora e le informazioni statistiche basate sulle meta–informazioni, visualizzate sotto forma di valori assoluti e normalizzati nell'apposito riquadro.

9. Conclusioni

Il progetto GEROM ha tra i suoi obiettivi la realizzazione di un'innovativa opera di consultazione lessicale che superi la tradizionale dicotomia che contrappone l'approccio lessicografico a quello terminografico. Destinata ad accogliere sia repertori lessicali specialistici che raccolte afferenti ai più generici ambiti della cultura, della società e della politica, la piattaforma adotta un principio di organizzazione dei dati basato sui synset, caratteristica che le permette di mantenere un approccio incentrato sul concetto, senza perdere di vista anche l'impostazione spiccatamente descrittiva delle sue raccolte. Questo strumento di consultazione vuole distinguersi dalle banche dati terminologiche tradizionali non solo in virtù della strutturazione dinamica e interattiva dei dati visualizzati, ma anche per un'interfaccia utente che, accanto ai termini registrati nelle schede, preveda, in entrambe le lingue del progetto, l'integrazione di un'ampia gamma di informazioni dinamiche, come i contesti d'uso, le statistiche e le informazioni enciclopediche.

Nel presente contributo sono stati illustrati l'architettura della banca dati e ci si è soffermati sulle scelte metodologiche che stanno alla base della raccolta dei dati lessicali e testuali, della loro organizzazione e rappresentazione. Grazie a questa banca dati, traduttori, interpreti, giornalisti e studiosi di varie discipline potranno accedere a informazioni linguistiche che permetteranno loro di comprendere meglio il lessico utilizzato in una grande varietà di dimensioni, da quella puramente denotativa a quella emotiva–valutativa e ideologica. Il carattere enciclopedico della piattaforma, inoltre, consentirà agli utenti di inquadrare il lessico nei rispettivi contesti d'uso, consentendo non solo di eseguire analisi più approfondite nei testi di partenza, ma anche di operazionalizzare tali conoscenze nell'attività di produzione testuale, ad esempio in ambito traduttivo.

Claudio Fantinuoli