VERSIONE ITALIANA

Ringrazio amici e colleghi che mi hanno invitato a fare questa presentazione.

Negli atti trovate il mio paper che ha solo valore storico: è la prefazione al libro in cui sono raccolti gli interventi al Seminario Introduttivo alla

Linguistica Computazionale. Il seminario si tenne nel 1982 al Centro di Calcolo di Ca Foscari con la partecipazione dei colleghi di Pisa che si interessavano alla LC. L'intervento più interessante per quanto dirò oggi fu quello di Zampolli che però non volle metterlo su carta - non amava tanto scrivere quanto parlare. Il senso lo riporto io nella prefazione ed è che è necessario distinguere tra chi fa solo Linguistica Quantitativa e chi invece fa Linguistica Computazionale, le due non sono intercambiabili: di Risorse Linguistiche in quel tempo ancora non si parlava. Oggi vorrei dire che non si fa più né l'una né l'altra: si fa ingegneria della lingua, ma di questo dirò più avanti.

Cosa si faceva allora a Venezia: sia l'una che l'altra negli anni 70 e 80 ma poi soprattutto Linguistica Computazionale. In particolare le specialità che la distinguevano erano Applicazioni per l'Insegnamento delle lingue, Applicazioni per l'analisi letteraria e testuale, Applicazioni per il TTS o text-to-speech. Per quanto riguarda la LC, oltre alla PROSODIA, Venezia ha implementato algoritmi di morfologia generativa e di analisi sintattica – parsers - nella prima metà degli '80. e algoritmi di analisi semantica e ragionamento nella seconda metà.

E qui vale la pena ripassare la lezione:

Definizione 1, LC è lo studio scientifico del linguaggio umano, serve a fornire modelli di fenomeni linguistici attraverso la creazione di algoritmi che richiedono diversi passaggi in prove ed errori;

poi gli scopi teorici della LC come l'implementazione di teorie sintattiche e semantiche in forme trattabili computazionalmente, la scoperta di tecniche di analisi linguistiche che combinino le proprietà distribuzionali a quelle strutturali - paradigmatico a sintagmatico, sino ai modelli cognitivi e neurologici di come il cervello funziona quando si utilizza la facoltà del linguaggio.

Torniamo quindi a Venezia e alla prima metà degli anni 70 quando durante la mia permanenza in Australia per il mio dottorato mi resi conto che dovevo imparare ad utilizzare il computer se volevo fare uno studio approfondito dell'opera completa del poeta su cui stato lavorando.

Di ritorno in Italia imparai il FORTRAN che serviva per comunicare con il Cyber CDC, il mainframe che si trovava al CINECA a Bologna. Questo era però insufficiente: mi serviva la risorsa linguistica cioè il testo dei Collected Poems in forma digitale, che poi scoprii essere composto da 62000 parole circa.

Riuscii ad ottenere finanziamenti dal CNR con ricerche dal titolo suggestivo Analisi Stilostatistica ecc. per due anni 1976/77 e 1977/78. Questo mi permise di ingaggiare le persone che lavoravano al Centro di Calcolo e al Centro Linguistico e così ottenere il sospirato risultato. Avevo due testi in forma digitale: un libro di letture di Inglese Scientifico - di cui dirò dopo - e i Collected Poems del mio poeta.

Quindi completai l'analisi quantitativa del corpus del mio poeta e feci stampare il libro nel 1979 che conteneva quindi un capitolo con tante tabelle statistiche e una serie di gaussiane su cui veniva plottata la distribuzione dei pronomi personali nelle sei partizioni in cui era stata suddiviso il corpus. Il lavoro quantitativo venne poi pubblicato nel 1980 sulla

Rivista del LASLA. . Il plotter si trovava alla Facoltà di Scienze o meglio Chimica come si chiamava e mi fece le gaussiane un mio collega di Fisica.

Questa è la copertina della seconda edizione del libro apparsa qualche mese fa con un nuovo capitolo di Linguistica Quantitativa e un altro capitolo di Analisi Computazionale che in quel tempo non si poteva fare

Insegnavo Inglese scientifico con il ruolo di Professore Incaricato a Titolo Gratuito - avete sentito bene. Più o meno nello stesso tempo feci una grammatica dell'inglese tecnico derivata dagli spogli elettronici, concordanze,

liste di frequenza, ordinamenti inversi, ecc. Tutti gli esempi e il dizionario accluso erano presi dalla lingua inglese scientifica. Il libretto che ne derivò venne presentato al Seminario di Lemmatizzazione Computerizzata all'Index Tomisticus, invitato da Padre Busa che in quel tempo si trovava a Venezia dai Gesuiti perché a Venezia c'era il Centro Scientifico dell'IBM dove lui svolgeva le sue ricerche e dove andai alcune volte. In quell'occasione conobbi Zampolli, e gli altri personaggi importanti che erano presenti.

