LA LINGUISTICA COMPUTAZIONALE A VENEZIA

Questo capitolo ha come argomento lo sviluppo della Linguistica Computazionale a Venezia con l'intenzione di mettere in luce le interrelazioni con gli altri componenti del Dipartimento di Linguistica, quelli di glottodidattiche e quelli della linguistica teorica con le quali ha interagito nel tempo. Lo sviluppo temporale permette anche di legare gli eventi locali all'avanzamento della tecnologia e della scienza linguistica sperimentale in ambito internazionale.
Quando alla fine del 1984 telefonai a Giovanni Freddi per chiedere se ci fosse un posto disponibile per un linguista computazionale nell'Istituto di Linguistica, la risposta fu subito positiva. Questa risposta orientò tutta la mia successiva carriera, visto che in quel momento dovevo decidere se accettare oppure no la richiesta ripetutami più volte di restare a Trieste dove insegnavo Lingua Inglese triennale nella Facoltà di Economia. Il 1984 aveva coinciso con la mia entrata in ruolo come professore associato, ma anche con il mio debutto ufficiale a livello internazionale, come dirò meglio più avanti.

Farò prima un passo indietro, all'inizio della mia carriera di ricercatore nel campo della linguistica computazionale, che si può far coincidere con il mio ritorno in Italia dal periodo di dottorato trascorso in Australia a Melbourne – metà degli anni '70. Il periodo in Australia aveva comportato due risultati fondamentali per me: una attività intensa nel campo della poesia, di lingua inglese ma anche di lingua italiana – interesse che non ho più abbandonato; la scoperta del computer come strumento di indagine scientifica in campo letterario e linguistico. Al mio ritorno in Italia, avevo deciso che ogni mio interesse di ricerca linguistica fosse legato alla sua sperimentabilità e il ruolo del computer – che avevo iniziato a utilizzare seguendo corsi in università - diventò per me lo strumento principe per la verifica empirica di ogni ipotesi linguistica teorica.

Questo valeva anche in generale per la Linguistica Computazionale (che indicherò da ora con LC), grazie anche ai cambiamenti in corso nella tecnologia con la nascita del desktop computer che stava sostituendo – anche in potenza - i main frames. La comodità di utilizzo era un motivo irresistibile per convertirsi al nuovo tipo di computer: questi stavano sul tavolo, mentre ai main frames si accedeva con il lettore di schede perforate dal Centro di Calcolo – negli anni '70 – poi con l'accesso da console in batch. La tecnologia ha giocato e gioca tuttora un ruolo importante nello sviluppo possibile del cosiddetto Natural Language Processing (che indicherò da ora con NLP) – la versione applicativa stretta della LC - quando si parla come si fa oggi di grandi quantità di dati: l'analisi linguistica richiede tempi di calcolo spesso incompatibili/inconciliabili con l'applicazione che dovrebbe utilizzarla. Se per ipotesi, ma cosa che accadrà in un futuro non lontano, le ricerche su web venissero fatte utilizzando NLP come strumento fondamentale, si potrebbero ricevere solo le risposte utili veramente. Questo attualmente non è possibile dal momento che il tempo medio di analisi di un documento di 2000 tokens è superiore esso stesso a quanto un utente è disposto ad attendere la risposta connesso al browser. E di documenti da 2000 tokens può accadere che il sistema ne debba analizzare migliaia prima di trovare le risposte.

Come altri ricercatori che operavano nel campo computazionale nella prima metà degli anni '70 anche io avevo ottenuto finanziamenti dal CNR per studi stilo-statistici. Questi non mi interessavano solo dal punto di vista letterario – come avevo fatto nella parte sperimentale della mia tesi Ph.D. – ma anche per poter utilizzare il risultato in sede didattica. Infatti, per il mio primo incarico di insegnamento a Venezia, Lingua Inglese, avevo organizzato una lista di frequenza di parole dell'inglese scientifico che gli studenti poi avevano nel loro libro di grammatica inglese e potevano consultare e studiare. Nel 1979 vengo invitato a presentare il mio lavoro al Seminario di Lemmatizzazione Computazionale organizzato da P.Roberto Busa all'Index Tomisticus di Venezia. Negli anni '60 e '70, la ricerca in campo computazionale era fortemente condizionata

dall'approccio quantitativo per quanto riguardava lo studio di corpora. Era consuetudine fare spogli di testi digitalizzati manualmente su mainframes. Da questi spogli si potevano costruire concordanze e liste di frequenza di vario tipo che permettevano di mettere in luce le caratteristiche lessicali dal punto di vista distributivo di un autore o di un libro. Il rapporto con la linguistica teorica era molto sentito nel campo formale da un lato e in quello dell'Intelligenza Artificiale - applicativo della LC in un senso più esteso e che ingloba la robotica e altre branche dell'ingegneria. Si faceva riferimento alla teoria chomskiana come quadro formale e se ne utilizzavano le intuizioni per creare dei "Parsers" parola inglese utilizzata per indicare in modo comprensivo tutto quello che serve per analizzare dal punto di vista sintattico e logico una frase.
Ma il vero mutamento in senso computazionale delle mie attività linguistiche avvenne quando sempre negli anni '70, al Centro di Fonetica Sperimentale del CNR di Padova – dove mi recavo frequentemente per aggiornarmi e scambiare idee, venni a contatto con le persone che lavoravano a Ingegneria. In questo modo scoprii che esisteva il Centro di Sonologia Computazionale frequentato da artisti della musica elettronica contemporanea di tutto il mondo, in cui c'era una apparecchiatura particolare: un sintetizzatore vocale analogico comandato da computer. Insomma un computer che parlava. A livello internazionale, la tecnologia si era sviluppata negli anni '70, quando vennero pubblicate le prime ricerche sul cosiddetto Text-to-Speech negli USA. Si trattava di sistemi computazionali molto complessi che inglobavano fonologia, fonetica, grammatica del discorso, sintassi, morfologia e lessico. Erano i primi sistemi di Text Understanding attraverso i quali un utente interagiva con un computer ponendogli delle domande scritte in linguaggio naturale e il computer rispondeva oralmente. Gli ingegneri di Padova, tra i quali Antonio Mian – morto di recente purtroppo – erano fortemente interessati ad avere un programma di Text-to-Speech o TTS per l'italiano per poter migliorare il proprio sintetizzatore e per poter quindi passare dalla lettura parola per parola a quella frase per frase – una trasformazione epocale.
Lavorando al programma scoprii che erano molto interessati anche gli ingegneri della STET-SIP – la Telecom di allora – che nel loro centro di ricerca a Torino, lo CSELT, avevano Piermarco Bertinetto come consulente. Bertinetto, che non aveva alcuna competenza computazionale, quando seppe che c'era un linguista che si interessava al TTS mi contattò. Facemmo alcune tesi assieme – a Torino – sul problema dell'assegnazione automatica dell'accento di parola in italiano. A quel punto il mio programma, che avevo chiamato PROSO, era nato: era la fine degli anni '70. Agli inizi degli anni '80 lo presentai ad alcune conferenze internazionali e risultò di grande interesse. Poi ci furono tesi di ingegneria a Padova e allo CSELT di Torino e varie richieste di utilizzarlo anche dall'estero – il CNRS di Parigi - che dimostrarono come il programma fosse all'avanguardia a livello internazionale. Fui anche invitato a presentarlo allo KTH di Stoccolma da Bjorn Granström.

