Traduzione Multilingue: Atamiri e... dintorni

di Leo Sorge

La traduzione veloce è un grosso problema. Dato che l'attuale mondo occidentale vede un monopolio culturale basato sulla lingua inglese, il trasferimento di tecnologia in nazioni del terzo e quarto mondo (America Latina, Paesi Arabi, Africa) è impossibile per la quasi completa mancanza di persone che conoscano la lingua inglese usata nelle documentazioni. Un altro campo d'applicazione è nel settore amministrativo per i contatti tra nazioni di lingua diversa (ONU, FAO, IBI, EC), ove è fondamentale — e al momento assai costoso — avere immediatamente disponibili in tante lingue le varie circolari.

Il primo metodo usato fu ovviamente basato sul solo lavoro umano: un traduttore esperto, con alti costi, realizza una versione pressoché definitiva alla velocità di 5-7 pagine al giorno per uomo. Il collo di bottiglia che limita la velocità è la presenza di termini tecnici, per i quali il traduttore deve effettuare una ricerca, spesso lunga e talvolta fastidiosa.

Si pensò quindi di affidarsi almeno in parte ai computer. Lo schema logico seguito in questo caso analizza le lingue a coppie, derivando un'analisi diretta della diversa costruzione del periodo (suddivisa in grammatica, sintassi e semantica) e attingendo ad un opportuno vocabolario. Il testo, elaborato, veniva tradotto in una prima forma, assai grezza ma generalmente comprensibile nei punti cardine, che poi andava rielaborata da un traduttore esperto per avere una versione definitiva. I grossi vantaggi sono nel tempo, che viene ridotto d'un fattore circa 10, e nei costi, sostanzialmente dovuti al solo tempo impiegato dal traduttore (più ammortamento e manutenzione del computer).

Dal punto di vista del metodo, però, questo sistema presenta almeno due grossi svantaggi:

1) la traduzione è in un solo verso, cioè da una lingua all'altra, ma non viceversa;

2) ogni volta che bisogna introdurre un nuovo linguaggio si devono fare 2 programmi per ciascuna lingua già usata (uno per ogni verso).

Il problema è stato affrontato qualche anno fa dalla Comunità Europea, che avendo all'epoca (1982) sette membri aveva bisogno di un sistema che potesse gestire la bellezza di 42 coppie di lingue: la formula che ci dà il C, il numero delle coppie, a partire da N, il numero delle lingue, è infatti:

$C = N \cdot (N-1)$.

È evidente che questo numero va moltiplicato per due, in entrambi i versi. La ricerca della E.C. aveva quindi come obiettivo un qualcosa che era composto da 84 diversi programmi!

Sulla scorta del Systran, un sistema già funzionante presso la stessa E.C., basato su tre coppie di linguaggi (Inglese-Francese, Francese-Inglese ed Inglese-Italiano), un team di ricercatori provenienti da 11 università europee modellarono il progetto Eurotra, capace di gestire le famose 42 coppie. Se Systran serviva per 12.000 pagine l'anno, con una velocità massima (per utenti esperti) di 25 pagine al giorno, Eurotra permetteva di arrivare fino a 50 pagine al giorno: un grande miglioramento.

Il progetto, partito nel 1982, aveva un budget di circa 16 milioni di sterline, che con il cambio di allora (2.400 lire, contro le 1.950 attuali) fa circa 38 miliardi di lire, e sarebbe durato 6 anni. Quando poi Spagna e Portogallo si sono uniti alla Comunità le coppie di linguaggi sono diventate 72, i miliardi 60 e gli anni 7,5. Di questo passo, il più serio ostacolo per l'Europa Unita sarà il sistema di traduzione! I Paesi occidentali usano infatti 14 lingue nazionali, escludendo nazioni come Islanda e Jugoslavia (praticamente convertita al capitalismo), e svariate lingue più o meno importanti come il catalano, il basco, il celtico, il gaelico, il ladino etc. ect.

In un articolo del novembrre 1985, il London New Scientist — una testata nota negli ambienti della ricerca d'oltremanica — concludeva un articolo precisando che Eurotra era all'epoca il più ambizioso progetto di traduzione automatica mai concepito. Ma i due problemi puntualizzati poco fa mostrano chiaramente che questa non è una strada praticabile per tante lingue.