Il libro conteneva un capitoletto che spiegava come era stato creato il corso- qui vedete il flowchart con la sequenza delle operazioni fatto a mano... non c'erano software di grafica.

Assieme al lavoro quantitativo stavo però elaborando il mio ingresso nel mondo della linguistica teorica e computazionale attraverso l'ambito a me più congeniale, cioè la prosodia, la scienza più vicina alla mia attività di musicista negli anni 60. La prosodia conteneva al proprio interno tutto quello che costituisce la scienza linguistica dalla fonologia alla pragmatica. All'inizio ero concentrato sulla fonologia, la fonetica, anche quella sperimentale con la sua parte di acustica. Utilizzavo queste informazioni che trovavo in riviste e libri al Centro di Fonetica del

CNR di Padova, dove mi recavo diverse volte per settimana, per costruire un programma per l'assegnazione automatica dell'accento di parola in italiano.

Questo mio lavoro arrivò alle orecchie dei colleghi di Ingegneria Elettronica, uno di questi Antonio Mian mi invitò al Centro di Sonologia Computazionale dove si faceva musica elettronica ma dove si trovava un sintetizzatore vocale dell'italiano che parlava con voce femminile - di cui conservo registrazioni. Guidato da un mainframe dell'IBM parlava però una parola alla volta. In quel momento iniziò la mia collaborazione con Ingegneria di Padova che durò fino al 1986 per la realizzazione di un TTS dell'italiano e il suo trasferimento su motherboard.

Mi venne chiesto di fare un programma che permettesse al sintetizzato di pronunciare una frase di seguito.

Il programma PROSO era pronto nel 1980 e funzionava perfettamente. La collaborazione fu finanziata da un progetto CNR finalizzato per le Tecnologie Biomediche e Sanitarie, il cui scopo era creare un sistema su scheda madre per persone con disabilità uditive.

PROSO conteneva al proprio interno le regole fonologiche e prosodiche di cui era composto, che vennero illustrate in un articolo pubblicato nel 1981 sulla rivista dell'Accademia della Crusca. Lo presentai al mio primo convegno internazionale, il FASE che si tenne a Venezia. Lì conobbi oltre a tutti i personaggi mitici della fonetica acustica e sperimentale come Gunnar Fant, anche i ricercatori dello CSELT di Torino il centro ricerche della STET, l'attuale TELECOM.

Il direttore mi convocò a Torino e anche loro lavoravano al TTS dell'italiano ed erano interessati al mio sistema che si chiamava ora PROSO. Fu deciso di attivare delle tesi di ingegneria sotto la mia direzione, con lo scopo di rendere più efficiente il codice e trasformarlo in forma tabulare. Uno dei tesisti era Maurizio Omologo che i colleghi dell'IRST conoscono - ora ad Amazon.

Il laborato di LC di Venezia era ora sparso su più luoghi ma si stava per concentrare a Venezia. Il sistema PROSO venne presentato a diverse conferenze internazionali tra cui il primo EACL di Pisa e l'ICASSP di San Diego. Ma forse il momento più importante fu la presentazione invitata come rappresentante italiano alla conferenza di Doug Applet che si tenne alla Stanford University.

In Italia il LLC divenne un punto di riferimento per la creazione di risorse linguistiche utilizzabili in vari contesti, Ca Foscari stipulò convenzioni con l'IRST - dove c'era Oliviero Stock, con la FUB - la Fondazione Ugo Bordoni di Roma con Andrea Paoloni, e soprattutto con lo CSELT di Torino.

Nella programma PROSO stava entrando la sintassi e la semantica. Negli anni '80 avvenne una rivoluzione nelle teorie sintattiche e nacquero le teoria della West Coast, in California, tra cui la LFG della Joan Bresnan che io iniziai da subito a seguire. Con la Bresnan ebbi uno scambio epistolare e poi la incontrai allo XEROX Park per presentarle una mia teoria sulla syntactic closure. La Bresnan venne poi Invited speaker a Venezia che lei amava, nel convegno TAG che Joshi mi aveva chiesto di organizzare inizio anni 2000.

Il lavoro al TTS che si riferiva all'analisi di testi senza limiti di vocabolario si arricchivano di un analizzatore morfologico basato su radici e morfemi, un piccolo tagger e un parser a cascata organizzato attorno a un chunker per regole context-free. A questo lavorò Giorgio Satta nella sua tesi fatta a Padova ma implementata a Venezia al Centro Calcolo.

I risultati delle ricerche in corso per PROSO vennero presentati in vari convegni internazionali e su riviste e come speaker invitato da Geoffrey Leech in UK, e da Doug Applet in USA alla Stanford University come rappresentante italiano della LC. Racconto questo aneddoto riguardo a Stanford perché quando arrivai mi resi conto che la cosa era stata organizzata da Zampolli senza ovviamente dirmelo. Zampolli era rimasto impressionato dalla demo che avevo fatto al primo convegno EACL tenutosi a Pisa ed era venuto a dirmelo. Poi però prese questa iniziativa e io senza saperne nulla preparai per Stanford una conferenza sulla sintassi che stava entrando in PROSO. Quando terminai. mi venne accanto e mi disse all'orecchio che loro si aspettavano che parlassi del sistema di TTS e non della sintassi!!