Alla metà degli anni '80 PROSO venne trasferito su chip assieme al sintetizzatore vocale di Padova che nel giro di pochi anni era diventato completamente digitale e costituì uno strumento di ausilio per le persone non vedenti. Nei primi anni '80 organizzai a Venezia un workshop sulla Linguistica Computazionale in cui venne presentato lo stato dell'arte nei vari campi della LC, dal lessico alla morfologia e dalla sintassi alla semantica, da cui pubblicai una miscellanea. Venni invitato a presentare le mie ricerche a Stanford nel 1984 al 2nd International Workshop on Language Generation, che faceva il punto sulla situazione della LC. Il Centro Europeo dell'Educazione di Frascati mi invitò a far parte del Comitato Scientifico Nazionale per l'organizzazione di un progetto pilota che aveva lo scopo di introdurre l'Informatica nelle scuole italiane, nel gruppo di interesse per gli aspetti interdisciplinari della linguistica e delle lingue straniere. Sempre nel 1984 venni invitato da Geoffry Leech a presentare un lavoro alla Conferenza "Computers in English Language Research" organizzato dall'ICAME. Nello stesso anno andai all'Università di Napoli a presentare un lavoro sul tema "Computer and Scientific English Teaching".
In quel periodo, la LC aveva fatto un salto teorico fondamentale, abbandonando definitivamente l'approccio chomskiano all'analisi del linguaggio naturale per abbracciare le teorie funzionali e sistemiche. Mi riferisco in particolare alla teoria LFG di Joan Bresnan e a quella meno nota di

Martin Kay che ne è una versione derivata ma più formalizzata dal punto di vista matematico. Vi erano anche altre teorie che dicevano cose simili in forma diversa, come ad esempio la teoria HPSG e la Systemic Grammar di Halliday. Queste teorie hanno costituito e costituiscono il fondamento di molti sistemi funzionanti anche al giorno d'oggi. Una teoria meno nota in quel momento ma già abbracciata da diversi sistemi era la Dependency Grammar teorizzata originariamente dal linguista francese Tesnière e trasformata in strumento computazionale dal russo-canadese Melçuk. Io avevo scelto la teoria LFG fin dal suo inizio, per il respiro sperimentale che la ispirava e che poneva tra i suoi principi la sua computabilità e la sua plausibilità psicologica. Il mio approccio alla linguistica teorica era però associato all'interesse per la linguistica del testo e del discorso, attraverso la fonetica e la prosodia, e l'avevo presentato in vari congressi. Nel mio viaggio a Stanford avevo presentato le mie posizioni teoriche alla Bresnan con la quale ero rimasto in contatto anche successivamente. Nel 1985, avevo poi pubblicato un lavoro che associava i miei vari interessi linguistici con la LC, dal titolo "Parsing Difficulties & Phonological Processing in Italian", che avevo presentato alla 2nd Conference of the European Chapter of ACL, a Ginevra. Due anni prima avevo presentato alla 1st Conference of the European Chapter of ACL a Pisa un lavoro dal titolo "A Phonological Processor for Italian", accompagnato da una demo del sintetizzatore vocale di Padova che pronunciava frasi utilizzando il mio programma di TTS.
E' a questo punto che si inserì la mia entrata in ruolo e il mio trasferimento da Trieste a Venezia da dove mi ero allontanato per necessità: non avevo più un lavoro a pagamento e il mio incarico di Lingua Inglese era stato cancellato. A Venezia entrai in contatto con un universo linguistico che mi attraeva e che servì a mettere a punto le mie posizioni teoriche che pubblicai su riviste italiane. Qui avvenne un fatto straordinario: la Digital Equipment, la seconda più grande produttrice di computer in America e anche quella più innovativa sotto tutti i punti di vista, mi contattò attraverso i responsabili italiani per darmi la direzione di un progetto di grande rilievo: produrre la versione italiana del loro sintetizzatore vocale DECTalk. A questo progetto partecipò tutta la compagnia dei linguisti che frequentavano l'Istituto di Linguistica o perché erano laureati, ricercatori o semplicemente simpatizzanti. Ecco l'elenco dei collaboratori al progetto DECTalk: Anna Cardinaletti, Giuliana Giusti, Roberto Dolci, Laura Brugé e Paola Merlo. Gli ultimi tre hanno poi continuato a collaborare fino alla metà degli anni '90 in altri progetti nazionali ed europei che richiedevano la predisposizione di strumenti computazionali per la lingua italiana, strumenti inesistenti in quel momento.
Per poter costruire un sintetizzatore vocale per l'italiano della potenza di DECTalk era necessario avere a disposizione tutta la gamma di programmi di base per l'analisi automatica della lingua italiana senza limiti di vocabolario, come era stato fatto per la lingua inglese. Si partiva dall'analisi morfologica per passare a quella sintattica. Anche se esistevano parsers ATN per l'taliano, decisi di basarmi su un parser meno potente ma più generale come l'RTN. In questo modo, avevo anche deciso di limitare l'analisi alla struttura frasale alla possibilità di separare il Sintagma Verbale dal soggetto, se espresso. A questo scopo era necessario avere una grammatica context-free della lingua italiana possibilmente con informazioni di tipo probabilistico sui costituenti più frequenti e sul loro comportamento contestuale. Ed ecco che si pensò alla annotazione linguistica di testi veri digitalizzati, di cui il Laboratorio Computazionale già disponeva per lavori svolti in precedenza. Questo fu il primo caso a livello mondiale di creazione di una Treebank, cosa divenuta oggi fondamentale per qualsiasi lingua allo scopo di creare qualsiasi tipo di strumento automatico di analisi, ma che in quel momento era, di nuovo, inesistente. Inesistente era anche il Lessico Computazionale di frequenza dell'italiano contenente i primi 6000 lemmi tratti da vari spogli elettronici pubblicati nel tempo sull'italiano, a cui venne associata la sottocategorizzazione. Questo lavoro fu svolto a cavallo della fine degli anni '80 inizio anni '90 e di nuovo fu il primo in Italia e tra i primi a livello mondiale. Ciò che lo caratterizzava e lo distingueva e lo distingue anche oggi, è il fatto che un programma crea una versione codificata che poi viene decodificata sempre da programma. In pratica, il lessico viene prodotto da un linguista attraverso una interfaccia e il programma registra tutti i dati in forma codificata numerica su un solo record. Questo programma

