Compilazione Automatica di Dizionari

Abstract in English:

The methodology for the compilation of dictionaries is in a profound change. On one hand, people in a globalized world need increasingly communication skills in special purposes and especially foreign languages for their competitiveness. On the other hand modern communication relies heavily on electronic processes. This pressure from both the quantity and quality side of the communication process make the traditional way of looking up words in print dictionaries look inadequate. Truly obsolete seems in this setting the traditional method of collecting a myriad of lemmas, to be published after years in a print dictionary. In fact modern approaches rely heavily on corpora, computation and modern media.
To what point can this process of automatization be taken? **Is automatic dictionary compilation possible** at all? What are today's most advanced techniques and what can we expect in the next decade? And if print dictionaries become historic, what will future dictionaries look like?

How dictionaries could be compiled:

- compile a corpus, starting e.g. from a term
- extract terms, classify their quality, and add them to the seed entries
- start bootstrapping process: go search new texts with the seeds,...

The programme should be able to complete the dictionary on its own.

1. a) starting out with a text: example-based term-extraction
   b) starting out with a seed entry, a term
2. From the term to a definition through a pattern search (ex: GOOGLE's „define"-funtion) („is defined as ~", „ ~ is a")
3. Corpus compilation: The texts that contain definitions are taken in.
4. Text usefulness for the dictionary is assessed through the density of terms in the texts. Texts can also be automatically classified according to their domain.
5. Different definitions for different domains can be saved
6. Term-extraction on the new texts in the corpus lead to new terms: bootstrapping process to iteratively build up corpus-based terminology
7. With quantity and quality of the corpus growing, corpus hits can be favoured before external hits.
8. With parallel corpora, also translation data can be acquired.
9. The user can choose the class of sources for the term (e.g. first in the corpus, then in certain classified sites)
10. In corpus-based terminology, a simple link to the occurrence suffices for term description. Whenever the corpus is refreshed, the term-text-references can be recalculated automatically.
11. Corpus quality improvement: Clustering and similar techniques allow to identify aberrant and low quality texts. The corpus should weed out its texts regularly.

For the single tasks **programmes** are already available and have been proven in practice: Text-classification (language, domain, …), term-extraction, description with definitions, contexts, translations, frequency etc., corpus enrichment, automatic arrangement of data according to an alphabetical or topical paradigm (and even a semantic order is possible through sense disambiguation using the MI in large corpora), formatting and adaptation to different user-devices.

The **difficulties** to resolve are the accumulation of errors through the bootstrapping, the ambiguity of language, the lack of (free, accessible, compatible) digital resources, esp. of parallel corpora (ex. Lexalp), copyright issues and the non-compatibility of software and/or data (ex. Windows).

The frequency of a term in a local LSP corpus can be compared to its frequency in an external general language corpus (e.g.: German legal data from BISTRO with general German from Wortschatzprojekt Leipzig). Such a quantitative tool is technically perfectly possible and would only require a computer-linguist to match data categories and regulate cooperation of the databases. The terminographer or lexicographer comes in only at the level of interpretation of results. Additional quantitative data will influence future dictionaries: They will be more flexible (able to be connected to a variety of external quantitative data) intelligent (additional comparative information not available so far), up-to-date (as external data changes, the comparative results in the dictionary itself will change) and adaptive (hands-on facility in electronic dictionaries allow personalized ad hoc terminology).
The paradigm publisher (lexicographer/terminologist) – user will give way to more flexible user participation (ex. Wiki-projects).

The automatic enrichment of terminology is possible. The necessity of a manual/intellectual feedback-loop depends on quantity and quality of the seed material and the desired target quantity and quality. Dictionary compilation around top quality seed-terms is also already possible.
The electronic dictionary will be different from traditional dictionaries. It will interconnect resources, also of different type (different languages and dictionary types). Automatic lexicography will not be central storage of data, but connection of decentred data.
In the long run, dictionaries for the main purposes and the main languages will be online. Print products will become niche products for LSP, special language combinations, and diachronic research.