Il sistema Atamiri

In spagnolo, lingua ufficiale in Bolivia, ATAMIRI è l'acronimo di Automata Traductor Algoritmico Multilingue Interactivo Recursivo Inteligente, espressioni che non necessitano di traduzione. Questa definizione è stata creata ad hoc, poiché Atamiri in Aymara vuol dire... interprete.

Atamiri si basa su più parti:

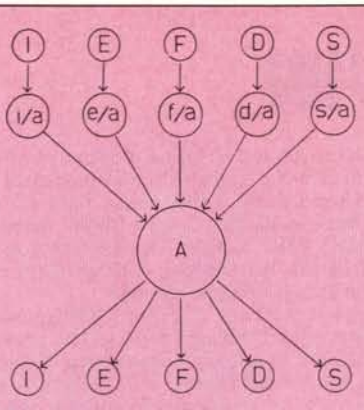
— un dizionario morfologico multilingue;

— un'analisi sintattica formale di Aymara;

— un analizzatore sintattico.

Il primo modulo non è solo un elenco di parole, bensì un vero e proprio strumento dedicato all'ingegneria del linguaggio. La classificazione dei vocaboli puó essere fatta in tre modi: morfologico, semantico e sintattico, il che vuol dire che il programma, dovendo analizzare una parola, ne determina il significato (semantica) dalla posizione nel periodo e dagli elementi che la circondano (sintassi), evitando di confondere sia il nome con il verbo di stessa forma, che i diversi significati d'uno stesso nome, che le espressioni tipiche o idiomi.

Il dizionario, che va realizzato immettendo le nuove parole in qualsiasi momento, necessita di circa un mese di lavoro per acquisire un numero di parole tale che quelle sconosciute siano in numero limitato, e 2-3 mesi per avere un dizionario esteso, di circa 50.000 parole per l'inglese, un po' meno per l'italiano e così via.



1

(1) Schema del funzionamento a ponte di Atamiri. Da una lingua si passa — tramite una trasformazione matematica — alla traduzione in Aymara. Da qui si passa alla lingua definitiva. Il metodo con cui si effettua il secondo passaggio non è illustrato, per non appesantire il disegno, e si basa su proprietà matematiche. Per andare dall'Aymara ad uno degli altri linguaggi, ad esempio lo spagnolo, basta invertire la matrice che fa andare nella direzione opposta, cioè da spagnolo ad Aymara.

Con Atamiri si abbrevia di molto la traduzione multilingue: dato che la conversione ad Aymara è comune a tutte le traduzioni, per ogni lingua oltre la prima serve metà del tempo totale.

2

3

(2) Metodi tradizionali nella traduzione. Il testo edito da un WP viene affidato ad un traduttore professionista (umano), che concepisce direttamente la versione definitiva; questa viene infine rielaborata per renderla esteticamente identica (paragrafi, note...) a quella d'origine. Questo sistema ha una velocità tra le 5 e le 7 pagine al giorno per uomo, ed un costo elevatissimo, in quanto il traduttore umano pesa sul 100% del tempo impiegato.
Così facendo, la traduzione multilingue richiede lo stesso tempo per ogni passaggio.

(3) Traduzione automatica con Atamiri. L'apparente maggiore complessità è in realtà dovuta alla maggiore articolazione del lavoro. Il testo inserito nel WP viene inizialmente tradotto da Atamiri in una forma grezza ed incompleta.

Il programma è stato scritto in P1 o Pascal su mini di opportune caratteristiche: al momento la principale limitazione che impedisce il trasferimento su PC o AT è la necessità di disporre di 100 MB di memoria di massa on-line, ovvero viste contemporaneamente dal sistema operativo, e di essere multiutente, dato il tipo di associazioni in cui Atamiri sarebbe davvero utile. Poiché in MS-DOS (l'ultima release è la 3.2) la massima estensione di un HD è di 33 MB, e di vera multiutenza non se ne parla proprio, questi sistemi non sono ancora sufficienti. È ipotizzabile andare sotto sistemi operativi più moderni, come Unix e il fratellastro Xenis, ma la mancanza di standard mondiali blocca l'implementazione con questo software.

Un'altra caratteristica che puó sembrare secondaria ma che non lo è, vede questo sistema accettare i testi forniti dai comuni WP in commercio per le grosse macchine.

A ruota libera su Atamiri (e sul suo ideatore)

Curiosando qua e là abbiamo trovato una valanga di spunti interessanti sulle tantissime facce di Guzman, Atamiri, Aymara, Quechua et cetera. Non potendo scrivere un libro sull'argomento ve ne citiamo una mezza dozzina in ordine sparso.