venne scritto in C da uno degli informatici che collaboravano con me a quel tempo. Il decodificatore scritto da me sempre in C, permette/va di tradurre i dati di partenza che codificano la sottocategorizzazione di verbi, nomi e aggettivi, in vari formati. I formati iniziali sono la traduzione semplice del contenuto della codifica; poi però si può passare a una versione dello stesso lessico in cui vengono prodotti automaticamente i Ruoli Semantici associati agli argomenti dei predicati; infine c'è la possibilità di produrre un'ultima versione in cui il lessico viene scritto nel formato delle Rappresentazioni Concettuali alla Jackendoff. Per produrre queste rappresentazioni semantiche vengono utilizzate classi semantiche e concettuali associate a quelle sintattiche e quelle aspettuali. Poi il decodificatore contiene regole linguistiche che associano ad esempio la presenza di una marca aspettuale a quella semantica, oppure la presenza nel soggetto del tratto inerente umano associata a una marca sintattica o a una certa preposizione sottocategorizzata dall'obliquo.

Il luogo in cui era sistemato il laboratorio con il MicroVax a Ca' Garzoni Moro era una stanzetta che mi era conquistato con i denti, lottando con i bibliotecari. Alla fine i libri che erano conservati nella stanza vennero trasferiti altrove e il laboratorio poté divenire anche aula didattica per i miei corsi. Ca' Garzoni Moro era un palazzo molto particolare di cui nessuno conosceva la struttura interna: era l'unione di due palazzi con in mezzo una scalinata e i soffitti erano non coincidenti. All'ultimo piano c'era Storia e lì conobbi l'altra persona che nel palazzo si interessava di computers, anche se in maniera quasi passiva: era Giovanni Stiffoni il docente con il quale avrei sviluppato nel tempo una stretta amicizia. Il motivo era semplice, anche lui preferiva i computer Apple, cosa in quel tempo rara. Per cui qualsiasi problema ci fosse, ci consultavamo. E la cosa più incredibile fu la costruzione fisica della prima rete che collegava computers credo di poter dire a Venezia: questa fu fatta con un cavo di più di venti metri - che Giovanni mi allungò dal quinto piano fino a piano terra dove si trovava il laboratorio. Io poi provvidi a bloccare il cavo al muro ai vari piani e feci un buco nella finestra per farlo entrare in laboratorio dove si collegò direttamente al Vax e da qui ai Mac che erano dislocati nel laboratorio ma anche nel mio studio – da cui ovviamente io ero connesso precedentemente nello stesso modo. Giovanni era spinto non solo da una curiosità scientifica e personale, ma anche dalla stima incommisurabile che mostrava quando parlava di suo figlio Francesco, il quale aveva una passione sfrenata per il computer e era a conoscenza dell'esistenza dei primi esperimenti di connessione Internet in America e a Ginevra. Il figlio che studiava informatica, fu poi da me assunto per lavorare al progetto SLIM di cui parlerò in basso.

Nella seconda metà degli anni '80 inizia una collaborazione fondamentale per la mia carriera, quella con Dario Bianchi e Ingegneria di Parma. Bianchi era fortemente innamorato della LC e delle sue applicazioni nel campo dell'IA e programmava in Prolog. Fino a quel momento io avevo imparato – a lezione – il Fortran, poi da autodidatta il linguaggio C. Questi linguaggi però non erano particolarmente adatti a manipolare testi, quindi mi interessai al LISP. Questo era il linguaggio utilizzato in USA per scrivere i sistemi di NLP più noti, ed era stato anche utilizzato da Stock e Ferrari per creare il parser ATN dell'italiano a Pisa. In realtà, per poter far funzionare in maniera adeguata il Lisp era necessario avere computer molto potenti come ne esistevano a Pisa, ma anche alla Fondazione Bordoni di Roma e al Centro di Psicologia del CNR dove lavorava Stock. Si usavano dei minicomputer dedicati, della marca Symbolics. La scoperta del Prolog che non avevo mai utilizzato prima e che era il linguaggio più adatto a fare della semantica – PROgram LOGic – come faceva intendere il suo nome, mi prese molto del mio tempo all'inizio. Era il linguaggio introdotto in Europa per fare quello che negli USA si faceva utilizzando il LISP. Per poterlo installare – era il Quintus - sul minicomputer che mi aveva assegnato la DIGITAL per il progetto DECTalk – un MicroVax – dovevo chiamare direttamente in California e farmi dare il numero di serie passando attraverso un piccolo questionario di verifica della mia identità.

La collaborazione con Parma era iniziata per produrre un prototipo di traduttore automatico richiesto dalla ditta di microchip SGS Thomson, prototipo e progetto che ci occupò per alcuni mesi

ma non sortì risultato positivo. Rimanemmo quindi legati con una prima implementazione che però non potevamo utilizzare per il suo scopo. Il progetto della DIGITAL mi aveva permesso di costruire e di sperimentare con tutti gli strumenti di base per NLP della lingua italiana. Negli anni successivi vennero verificati altri approcci allo scopo di testare sia la validità teorica che l'efficienza degli algoritmi. Ad esempio l'analizzato morfologico per l'italiano subì vari cambiamenti vista l'inadequatezza della struttura rigida imposta dal database su cui si basava. Si trasformò in un algoritmo a regole, con un lessico di morfemi, un radiciario, prefissi suffissi e enclitici che controllavano la buona formazione della parola e licenziavano una sua decomposizione sulla base di paradigmi. La sintassi, passò dalla struttura a Recursive Transition Network alla versione Augmented, quindi ATN ma non nella versione classica che aveva molte rigidezze, ma grazie al Prolog, in una versione dichiarativa di facile implementazione.