Le presentazioni in USA ebbero come risultato importante la richiesta da parte della seconda più grande company di computer la Digital EQ. di portare il loro sistema di TTS che si chiamava DecTalk, in italiano. Questo progetto segnò l'inizio del Laboratorio di Linguistica Computazionale nel palazzo Ca Garzoni e Moro.

Ca' Foscari ricevette un finanziamento di 200milioni una Vaxstation II e una programmatrice stipendiata - una mia laureanda a Padova in ingegneria. Ma soprattutto mi venne finalmente concessa una stanza dove lavorarono gli studenti di linguistica di quel tempo - che ora sono i docenti di linguistica a Ca Foscari - Cardinaletti, Giusti, Brugè, e la Paola Merlo che ora lavora a Ginevra. La Merlo lavorava a un parser scritto in linguaggio C, a quel punto tutti iniziammo ad utilizzare il C.

Nel LLC venne rielaborato il sistema PROSO, ora in linguaggio C, ma venne soprattutto creato il primo TREEBANK a livello internazionale, annotando sui moduli continui in cui era stampato il testo, la struttura sintattica a costituenti. Questa struttura venne poi utilizzata per creare delle statistiche utili al parser.

Sempre nel LLC venne creato il primo lessico sottocategorizzato dell'italiano annotando i 7000 lemmi - nomi verbi e aggettivi - più frequenti, con il tramite di una interfaccia scritta in C.

Nel 1986 Dario Bianchi della Facoltà di Ingegneria dell'Università di Parma viene a Venezia per una collaborazione in un progetto di traduzione automatica. Inizia un nuovo capitolo del LLC grazie all'apporto scientifico che lui porterà che va dall'uso del Prolog per il parser dell'italiano - poi diventato sistema completo GETARUN - alla necessità di unire alla semantica il ragionamento, realizzato con l'uso del sistema KL-ONE reso disponibile dall'Università di Berlino.

Dal 1987 al 1991 il LLC partecipò al progetto europea EUREKA-PROMETHEUS, Pro-Art, Man-Machine Interface, che vide per la prima volta la partecipazione di tutte le unità italiane attive nel campo della LC. Il compito dell'Unità di Venezia era quello di produrre un algoritmo per la generazione di dialoghi parlati per la comunicazione con il co-pilota nel dominio del traffico, del turismo e delle previsioni del tempo. A questo scopo venne fatta una analisi dettagliata di un corpus di parlato trascritto che realizzò Laura Brugè. L'algoritmo per la generazione del linguaggio - nonché il modulo per la Situation Semantics nel Discourse Model - fu creato da un mio studente destinato poi a diventare un eccellente scolaro, Emanuele Pianta, purtroppo deceduto prematuramente.

Nel 1989 il sistema completo che utilizzava KL-ONE era in funzione e fu descritto in un importante lavoro pubblicato nel 1990 su Computer and the Humanities. Nei primi anni '90 Ca Foscari firmò accordi internazionali a livello europeo che permettevano a ricercatori di paesi come la Romania di venire a lavorare a Venezia finanziati. Ricevemmo da Iasi quattro ricercatori programmatori informatici uno dei quali era Dan Cristea. Come membro del Consiglio di Amministrazione proposi e riuscii a far finanziare un progetto chiamato SLIM con il quale si intendeva creare un sistema per l'auto-apprendimento della lingua inglese che utilizzava il riconoscimento e la sintesi del parlato. Sistema molto innovativo per quel tempo e che venne dimostrato in varie conferenze internazionali dal 1995 in poi. Il team del LLC si componeva ora di più di 12 persone, 5 madre lingua per inglese e francese, 3 programmatori, quattro linguisti.

Nei primi anni ‘90 le pubblicazioni aumentarono sensibilmente e spaziavano dallo speech e la fonetica sperimentale all’analisi dei quantificatori e l’implementazione di un algoritmo per il loro riconoscimento secondo gli schemi proposti da LFG, all’analisi dei pronominali e della struttura del discorso.

Di seguito per finire vi mostro l'elenco delle unità che si occupavano di LC in Italia compilato dall'AIIA e pubblicato in un libretto. Sempre di seguito la lista di alcune delle persone che si sono laureate o hanno svolto ricerche per la tesi nel LLC sotto la mia direzione. Poi un elenco delle 11 conferenze inter/nazionali organizzate dal LLC nel tempo. E ora ritorno sulla questione di fondo: esiste ancora la LC? Stiamo assistendo al trionfo dei LLM e delle Deep NNs. Se però è morta auguriamo lunga vita alla LC, perché la LC è una passione e vi auguro di essere tutti molto appassionati.