I. Introduzione:

Il metodo per la compilazione di dizionari sta cambiando. Il primo motivo è che la globalizzazione genera il bisogno di più dizionari, in più combinazioni di lingue, e di dizionari per la comunicazione specialistica. Il secondo motivo è che il dizionario elettronico, al contrario di quello cartaceo, si integra perfettamente nel processo comunicativo moderno.
Alla luce di questi cambiamenti sia quantitativi che qualitativi, il metodo tradizionale sembra sempre più anacronistico. Di fatto la terminografia moderna utilizza corpora, computer e comunicazione moderno.
Fino a che punto si può automatizzare? È fantascienza la “macchina del dizionario”, che basta attivare per ricevere risultati senza fatica, o si può già concepire? Quali caratteristiche avrà un dizionario elettronico compilato automaticamente?

II. Un nuovo metodo per compilare dizionari

Eccovi un esempio di compilazione automatica di dizionario. Mettiamo che si ha una terminologia descritta con lemma, definizione, settore e contesto, e tutto questo in due lingue. Il compito sarebbe di aggiornare e allargare questa banca dati, partendo da un nuovo testo elettronico. La colonna a sinistra descrive il procedimento, mentre nella colonna di destra si commenta la tecnica da utilizzare.

Procedimento	Commento
1. Serve un corpus, cioè testi elettronici ordinati secondo lingua e settore specifico.	L’acquisizione di corpora è ormai diventato quasi una routine.[1] L’identificazione della lingua riesce con pochissime eccezioni.[2] L’identificazione del settore funziona altrettanto bene.[3]

[1] Streiter&Voltmer, Document Classification for Corpus-based Legal Terminology, 8. International Conference of the International Academy of Linguistic Law, Iaşi, Romania, http://dev.eurac.edu:8080/autoren/publs/iasi/ : 11.2.2006. Fletcher, Facilitating the Compilation and Dissemination

2. Il corpus è la base per un'estrazione di termini, o di candidati di termini. Il risultato è la stringa del testo che è più probabilmente un nuovo termine.	Uno dei metodi per l'estrazione è l'estrazione di termini sulla base di esempi[4]: Il programma sceglie la stringa più simile ai lemmi già contenuti nella banca dati. Con questo metodo è rarissimo che il migliore "candidato di termine" non è un lemma valido. I metodi basati sulle regole grammaticali e sulla statistica funzionano altrettanto bene, ma hanno più presupposti.[5]
3. Un programma cerca nel corpus compilato in 1 una definizione per il lemma estratto in 2.	L'idea dietro un tale programma è semplice: si creano matrici di definizioni come: "un [LEMMA] è un", "definizione [LEMMA]:" Per [LEMMA] si inserisce il lemma ricercato. Una ricerca di definizioni è implementata in GOOGLE; basta ricercare "define:…". Funziona anche in lingua italiana: http://www.google.com/search?q=define:italia&defl=it .[6]
4. Un programma cerca nel corpus compilato in 1 tutti contesti per il lemma estratto in 2, li mette in ordine di pertinenza e salva la loro posizione.	Occorrono criteri per la pertinenza dei contesti. È possibile usare la similitudine dei contesti ai dati già contenuti nella banca dati: Più lemmi nel contesto, più vicino è considerato il contenuto. "Salvare la posizione" significa che il (con-)testo non viene copiato ma collegato tramite link. Così si evitano inconsistenze tra corpus e dizionario, si alleggerisce la gestione dei dati e i contesti e la loro ordine possono essere aggiornati con molto facilità.
5. Il processo ricomincia con 2 e continua finché si è raggiunto il numero richiesto di nuovi lemmi.	L'iterazione del processo è molto efficace, ma rischia di moltiplicare gli errori fatti, che diventano la base per i nuovi cicli. Sarebbe opportuno introdurre un controllo dei lemmi.
-	Questo procedimento è solo un esempio per web-based term mining.[7]
A questo punto manca la descrizione nella lingua d'arrivo. Chiaramente bisogna presentare l'informazione al sistema, ma anche qui si può automatizzare.	Idealmente si dispone di corpora paralleli.
Se esistono corpora paralleli in internet, il computer li può trovare, salvare, allineare ed utilizzare per identificare come vengono tradotti i lemmi aggiunti.	1. Esistono corpora presso le associazioni per la distribuzione delle risorse linguistiche (ELRA) ed altri istituti di ricerca che distribuiscono corpora paralleli gratuitamente.[8] 2. I testi forniti dall'UE costituiscono una ricca fonte di testi paralleli, dai quali si possono produrre corpora paralleli con l'aiuto di codifica ed analisi.[9] 3. Si possono trovare con un motore di ricerca: a) ricerca di un lemma, ristretto a siti che contengono la sigla della lingua, p.e. "orzo si-