Alla fine degli anni '80 mi riuscì anche attraverso Ingegneria di Parma di entrare in un progetto europeo di lungo respiro, il progetto EUREKA-PROMETHEUS, Subproject PROART-Man Machine Interface, NLP Unit, Prosody and Natural Language Generation. Questo progetto aveva come obiettivo finale la produzione di un co-pilota automatico che rispondesse a domande del pilota umano, riguardanti le località e il tempo atmosferico, a scopo turistico. Nel progetto c'erano tutte le case automobilistiche europee, dalla FIAT alla Mercedes, dalla Renault alla BMW. Noi avevamo il compito di produrre il generatore di risposte in linguaggio naturale. Bisogna pensare che in quel periodo, la generazione era una tecnologia ancora molto rudimentale e alle prime armi e funzionava essenzialmente sulla base di informazioni analogiche o da database che veniva trasformate in frasi ben formate concatenando "canned text". Noi partimmo subito con l'idea che si poteva arrivare a generare frasi ben formate sulla base di restrizioni linguistiche e semantico logiche. Quindi da un lato si lavorava alla costruzione del generatore, e dall'altra si cercavano le strutture linguistiche più ricorrenti nei testi che ci interessavano. Al generatore lavorava Bianchi e uno studente di Padova, Emanuele Pianta che è attualmente direttore del CLECT di Trento – un importante centro di ricerca. Alle strutture linguistiche lavoravano Laura Brugé e Roberto Dolci: in particolare Roberto aiutava anche a gestire il Vax, cosa non facile considerando le decine di manuali e le migliaia di pagine. Roberto lavorava anche alla prosodia assieme a me. I risultati sono stati poi trasferiti in prodotti, che però essendo molto all'avanguardia, non riscossero una accoglienza molto favorevole.
Uno dei risultati più importanti del lavoro che si stava portando avanti assieme fu l'utilizzo e l'adattamento del sistema KL-One implementato col nome BACK in Prolog e messo a disposizione gratuitamente dall'Università di Berlino. KL-One era un sistema per la Rappresentazione della Conoscenza in grado di funzionare come Reasoner e Theorem Prover, che funzionava partendo da una rappresentazione logica – una Forma Logica – da cui costruiva la T-Box e la A-Box, cioè il database estensionale e quello intensionale, che poi utilizzava per fare inferenze. A questo sistema Bianchi adattò il lavoro lessicale e il lavoro che si stava facendo sul ragionamento temporale e spaziale nel campo dell'Intelligenza Artificiale. In questo modo era possibile analizzare un piccolo testo e poi porre al sistema delle domande in linguaggio naturale ottenendo delle risposte logicamente verificate. Il lavoro su KL-One aveva permesso di affrontare problemi teorici che ci venivano posti nel momento in cui si voleva sostenere la validità del nostro approccio fondato su una analisi linguistica profonda – Deep Processing – nel campo del NLP. Non era una visione condivisa che da pochi ricercatori a livello internazionale, buona parte dei quali operavano nella sfera di influenza delle grammatiche funzionali-lessicali come LFG, HPSG ecc. Il campo del NLP era lentamente assorbito dalla IA a scapito della LC. Questo si cominciò a percepirlo agli inizi degli anni '90 quando le applicazioni fondate su basi di LC soffrivano dei difetti tipici degli approcci rigidamente teorici, cioè mancavano di estendibilità, scalabilità e erano fragili: cambiando di dominio o introducendo strutture non conosciute si bloccavano.

Il passaggio nel nuovo decennio segnò in maniera decisiva il campo della LC e del NLP in senso più IA e matematico. Questo non avvenne per caso. Nel '89 partecipai alla scuola NATO-ASI che

si tenne a Cetraro in Calabria. In quell'occasione l'IBM presentò i risultati degli esperimenti compiuti nel campo del riconoscimento vocale e il mondo della LC non fu più lo stesso. In pratica, Jelinek, il padre del progetto statistico e probabilistico, lavorando su corpora di milioni di parole, inventò il cosiddetto Language Model, un modello probabilistico delle occorrenze di forme di parole in contesto. Il LM era la base per un riconoscitore vocale, una volta che il sistema di riconoscimento era attrezzato con un database di parlato dello stesso parlante, in cui erano registrate tutte le possibili realizzazioni vocali delle forme di parole nei vari contesti. Le parole in realtà non erano registrate nude e crude, ma in forma decomposta, utilizzando il concetto dei difoni che esisteva in sintesi vocale. Per ogni segmento di parola era poi realizzato un modello probabilistico che permetteva di emettere una previsione di quale fosse il segmento possibile successivo. Questo modello probabilistico è chiamato HMM, o Hidden Markov Model. L'IBM era in quel momento la casa di computer più ricca e quella che aveva speso di più nella sperimentazione in campo linguistico. Come conseguenza di questa presentazione e pubblicazione degli atti alcuni anni dopo, in USA alcuni ricercatori in campo NLP applicarono lo stesso procedimento al corpus di 1 milione di parole, il treebank costruito dalla Penn University, chiamato quindi PennTreebank. I primi risultati furono esaltanti: era possibile prevedere la categoria grammaticale apprendendo le informazioni utili a livello contestuale automaticamente e applicando questa "grammatica" a un nuovo testo qualunque con una approssimazione del 92%. Poi la percentuale salì al 95% e infine oggi si è raggiunto anche il 98%.
La possibilità di apprendere dai dati era entrata subito nel DNA della nuova stagione del NLP che faceva capo alla IA, ma anche chi lavorava in LC dovette adeguarsi. Il fatto è che ora il risultato di questa decisione nei vari campi a cui è stato applicato è la verifica definitiva della sua insufficienza. Nel campo della sintesi vocale si è abbandonato completamente lo studio della sintesi a formanti che richiedeva conoscenze accurate della transizioni e dei meccanisci di produzione del parlato; nel campo del riconoscimento si è giunti a riconoscere il parlato attraverso sistemi chiamati Spoken Dialog Systems, in cui i turni e gli argomenti sono prefissati, ma per poter dettare un testo al computer vocalmente è indispensabile fare il training della propria voce; nel campo del NLP si è finalmente tornati a parlare da due o tre anni di semantica e magari di pragmatica. Dopo l'ubriacatura stocastica si è capito che questa può servire si e no – nel senso che lo stesso lavoro lo sanno fare egregiamente anche i sistemi a regole – per gli ambiti linguistici in cui gli oggetti da modellare sono segmenti adiacenti e ricorrenti con una buona frequenza. Fenomeni linguistici come la prosodia e tutto quello che attiene alla semantica proposizionale richiedono altri approcci.

Fu proprio del '92 l'assegnazione al mio laboratorio da parte del Consiglio di Amministrazione di Ca' Foscari di un cospicuo finanziamento per un progetto molto ambizioso ma concettualmente semplice. La creazione di un sistema per l'apprendimento dell'inglese che potesse essere utilizzato anche autonomamente dagli studenti e che fosse basato sulle più moderne tecnologie informatiche e linguistiche. Il progetto, chiamato da me SLIM, fu realizzato grazie alla presenza in Università di informatici romeni provenienti dall'Università di Iasi: Dan Cristea, Mirela Petrea, Luminita Chiran, Ciprian Bacalu erano i principali. Questi informatici lavorarono alla creazione di un sistema all'avanguardia per quei tempi, coadiuvati anche da Francesco Stiffoni. La parte più creativa del sistema era il modulo prosodico, che controllava automaticamente la prosodia a livello di parola, di gruppo fonologico e di gruppo intonativo. Per poter funzionare utilizzava le registrazioni in camera silente della produzione linguistica di alcuni lettori di madrelingua inglese. Furono registrate anche le ripetizioni a ritmo controllato di tutti gli enunciati contenuti in tre corsi di lingua inglese seconda, dei quali vennero anche digitalizzati i contenuti multimediali. C'erano poi un modulo che utilizzava il riconoscitore vocale fornito sui computer della Apple, che quindi divenne il nostro computer di riferimento. Questa cosa fu aspramente osteggiata dagli informatici che in quel tempo lo consideravano un computer giocattolo e non affidabile. La Apple aveva anche inventato un linguaggio HyperCard che serviva come metafora per elaborare strumenti per l'apprendimento. Questo linguaggio ebbe un grande impatto su Internet perché ha poi ispirato l'HTTP e il JavaScript – della SUN - che fu il punto di partenza per l'HMTL e Internet in definitiva. Il suo creatore Bill

Atkinson disse con rammarico che:

“I have realized over time that I missed the mark with HyperCard, I grew up in a box-centric culture at Apple. If I'd grown up in a network-centric culture, like Sun, HyperCard might have been the first Web browser. My blind spot at Apple prevented me from making HyperCard the first Web browser.”