Qualche anno fa, Rojas cercava un metodo per insegnare ai ragazzi boliviani i principi dell'informatica. Si avvicinò così all'Aymara in modo analitico, dato che sulle Ande tra Bolivia e Perù sono molti coloro i quali parlano questo linguaggio od altri assai simili. Studiandolo a fondo, scoprì svariate cose. Innanzitutto che era perfettamente regolare; poi che la sua grammatica, sintassi e semantica erano altrettanto rigide; terzo, che si basava su tre possibili situazioni: vero, falso e incerto. Le prime due caratteristiche ne rendevano possibile la descri-

Agglutinanti e Flessive

Parlando di lingue, per di più così diverse da quelle d'oggi, viene naturale parlare della scienza che le descrive, la linguistica. Orecchie informatiche possono non esser d'accordo sul termine «scienza», ma questa non descrive solo fisica ed elettronica, bensì qualsiasi classificazione organica d'un settore del sapere, quindi anche lingue e dialetti.

Tornando alla lingustica, se la moderna tendenza è definire le lingue con i parametri vocali, per cui assumono importanza i singoli mattoncini della pronuncia, ovvero fonemi, allofoni e difoni, fino ai primi di questo secolo la classificazione era fatta con riferimento al significato delle varie parti della parola.

In ogni parola, infatti, è sempre possibile individuare un nucleo centrale, che contiene il concetto, e una o più parti, che definiscono se si tratta d'un nome, d'un verbo o d'un avverbio, ed eventualmente ne specificano il genere e il numero (per i nomi e i pronomi), oppure il tempo e la persona (per i verbi).

Per fare degli esempi, la parola
ANDIAMOCI
contiene la radice AND, le vocali IA che indicano il presente, il suffisso MO che indica la prima persona plurale e l'altro suffisso CI che indica il luogo (andiamoci = andiamo lì).

Secondo questa scuola, che classificava le lingue in base a radici e suffissi, ce n'erano tre tipi fondamentali:

1) monosillabiche;
2) agglutinanti;
3) flessive.

Nel primo caso i suffissi vengono semplicemente dopo la radice, senza essere scritti o letti insieme, e quindi senza subire modifiche di alcun genere. A questa categoria appartengono tutte le lingue dell'estremo oriente, come cinese, giapponese e lingue indocinesi.

Le lingue agglutinanti, invece, uniscono formalmente radice e suffissi, ma in generale senza intaccarne le caratteristiche (scrittura e pronuncia), per cui i vari componenti sono facilmente riconoscibili ed agglomerabili.

Tipici esempi di questa categoria sono altre lingue poco note in Italia, come quelle turche e l'ugro-finnico.

Le lingue europee, derivanti dall'ultimo sanscrito, sono quelle flessive, ove la radice viene alterata dai suffissi, che a loro volta si confondono nel formare la nuova unità lessicale, la parola completa, nella quale la provenienza è riconoscibile a fatica.

Da quanto detto nel resto dell'articolo, l'Aymara è un perfetto esempio di agglutinante: le parole, infatti, vengono formate per giustapposizione delle unità di base, che sono 16.000 e possono essere combinate in 400.000 modi con significato.

Come ultimo capoverso, sperando nella benevolenza del direttore, ci sia consentita una pignoleria: sebbene corretto ed accettato dai dizionari, il termine «linguistica» è un francesismo, come avverte anche il Dizionario Treccani, notoriamente curato dal Migliorini, e come tale andrebbe evitato. Nonostante possa sembrare brutto e fuori moda, il termine storicamente italiano è «glottologia» (o glossologia), nome composto dalle parole greche «glossa» (o glotta: in quella lingua, doppia S e doppia T erano interscambiabili) che vuol dire lingua, e «logos» che indica non solo il letterale discorso, ma anche tutte le scienze dell'epoca.

zione in termini formali con la moderna algebra, e quindi implementabili su calcolatore; la terza invece rendeva possibile ottenere una conclusione certa da dati iniziali incerti, risultato impossibile da ottenere con le lingue occidentali, tutte basate sulla sola scelta vero-falso, non ammettendo indeterminazione.

In quel momento Ivan ricordò le parole del padre, il pittore Cecilio, che gli ricordava sempre una cosa sulla tradizione andina: «Questa cultura è assai ricca», diceva; «non farti ingannare dalla sua apparenza povera».