of Ad-Hoc Web Corpora, in: Aston, Bernardini, Stewart, Papers from the 5. International Conference on Teaching and Language Corpora, Amsterdam, Benjamins, 2004, http://citeseer.ist.psu.edu/727859.html : 11.2.2006.

[2] Langer, Grenzen der Sprachenidentifizierung, Tagungsband KONVENS 2002, Saarbrücken, p. 99-106, http://www.cis.uni-muenchen.de/people/langer/veroeffentlichungen/grenzen_der_sprachenidentifizierung.pdf : 11.2.2006.

[3] Voltmer, Computerlinguistik für die Terminografie im Recht, Narr 2006, vedi capitolo 2.

[4] Streiter et al., Term Extraction for Ladin: An Example-based Approach, TALN 2003, Batz-sur-Mer.

[5] Streiter&Voltmer, "Example-based Term Extraction for Minority Languages: A case-study on Ladin" in: Proceedings of the International Conference on Corpus planning und Sociolinguistics, Urtijëi 2003, http://dev.eurac.edu:8080/autoren/publs/termex5.pdf .

[6] Walter, Computational Linguistic Support for Legal Ontology Construction, The Language and Law Conference Düsseldorf 2006, http://www.coli.uni-saarland.de/projects/corte/icail.pdf : 19/04/2006; Walter&Pinkal, Computerlinguistische Methoden für die Rechtsterminologie, DGfS-AG Sprache und Recht, Universität Bielefeld 2006, http://web.uni-frankfurt.de/fb10/rathert/forschung/pdfs/walter.pdf : 19/04/2006.

[7] Ciola&Ralli, Web-based term mining tra terminologie e memorie di traduzione, Seminario sulle memorie di traduzione, Roma, 30 settembre - 1 ottobre 2003, http://dev.eurac.edu:8080/autoren/publs/contributo-ciola-ralli.pdf : 11.2.2006.

[8] Per esempio http://logos.uio.no/opus/ : 11.2.2006.

[9] Sui strumenti informatici per l'acquisizione, la codifica e l'analisi di testi paralleli vedi: http://www.sitlec.unibo.it/lingue/ita/Ricerca/RicGiov/Bernardini2000.asp : 11.2.2006.