Noi utilizzammo HyperCard per la creazione di tutto il sistema e in particolare dell’interfaccia. Il riconoscitore interagiva con lo studenti in varie fasi, quella più avanzata è il Role-Play, in cui cui il sistema alternava la parte di uno dei protagonisti con quella dello studente che prendeva il ruolo dell’interlocutore. Ma il riconoscitore veniva utilizzato anche per altri esercizi, tra cui rispondere a domande basate sui dialoghi dei corsi registrati sui quali lo studente avrebbe dovuto esercitarsi. La sintesi vocale invece veniva utilizzata per fare dettati a varie velocità, anche questo un modo assolutamente innovativo che fu recepito in maniera entusiastica negli USA. Entrai nel gruppo dei ricercatori che a livello mondiale si interessava dell’inserimento delle tecnologie della voce negli strumenti per l’apprendimento linguistico.
In Italia la Olivetti era interessata ancora a livello di ricerca in questi nuovi campi di studio avanzati e mi contattò per interagire con loro e creare piccole applicazioni in campo didattico. Ca’ Foscari divenne il referente per la verifica e la creazione di strumenti didattici all’avanguardia entrando in diversi progetti europei che erano in corso dei quali noi dovevamo decidere la validità. Purtroppo poi, ci furono ostacoli interni alla realizzazione di un prodotto che l’Università non voleva o non poteva commercializzare.

Le tematiche di ricerca del gruppo di LC nella seconda metà degli anni '90 erano ben definite e spaziavano nei seguenti campi:

a. studi, analisi e applicazioni di fonetica sperimentale – soprattutto nella prosodia - dal punto di vista teorico in ambito fonologico e dal punto di vista applicativo nel campo della glottodidattica con le tecnologie informatiche;
b. studi, analisi e applicazioni in campo lessicale e linguistico, dal punto di vista delle loro rappresentazioni lessicali e concettuali;
c. pubblicizzazione del sistema SLIM e degli approcci innovativi utilizzati, in particolare nell'uso del database linguistico annotato a vari livelli – fino a quello semantico e pragmatico – per la creazione automatica di esercizi;
d. lavori nel campo del parsing profondo e della risoluzione dell'anafora;
e. lavori nel campo del ragionamento e dell'analisi semantica e inferenziale da utilizzare nell'analisi testuale con il sistema GETARUNS;
f. analisi morfologica e analisi degli errori per uno spelling checker e un grammar checker dell'italiano.

Alcune nuove tendenze stavano emergendo, che avevano come riferimento i lavori emergenti a livello internazionale ed erano queste:

a. il parsing “shallow” e il “tag disambiguation” utilizzando procedure a regole mescolate con procedure stocastiche;
b. la generazione di domande e risposte da analisi profonda e il “discourse model” di un testo di riferimento;
c. la creazione di treebank sintattici di testi scritti e di dialoghi trascritti – attività questa svolta all'interno di progetti nazionali.

Queste nuove attività ci permettono di mantenere il gruppo in funzione fino alla seconda metà degli anni 2000. Ma nel frattempo, nel 2000 appunto nasce il Dipartimento di Scienze del Linguaggio e si prefigura la possibilità di attivare un corso di Laurea all'interno del quale creare un percorso

completo di LC. I corsi che venivano impartiti precedentemente potevano essere scelti anche da studenti di Informatica, e questo fatto permetteva di creare delle sinergie inedite e proficue tra studenti di Lingue e studenti di Scienze. Purtroppo poi la rigidezza dei corsi di laurea bloccò questa possibilità. Ma la LC sta iniziando un percorso nuovo che la porterà un po' alla volta nel giro di un decennio a riappropriarsi degli aspetti più propriamente linguistici. Questa inversione di tendenza o sviluppo obbligato è dovuto - com'è giusto nel caso di una scienza applicativa come la LC - alla necessità di soddisfare meglio lo sviluppo delle tecnologie e le richieste del mercato nel campo dell'Information Technologies (IT). Lo sviluppo di internet e dei motori di ricerca sta invadendo tutti gli ambiti scientifici e non può non condizionare la LC che si basa appunto sull'analisi e la generazione delle lingue naturali o NLP. Si sente sempre più la necessità di affinare le tecniche di Information Retrieval basata sino a quel momento su una ricerca per parole chiave, che estromette le parole funzionali o stopwords dal calcolo e quindi cancella tutte le relazioni sintattiche e semantiche che intervengono nella frase e nel discorso. Questa situazione viene affrontata in maniera radicale quando agli inizi del decennio dei visionari introducono la cosiddetta Semantic Web, o meglio una rete fondata sulla condivisione del contenuto semantico delle informazioni nonché della loro rintracciabilità nel mondo reale. La Semantic Web si fonda su principi semantici e sulle cosiddette RDF che sono strutture logiche ispirate alla logica dei predicati di prim'ordine.

Le tematiche scientifiche legate a questa nuova ondata di sviluppo della LC sono quindi tutte improntate sulla necessità di recuperare la sfera semantica e quella pragmatica. Questo avviene su vari fronti, ma quello più interessante è lo sviluppo sempre maggiore di Challenges e Competitions su scale mondiale che mettano in gioco gli attori che sono in grado di sviluppare tecniche all'avanguardia e costituiscano una finestra verso il mondo scientifico per mettere in luce i migliori approcci e le migliore tecnologie. Leader in questo campo sono ovviamente gli USA che con il NIST – il famoso istituto degli standards nelle tecnologie indìce queste gare e si preoccupa di organizzare i dati sperimentali per lo sviluppo e la fase di test. La verifica dei risultati avviene poi in un workshop aperto a tutti i team che hanno partecipato su base volontaria alla challenge, organizzato di solito a Washington nella sede governativa del NIST.

In questo modo si mettono in circolazione dati condivisibili sul Question/Answering che hanno lo scopo di migliorare il funzionamento dei motori di ricerca nel senso linguistico, che tutti possono utilizzare per valutare la qualità del proprio sistema. Si tratta di migliaia e migliaia di domande e risposte raccolte manualmente. Un altro campo che vede uno sviluppo enorme grazie a queste iniziative è quello della cosiddetta Text Summarization fatta su basi estrattive che ci vede come protagonisti. Partecipano a queste challenges una ottantina di team da tutto il mondo, provenienti da università, software house, grandi multinazioni, centri di ricerca di ogni tipo. Dalla valutazione dei risultati ottenuti dal nostro sistema, per due volte siamo rientrati nella prima metà, cioè 34/35esimi.