E così venne fuori la dimensione di Aymara, un ramo del Quechua, la lingua più parlata dagli indi del Sudamerica da cinquemila anni a questa parte. Studiandolo, Rojas formulò un'altra teoria: Aymara non sarebbe una lingua come le altre, evolutasi con l'uso, bensì un linguaggio artificiale, progettato appositamente da sofisticatissimi ingegneri che nel 3000 avanti Cristo modellarono un perfetto esperanto andino. Un'impressione tanto profonda quanto affascinante, perché completamente staccata dagli schemi mentali usuali, in cui solo l'Europa degli ultimi duemila anni, seppure con alterne vicende, ospitava civiltà. Un punto di vista contestatissimo, ma dato per scontato dalla maggior parte degli europei, che ignorano del tutto non solo le civiltà precolombiane, ma anche l'Impero Egizio, le città-stato Accadiane, le dinastie Veda e la cultura del Sol Levante.

«È ironico» commenta Ivan, «che un'opportunità del genere sia fornita dal disprezzato linguaggio d'una popolazione per lo più analfabeta!».

La reazione del mondo scientifico internazionale alle idee di Rojas fu, come al solito, di derisione e scherno, ma anche di paura, concretizzata nel rifiuto di capire. La portata dei suoi risultati era così grande da sconvolgere buona parte delle teorie, e quindi delle reputazioni di coloro che le propugnavano. Pressioni politiche lo costrinsero a lasciare l'Istituto di Ricerche Scientifiche che lui stesso aveva fondato all'Università di La Paz. Continuando a lavorare nel tempo libero, e con i computer messigli a disposizione dai suoi clienti — banche, centri di calcolo — Rojas continuò gli studi sull'Aymara e su Atamiri, scrivendo una monografia di 150 pagine che fu pubblicata dal Centro Internazionale di Ricerca del Canada. Un piccolo contributo economico venne dal Centro Culturale, Scientifico e Didattico delle Nazioni Unite. Questi primi risultati non facevano che aumentare il numero degli scettici.

Cos'è l'IBI

L'associazione che ha portato in Italia Atamiri e il suo creatore è l'Intergovernmental Bureau for Informatics, con sede in Roma, Via Civiltà del Lavoro 23, tel. 5916041.

Le radici dell'IBI sono remote. Subito dopo la Grande Guerra, infatti, l'ECOSOC — Consiglio Economico delle Nazioni Unite — iniziò a discutere la realizzazione di strutture per sorreggere i vari organismi nazionali.

Nel 1961, dopo dieci anni di discussioni, viene fondato il CID, Centro Internazionale di Calcolo, in cui USA, GB e Francia mettevano le loro conoscenze a disposizione di tutti.

Nel 1974 il CIC si riorganizzava come IBI, mantenendo solidi contatti con l'UNESCO, ma sviluppando una propria politica come centro di riferimento non solo per i Paesi dell'America Latina, ma per tutti quelli del Terzo Mondo. Attualmente i Paesi membri dell'IBI sono una quarantina, tra europei, sudamericani ed africani, più qualche asiatico.

Una delle linee d'azione dell'Organizzazione è identificare i progetti che possano aiutare lo sviluppo economico, sociale e culturale dei Paesi, e lo scambio di esperienze e tecnologie: in questo contesto s'inquadra perfettamente la realizzazione del Seminario su Atamiri.

L'obiettivo finale dell'IBI è sviluppare le risorse dei Membri in modo organico, quasi fisiologico, nonostante la confusione e competizione attuali.

Intervista a Ivan Guzman De Rojas *19-11-'86*

Ivan Guzman de Rojas parla del suo lavoro senz'altra enfasi che quella d'un uomo tranquillo. Fa pochi paragoni, e li mimetizza nel discorso. Nel mostrare il nucleo di Atamiri sembra quasi un semplice addetto alle dimostrazioni. A domanda risponde prima direttamente, poi espandendo il discorso, e sempre in modo estremamente chiaro, come vedrete dalle poche domande che gli abbiamo rivolto.

Cos'è per Lei Atamiri?

Un gioco, molto bello. Tutti noi siamo bambini, almeno in parte, e vogliamo giocare. Questo è il mio.

Cosa vuole fare con questo sistema?

Trasferire le conoscenze a tutti. Vede, la maggior parte della letteratura tecnica è scritta in inglese, e molte persone nel mondo non conoscono questa lingua. Per conoscere i contenuti di un testo è possibile usare direttamente Atamiri, che fornisce assai velocemente una traduzione grezza ma assai comprensibile.