	te:.it/it“ in GOOGLE. Si trova per esempio http://www.hotel-leitner.it/de/221_kuchllkastl.html . b) si sostituisce la sigla della lingua con la sigla della lingua d’arrivo. Se la pagina esiste (nell’esempio http://www.hotel-leitner.it/de/221_kuchllkastl.html), si possono introdurre ulteriori controlli come una comparazione della struttura o del numero di caratteri nei due testi trovati. Questo metodo è effettivo quando utilizzato dal computer. Nell’esempio esistono italiano e tedesco, in altri casi ci sarà italiano ed inglese. Per combinazioni di lingue meno frequenti è poco probabile che ci siano testi paralleli, ma la richiesta è anche minore. Rimane il problema dell’affidabilità dei testi in internet. Bisogna caricare alcuni testi paralleli per trovare appoggio dalla statistica, che eliminerà le traduzioni sbagliate.
L’organizzazione dei corpora può tenere conto dell’affidabilità delle loro fonti, della loro pertinenza riguardo al settore e della qualità linguistica dei termini nel dizionario. In altre parole: più termini del dizionario si trovano in un testo, più interessante è.	Finora tutti testi del corpus erano considerati equivalenti. La qualità del processo e del risultato si può aumentare con una selezione preliminare del corpus: prima di estrarre un lemma, una definizione e un contesto si limita il corpo dei testi ai più “validi”. Un criterio formale di “valido” può essere la similitudine con i lemmi e contesti già presenti nella banca dati. (Vedi la descrizione sotto punto 4 per definire la pertinenza dei contesti.)
Visto che tutto il processo è automatico, la banca dati si può aggiornare automaticamente. Si possono escludere testi ormai vecchi e sostituirli con testi più nuovi.	Come menzionato in 4, il facile aggiornamento dei contesti è uno dei vantaggi di un dizionario moderno e dinamico. Per aggiornare i contesti, bisogna aggiornare il corpus e riavviare la ricerca di contesti. Il corpus cambia per forza quando scadono i diritti d’uso o quando un testo, specie una norma, è ufficialmente sostituito. Il corpus cambia anche crescendo. Se si salvano i vecchi link si crea una vista diacronica del dizionario. Qualcuno sostiene che un vantaggio dei dizionari stampati sarebbe che essi calcano il livello di conoscenza di un certo momento. L’aspetto diacronico non è invece limitato al dizionario stampato; anzi, la quantificazione del cambiamento è più facile in un dizionario elettronico. Ci possono essere vari parametri per scartare testi dal corpus: Testi troppo vecchi, testi non più usati dal dizionario, testi che gli utenti non vogliono più…
Tutte le definizioni e i contesti sono un rinvio verso un testo.	Ci sono programmi che accertano se un link funziona ancora. Se il testo del corpus è stato eliminato, il sistema può avviare il programma di cui sopra e trovare un altro testo. Il sistema funziona quindi non solo con un corpus interno, ma anche in collegamento con corpora esterni, il che riduce il costo di mantenimento.
La terminologia meno usata o meno integrata si può individuare automaticamente e, in extremis, scartare.	Conviene controllare non solo la qualità del corpus, ma anche dei lemmi. Analizzando l’uso del dizionario, i lemmi meno visti possono essere tolti dalla banca dati.
La parte del corpus meno integrata o fuorviante può esse-	Con il metodo c*lustering* si trovano i testi che, in funzione alla loro similitudine con il resto del corpus, sono i più “strani” o diversi.[10]

[10] „Computerlinguistik für die Terminografie im Recht“, G. Narr Verlag Tübingen, „Forum für Fachsprachen-Forschung“ (FFF), 2006 (in stampa), capitolo 2 (Fachgebietserkennung für Terminografen). Streiter&Voltmer “Les domaines du droit se reflètent-ils dans le langage juridique ?” in: Col-

re automaticamente individuata ed esclusa o sostituita.	Forse questi testi contengono anche lemmi e/o contesti della banca dati, e sfuggono quindi alle prova di qualità di cui sopra. Di conseguenza non basta sostituire un testo con uno più tipico per il settore specifico, ma il nuovo testo deve anche fornire il lemma o un nuovo contesto. Dopo un *clustering* bisogna quindi riavviare la compilazione di un corpus (vedi 1). Alla fine di questo processo è aumentata la qualità del dizionario.

Ecco passo per passo la possibile automatizzazione. I programmi che occorrono sono standard o tali che ogni linguista computazionale potrebbe programmare.[11] I compiti sono tutti quanti fattibili:

- Classificazione di testi (per lingua, per settore specifico,…)
- Estrazione di termini
- Inserimento automatico in una banca dati terminografica con definizioni, contesti, traduzioni, numero di occorrenze etc.
- Aggiornamento automatico di un corpus

Rimane il problema che un programma si incatena all'altro, perché:

- gli errori si potenziano nel corso del processo automatico,
- i programmi lavorano sulla base del segno e non del significato,[12]
- non sempre si trovano risorse digitali, e in specifico mancano spesso corpora paralleli,[13]
- il copyright di testi digitali rimane un problema,
- i programmi modulari devono collaborare, cioè l'output dell'uno costituisce senza controlli l'input per l'altro.[14]

Un dizionario compilato automaticamente ha invece non solo vantaggi nella produzione, ma anche nell'uso:

Vantaggi nella produzione	**Vantaggi nell'uso**
1. L'ordine del dizionario può essere alfabetico, tematico, secondo settore o significato,…	1. I dati compilati costituiscono la base per la composizione *ad hoc* di una pletora di dizionari. In altre parole, il dizionario diventa completamente dinamico.[15]
2. Formattazione ed adattazione all'utente e il suo dispositivo di output	2. Una rappresentazione adattiva è necessaria per certi gruppi di utenti (non udenti o vedenti, persone con difficoltà di concentrazione o di lettura, persone disabili ed inesperti) e per la varietà dei dispositivi di output (browser nuovo o vecchio; collegamento lento o veloce; visualizzazione, vocalizzazione o rappresentazione a rilievo per non vedenti; rappresentazione in PDA,

loque International "Interpréter et Traduire", Centre d'Etudes et de Recherches sur les Contentieux (C.E.R.C), Faculté de Droit de Toulon, Editore Bruylant 2006 (in stampa).