Un ultimo tema di ricerca nato negli ultimi cinque anni è il cosiddetto Text Entailment, che ha come scopo l'individuazione automatica delle relazioni semantiche di entailment testuale tra un piccolo testo e una frase "ipotesi", che rappresenta la forma estesa di una domanda. Si tratta cioè di decidere automaticamente se il testo risponde alla domanda oppure no. Anche qui il nostro sistema ha ottenuto dei risultati elevati, sempre nella prima metà – sesto/settimo - dei partecipanti che però in questo caso non superano la trentina.

Le attività più recenti sono appena iniziate, e si rivolgono ai social network – Facebook, Twitter ecc. – da cui è possibile evincere l'opinione dei frequentatori su varie questioni, tra cui ovviamente anche prodotti commerciali. E' il cosiddetto "opinion mining" detto anche "sentiment analysis". Si tratta di analizzare i testi per indovinare se l'atteggiamento di chi scrive è positivo oppure negativo rispetto all'argomento di riferimento. Negli ultimi anni si è lavorato assieme a ricercatori di altre università come quella di Ginevra, per l'annotazione automatica dei dialoghi su basi argomentative, allo scopo di individuare le parti di dialogo di maggior interesse dal punto di vista del contenuto. Questo lavoro è partito dall'analisi dei dialogo multiparty – cioè a più interlocutori – realizzati e trascritti dall'ICSI di Berkeley sui quali era necessario anche tenere sotto controllo le sovrapposizioni di turno, che segnalano in maniera inequivocabile quali sono i parlanti più

competitivi rispetto agli altri. Analizzare dialoghi trascritti è di un livello di difficoltà di gran lunga superiore a dialoghi che avvengono con la tecnica del chat o nel forum di un blog dove ognuno scrive il proprio enunciato da tastiera senza esitazioni, ripetizioni, frammenti, nonparole, intercalari e appunto non ci sono sovrapposizioni da analizzare e situare temporalmente nel flusso del discorso. Da ultimo, sulla scia della Semantic Web, si sono sviluppate nel mondo le cosiddette Web Ontologies, anche nella forma di un Cloud o nuvola di ontologie interconnesse, che non sono altro che delle enormi enciclopedie accessibili direttamente da computer attraverso un qualsiasi programma e sono scritte per essere interpretabili da programmi di computer.
E' chiaro a questo punto che l'utilizzo dei telefonini diverrà sempre più diretto a soddisfare la sete di informazione che utenti sempre più esperti vorranno dalla rete. Ma questo non sarà con il web attuale e con i computer attuali: le analisi linguistiche richiedono come ho detto all'inizio grande velocità di elaborazione cosa che con i computer attuali non sarà possibile fare. La Linguistica Computazionale sarà sempre più una disciplina d'avanguardia ma anche di routine per quei linguisti che vorranno cimentarsi con il mondo reale. A questo scopo ho pubblicato in due volumi la mia Weltanschauung sulla linguistica computazionale nell'analisi di testi accompagnati da un CD in cui sono contenute tutte le diverse versioni del sistema di analisi costato anni di ricerca nel mio laboratorio, e la collaborazione di decine di ricercatori e scienziati del linguaggio.

Delmonte R., 2007. **Computational Linguistic Text Processing – Logical Form, Semantic Interpretation, Discourse Relations and Question Answering**, Nova Science Publishers, New York.
Delmonte R., 2008. **Computational Linguistic Text Processing – Lexicon, Grammar, Parsing and Anaphora Resolution**, Nova Science Publishers, New York.

In basso invece ho messo tutte le pubblicazioni più importanti dall'inizio della mia carriera in modo che si possa verificare con mano e in dettaglio la bontà dell'approccio alla studio linguistico da me sostenuto negli anni.

PUBBLICAZIONI

1980
Computer Assisted Literary Textual Analysis with Keymorphs and Keyroots, REVUE-Informatique et Statistique dans les Sciences humaines,1, 21-53.

1981
L'accento di parola nella prosodia dell'enunciato dell'Italiano standard, Studi di Grammatica Italiana, Accademia della Crusca, Firenze, 69-81.
An Automatic Unrestricted Tex-to-Speech Prosodic Translator, **Atti del Convegno Annuale A.I.C.A.,** Pavia, pp.1075-83.

1982
Automatic Word-Stress Patterns Assignment by Rules: a Computer Program for Standard Italian, Proc. IV F.A.S.E. Symposium, 1, ESA, Roma, 153-156.

1983
A Phonological Processor for Italian, Proceedings of the 1st Conference of the European Chapter of ACL, Pisa, 26-34.

1984
Complex Noun Phrases in Scientific English, Proceedings of the ICAME 84 - Conference on Computers in English Language Research, ICAME, Windermere(UK), 174-176.
"La 'syntactic closure' nella Teoria della Performance", in *Quaderni Patavini di Linguistica*, n. 4, Padova, pp. 101-131.
con G.A.Mian,G.Tisato,A Text-to-Speech System for the Synthesis of Italian, in *Proceedings of ICASSP'84*, San Diego(Cal), 291-294.

1985
Parsing Difficulties & Phonological Processing in Italian, Proceedings of the 2nd Conference of the European Chapter of ACL, Geneva, 136-145.
Sintassi, semantica, fonologia e regole di assegnazione del fuoco, Atti del XVII Congresso SLI, Bulzoni, Urbino, 437-455.
con G.A.Mian, G.Tisato, Un riconoscitore morfologico a transizioni aumentate, Atti Convegno Annuale A.I.C.A., Firenze, 100-107.

1986
A Computational Model for a text-to-speech translator in Italian, Revue - Informatique et Statistique dans les Sciences humaines, XXII, 1-4, 23-65.
con G.A.Mian, G.Tisato, A Grammatical Component for a Text-to-Speech System, Proceedings of the ICASSP'86, IEEE, Tokyo, 2407-2410.

1987
The Realization of Semantic Focus and Language Modeling, in Proceeding of the International Congress of Phonetic Sciences , Tallinn (URSS), 100-104.
Grammatica e ambiguitˆ in Italiano, **Annali di Ca' Foscari** XXVI, 1-2, pp.257-333.
Il principio del sottoinsieme e l'acquisizione del linguaggio, in P.Cordin(ed),Ipotesi e Applicazioni di Teoria Linguistica, - dal XIII Incontro di Grammatica Generativa, Trento, 47-64.

1988
Focus and the Semantic Component, in *Rivista di Grammatica Generativa*, 81-121.
Appunti per un corso di Grammatica Lessico-Funzionale, in *Annali di Ca' Foscari*, XXVII, N.1-2, pp.51-110.