In questo modo è effettivamente possibile trasferire la conoscenza.

Che usi prevede per il suo traduttore?

La traduzione grezza fornita da Atamiri può essere davvero utile in un qualsiasi posto dove le fonti siano scritte in lingue diverse da quelle conosciute: come ho già detto, l'uscita diretta è già comprensibile.

C'è una seconda applicazione, a livello più alto. Attualmente le traduzioni ad alto livello vanno avanti a 5-7 pagine per giorno e per persona: usando Atamiri il sistema raggiunge subito le 40 pagine per giorno per persona. Successivamente il sistema impara nuove parole, e i tempi decrescono ulteriormente. E bisogna ricordare anche che l'elaboratore necessario lavora in multiutenza, per cui ha più posti di lavoro che possono portare avanti la traduzione contemporaneamente.

Quali sono i prossimi sviluppi?

In questi giorni (19-11-1986) stiamo educando il sistema anche alla lingua italiana.

Una sfida che m'attira molto è lavorare sull'arabo, una lingua assai difficile. Non pensate solo alla tastiera. L'arabo viene letto da destra a sinistra, ovvero nel modo opposto delle lingue europee. Questo non è un grosso problema logico, ma ci vorrà un po' di tempo per riscrivere e modificare tutte le routine che scandiscono il testo.

Questo sistema, che è multilingue, lavora con almeno 100 megabyte su hard disk. Non ha pensato ad una versione semplificante, che traduca soltanto tra due lingue e con un numero limitato di vocaboli?

Sì, ma è un progetto lontano. Sarebbe una cosa estremamente utile per la didattica. Nonostante siano scesi molto, i prezzi sono ancora alti. Forse nel prossimo anno il prezzo dell'hardware sarà abbastanza basso per consentirci di realizzare una versione che giri su personal computer.

Atamiri fu poi installato negli uffici della Commissione per il Canale di Panama, dove tuttora funziona assai meglio dei sistemi usati in precedenza. Ciò diede un grande impulso al metodo, ed interessò l'OAS, un'organizzazione statunitense con sede nella città di Washington, ove sono allo studio uffici per lo sviluppo del sistema in modo finalmente adeguato. Ad occhio e croce, Atamiri ha vinto una guerra che non pensava di combattere.

Per inciso, le affermazioni di Rojas sono tutt'altro che casuali. L'algebra moderna cerca di definire le proprietà dei vari insiemi di elementi, e classifica questi insiemi a seconda delle operazioni possibili. L'Aymara è in pratica una struttura ad anello A, composta da un insieme (non vuoto) di elementi su cui sono definite due operazioni che indichiamo con i simboli $+,*$ per le quali valgono le proprietà seguenti:

1) $(A,+)$ è un gruppo abeliano;
2) $(*)$ è associativa;
3) $*$ è distributiva rispetto a $+$.

Un esempio quotidiano di anello è l'insieme dei numeri interi.

La presenza di una struttura algebrica definita è fondamentale per la traduzione a ponte. Infatti normalmente per andare da una lingua ad Aymara bisogna avere un programma, che solitamente è diverso da quello opposto che porta da Aymara al linguaggio scelto. Se però le proprietà di Aymara sono rigorose, è possibile individuare un procedimento T che porta a lui, e con un semplice calcolo trovare da T il procedimento inverso che va nella direzione opposta.

Se quindi A è Aymara, I è italiano e T è la trasformazione che porta da I ad A, avremo che

$$A = T \cdot I$$

e quindi, nel verso opposto, che

$$I = A / T.$$

Poiché nell'algebra le trasformazioni si ottengono con matrici e non con numeri, l'espressione formale della seconda espressione va modificata in

$$I = (T^{\hat{}}(=1)) \cdot A.$$

Per quanto riguarda la possibilità di usare una logica di tipo vero-falso-incerto, o a tre stati, al posto d'una limitata ai primi due, è stato dimostrato che per strutture formali, come i computer e l'elettronica numerica, i due metodi sono equivalenti, nel senso che portano allo stesso risultato. È ovvio che la logica a tre stati ci arriva prima.

Sta di fatto, però, che in generale una lingua NON È una struttura formale — Aymara a parte — e quindi gl'idiomi europei mal si adattano a risolvere situazioni in modo logico.

MC