[11] Alcuni programmi liberi: Per l'estrazione di termini: http://dev.eurac.edu:8080/perl/all.tar.gz . Per calcolare la similitudine sulla base di n-caratteri serve il pacchetto statistico per PERL: http://www.d.umn.edu/~tpederse/nsp.html . Per compilare un corpus: Open Source Robot wget : http://www.gnu.org/manual/wget .

[12] I programmi rischiano quindi di sbagliare con omonimi. Servirebbe una disambiguazione con informazione mutua da un corpus molto grande.

[13] Per esempio nel progetto LEXALP che compila terminologie in quattro lingue (F, D, I, SLO) http://www.alpinespace.org/approved-projects+M570e93bf0a6.html .

[14] Chi ha avuto il problema di programmi incompatibili p.e. in Windows ha un'idea dell'estensione del problema.

[15] Ralli, Ties, Streiter, Voltmer "BISTRO, the online platform for terminology management: Structuring terminology without entry structures"in: R. Temmermann e U. Knops (ed.), The Translation of Domain Specific Languages and Multilingual Terminology Management, Linguistica Antverpiensia New Series, Hoger Instituut voor Vertalers en Tolken, Hogeschool Antwerpen 3/2004, 203-215.

	WAP, o *touch screen*).
3. I dati di un dizionario completamente elettronico costituiscono una risorsa modulare per altre banche dati.	3. I dati non formattati servono invece nel caso di collegamento diretto con altre banche dati.

Lo scopo di questo intervento non è quello di dire che la compilazione di dizionari sarà fatta così, ma quello di dire che gli strumenti per una quasi totale automatizzazione ci sono già e di far vedere che questo comporta una trasformazione del prodotto "dizionario" stesso. La lessicografia automatica non sarà più la raccolta dei pezzi d'informazione che servono per un certo prodotto predefinito, ma un processo continuo verso un collegamento di dati decentrali.[16]
L'uso di un dizionario dinamico è molto più ampio e si lascia estendere in qualsiasi momento. Se a un certo punto interessa la frequenza di un termine, in un dizionario digitale si può generare: la frequenza nel corpus del settore è comparata con la frequenza in un corpus di lingua non settoriale, p.es. la frequenza di una parola giuridica nel corpus della banca dati BISTRO viene comparata con la sua frequenza nella banca dati del *Wortschatzprojekt Leipzig*[17]. Questa operazione può diventare una nuova informazione per tutti lemmi della banca dati; basta scrivere un programma che collega le due banche dati, con vantaggi per entrambi progetti. Il terminografo non serve più per compilare i dati, ma per definire il compito e per interpretare i risultati.
Questo stimola la ricerca quantitativa. Nasceranno sempre nuove utilità per scopi diversi. Se già adesso ci sono tantissimi dizionari di utilità diversa (rimario, dizionario storico, dialettale,…). Nel futuro dizionario tutto sarà fattibile, anche se certamente non tutto sarà utile. Proprio per tenere conto di questo serve la flessibilità. Aumenterà l'informazione contrastiva, l'attualità dell'informazione (appena le risorse decentrali cambiano sono già integrati nel dizionario collegato) e l'eterogeneità degli utenti (con l'output adattivo).
Infine il dizionario elettronico si presta anche all'integrazione dell'utente (progetti *Wiki*), aprendo un dialogo laddove fin'ora regnava un monologo patriarcale.

[16] Projekt Digitales Wörterbuch, http://www.dwds.de/ueber : 19.2.2006.
[17] http://wortschatz.uni-leipzig.de/ : 15/03/2006.