1989
Computational Morphology for Italian, in AA.VV., Studi di Linguistica Computazionale, UNIPRESS, Padova, Chapt.I,1-20.
con R.Dolci, Parsing Italian with a Context-Free Recognizer, Annali di Ca' Foscari XXVIII, 1-2,123-161.

1990
Semantic Parsing with an LFG-based Lexicon and Conceptual Representations, Computers & the Humanities, 5-6, 461-488.

1991
Linguistic Tools for Speech Understanding and Recognition, in P.Laface, R.De Mori(eds), Speech Recognition and Understanding: Recent Advances, NATO ASI Series, Vol. F 75, Springer -Verlag, 481-485.
Empty Categories and Functional Features in LFG, Annali di Ca'Foscari, XXX, 1-2,79-140.
Grammatica e Quantificazione in LFG, Quaderni Patavini di Linguistica, 10, 3-71.
con D.Bianchi, Binding Pronominals with an LFG Parser, Proceeding of the Second International Workshop on Parsing Technologies, Cancun(Messico), ACL 1991, pp. 59-72.
con R.Dolci, Computing Linguistic Knowledge for text-to-speech systems with PROSO, Proceedings 2nd European Conference on Speech Communication and Technology, Genova,ESCA.

1992
con D.Bianchi, Quantifiers in Discourse, in Proc. ALLC/ACH'92, Oxford(UK), OUP, 107-114.
con D.Bianchi, E.Pianta, GETA_RUN - A General Text Analyzer with Reference Understanding, in Proc. 3rd Conference on Applied Natural Language Processing, Systems Demonstrations, Trento, ACL, 9-10.
Relazioni linguistiche tra la struttura intonativa e quella sintattica e semantica, in E.Cresti et al.(eds), Atti del Convegno Internazionale di Studi Storia e Teoria dell'nterpunzione", Roma, Bulzoni, pp. 409-441.

1993
con Bianchi D., E.Pianta, Understanding Stories in Different Languages with GETA_RUN, Proc. EC of ACL, Utrecht, 464.
GETA_RUN: A fully integrated system for Reference Resolution by Contextual Reasoning from Grammatical Representations, ACL-93, Exhibitions and Demonstrations, Columbus.

1994
con D.Bianchi, Computing Discourse Anaphora from Grammatical Representation, in D.Ross & D.Brink(eds.), Research in Humanities Computing 3, Clarendon Press, Oxford, 179-199.
con E.Pianta, Discourse Structure and Linguistic Information, ACH/ALLC '94, Consensus Ex Machina Paris, 61-62.
Inferences and Discourse Structure, in G.Ferrari(ed), Proc.IV Conference of the Italian Artificial Intelligence Association - AIIA, 11-14.
Analisi pragmatica e prosodica dell'enunciato "Vabbene?!", Atti del Convegno AIA - Gruppo Fonetica Sperimentale, Torino - Roma, 163-176.

1995
Lexical Representations: Syntax-Semantics interface and World Knowledge, in Notiziario AIIA (Associazione Italiana di Intelligenza Artificiale), Roma, pp.11-16.
con Dibattista D. Switching from Narrative to Legal Genre, Working Papers in Linguistics, 5-1, University of Venice, 1-41.
con F.Greselin, How to create SLIM courseware, in Yeow Chin Yong & Chee Kit Looi(eds.),Proceedings of ICCE '95, Singapore, Applications Track, 206-213.
curatore di How to create SLIM courseware - Software Linguistico Interattivo Multimediale, Unipress, Padova.
Understanding texts in different languages with Geta_Run, Proc. JADT'95, Roma, 279-286.
con F. Stiffoni, SIWL - Il Database Parlato della lingua Italiana, Convegno AIA - Gruppo di Fonetica Sperimentale, Trento, 99-116.

1996
con Dan Cristea, Mirela Petrea, Ciprian Bacalu, Francesco Stiffoni, MODELLI FONETICI E PROSODICI PER SLIM, Atti 6° Convegno GFS-AIA, Roma, 47-58.
con Andrea Cacco, Luisella Romeo, Monica Dan, Max Mangilli-Climpson, Francesco Stiffoni, SLIM - A MODEL FOR

AUTOMATIC TUTORING OF LANGUAGE SKILLS, Ed-Media 96, AACE, Boston.
Contextual Reasoning and Inferential Processing, in Proc.SIMAI'96, 229-231.
con Bianchi D., Temporal Logic in Sentence and Discourse, in Proc.SIMAI'96, 226-228.
con E.Pianta, IMMORTALE - Analizzatore Morfologico, Tagger e Lemmatizzatore per l'Italiano, in Atti V Convegno AI*IA "Cibernetica e Machine Learning", Napoli, 19-22
con G.Ferrari, A.Goy, L.Lesmo, B.Magnini, E.Pianta, O.Stock, C.Strapparava, ILEX - Un dizionario computazionale dell'italiano, Proc. of the Fifth Convegno Nazionale dell'AI*IA "Cibernetica e Machine Learning", Napoli, 27-30.

1997
Lexical Representations, Event Structure and Quantification, Quaderni Patavini di Linguistica, 15, 39-93.
Learning Languages with a "SLIM" Automatic Tutor, in Asiatica Venetiana 2, pp.31-52.
con M.Petrea, C.Bacalu, SLIM Prosodic Module for Learning Activities in a Foreign Language, Proc.ESCA, Eurospeech97, Rhodes, Vol.2, pp.669-672.
con R.Dolci, Sound Parsing and Linguistic Strategies, Atti Apprendimento Automatico e Linguaggio Naturale, Torino, pp.1-4.
con Bianchi D., Rappresentazioni concettuali nella comprensione di storie, Atti Apprendimento Automatico e Linguaggio Naturale, Torino, pp.95-98.
Rappresentazioni lessicali e linguistica computazionale, Atti SLI, Lessico e Grammatica - Teorie Linguistiche e applicazioni lessicografiche, Roma, Bulzoni, pp.431-462.

1998
Le nuove tecnologie e l'insegnamento della lingua straniera, Periplo, Venezia.
Prosodic Modeling for Automatic Language Tutors, Proc.STiLL 98, ESCA, Sweden, 57-60.
con D.Bianchi, Dialogues From Texts: How to Generate Answers from a Discourse Model, Atti Convegno Nazionale AI*IA, Padova, 139-143.
Phonetic and Prosodic Activities in SLIM, an Automatic Language Tutor, Proc.EUROCALL, Leuven,77-78.
con E.Pianta, Immortal: How to Detect Misspelled from Unknown Words, in BULAG, PCUF, Besançon, 1998, 193-218.
L'apprendimento delle regole fonologiche inglesi per studenti italiani, in Atti 8° Convegno GFS-AIA, Pisa, 177-191.

1999
con Bacalu C., Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database, in Atti 9° Convegno GFS-AIA, Venezia.
A Prosodic Module for Self-Learning Activities, Proc.MATISSE, London, 129-132.
con Bistrot A., Il MUSEIKA in giapponese: desonorizzazione, devocalizzazione o elisione vocalica?, in Atti 9° Convegno GFS-AIA, Venezia.
La variabilità prosodica: dalla sillaba al contenuto informativo, in Atti 9° Convegno GFS-AIA, Venezia, 133-146.
con E.Pianta, Tag Disambiguation in Italian, in Proc. Treebank Workshop ATALA, Paris, pp.43-49.
con D.Bianchi, Determining Essential Properties of Linguistic Objects for Unrestricted Text Anaphora Resolution, Proc. Workshop on Procedures in Discourse, Pisa, pp.10-24.
con D.Dibattista, E.Pianta, Parsing and Interpreting Quantifiers with GETARUN, Proc. VEXTAL, Unipress, pp.215-225.
con Dario Bianchi, Reasoning with A Discourse Model and Conceptual Representations, Proc. VEXTAL, Unipress, pp. 401-411.
From Shallow Parsing to Functional Structure, in Atti del Workshop AI*IA - "Elaborazione del Linguaggio e Riconoscimento del Parlato", IRST Trento, pp.8-19.
con Bacalu C., Prosodic Modeling for Speech Recognition, in Atti del Workshop AI*IA - "Elaborazione del Linguaggio e Riconoscimento del Parlato", IRST Trento, pp.45-55.
Grammar and Structure, BULAG 24, PUFC, 19-37.
curatore con A.Bristot, Aspetti computazionali in fonetica, linguistica e didattica delle lingue: modelli e algoritmi, Atti delle IX Giornate di studio del GFS(AIA), Prefazione, Venezia, pp.iii.-v, Unipress.

2000
SLIM Prosodic Automatic Tools for Self-Learning Instruction, Speech Communication 30, 145-166.
con Luminita Chiran, Ciprian Bacalu, Elementary Trees For Syntactic And Statistical Disambiguation, TAG+5, Paris, pp.237-240.
con L.Chiran, C.Bacalu, Towards An Annotated Database For Anaphora Resolution, LREC, Atene, pp.63-67.
Shallow Parsing And Functional Structure In Italian Corpora, LREC, Atene, pp.113-119.
Speech Synthesis for Language Tutoring Systems - Some Examples, Proc. InSTIL2000, Dundee, pp. 1-25.
Generating and Parsing Clitics with GETARUN, Proc. CLIN'99, Utrech, pp.13-27.
Parsing Preferences and Linguistic Strategies, in LDV-Forum - Zeitschrift fuer Computerlinguistik und Sprachtechnologie - "Communicating Agents", Band 17, 1,2, pp. 56-73.
con Montemagni et al., The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation, LINC, ACL, Luxembourg, pp.18-27.
Parsing with GETARUN, Proc.TALN2000, 7° confèrence annuel sur le TALN, Lausanne, pp.133-146.
Generating from a Discourse Model, Proc. MT 2000 - MACHINE TRANSLATION AND MULTILINGUAL APPLICATIONS IN THE NEW MILLENNIUM, BCS, Exeter(UK), pp.25-1/10.

2001
Tecniche e Strumenti per una Scomposizione e Rappresentazione Multilivello del contenuto linguistico di Dialoghi Spontanei (Tecniche e Strumenti per la rappresentazione di dialoghi), CNR, Roma.
con Luminita Chiran, Ciprian Bacalu, HOW TO INTEGRATE LINGUISTIC INFORMATION IN FILES AND GENERATE FEEDBACK FOR GRAMMAR ERRORS, Workshop on Sharing Tools and Resources for Research and Education, ACL, Toulouse, 10-14.

How to Annotate Linguistic Information in FILES and SCAT, in Atti del Workshop "La Treebank Sintattico-Semantica dell'Italiano di SI-TAL, Bari, pp.75-84.

2002

A Prosodic Module for Self-Learning Activities, Proc.SPEECHPROSODY2002, Aix-en-Provence, 243-246.
Relative Clause Attachment And Anaphora: A Case For Short Binding, Proc.TAG+6, Venice, pp.84-89.
From Deep to Shallow Anaphora Resolution: What Do We Lose, What Do We Gain, in Proc. International Symposium RRNLP, Alicante, pp.25-34.
con D. Bianchi From Deep to Partial Understanding with GETARUNS, Proc.ROMAND2002, Universita' Roma2, Roma, pp.57-71.
con Bianchi D., Tecniche di apprendimento applicate al problema del tagging: una prima valutazione per l' Italiano, Workshop "NLP E WEB: LA SFIDA DELLA MULTIMODALITA' TRA APPROCCI SIMBOLICI E APPROCCI STATISTICI", Convegno Nazionale AI*IA, Siena, pp.20-34.
con D. Bianchi, Reasoning On Mistakes For Feedback Generation, Workshop "NLP E WEB: LA SFIDA DELLA MULTIMODALITA' TRA APPROCCI SIMBOLICI E APPROCCI STATISTICI", Convegno Nazionale AI*IA, Siena, pp.40-48.
From Deep to Shallow Anaphora Resolution:, in Proc. DAARC2002 , 4th Discourse Anaphora and Anaphora Resolution Colloquium, Lisbona, pp.57-62.
GETARUN PARSER - A parser equipped with Quantifier Raising and Anaphoric Binding based on LFG, Proc. LFG2002 Conference, Athens, pp.130-153, at http://cslipublications.stanford.edu/hand/miscpubsonline.html.
Relative Clause Attachment And Anaphora: Conflicts In Grammar And Parser Architectures, in A.M. Si Sciullo(ed), Grammar and Natural Language Processing, UQAM, Montreal, pp.63-87.
Feedback generation and linguistic knowledge in 'SLIM' automatic tutor, ReCALL 14 (1): Cambridge University Press, 209-234.
Linguistic Knowledge and Reasoning for Error Diagnosis and Feedback Generation, in Trude Heift and Mathias Schulze(eds.), Error Analysis and Error Correction in Computer-Assisted Language Learning CALICO Spring 2003 special issue, pp.513-532.
A Prosodic Module for Self-Learning Activities, Proc.SPEECHPROSODY2002, Aix-en-Provence, 243-246.
con Bianchi D., Reasoning On Mistakes For Feedback Generation, Workshop "NLP e Web: la Sfida della Multimodalita' tra Approcci Simbolici e Approcci Statistici", Convegno Nazionale AI*IA, Siena, pp.40-48.

2003

Getaruns: a hybrid system for summarization and question answering, in Proc. Workshop "Natural Language Processing for Question Answering" in EACL, Budapest, pp.6.
con Bianchi D., NLP e ragionamento per la diagnosi degli errori e la generazione di feedback, in AI*IA Notizie, XVI, 1, p.61-66, Milano.
Trasduttori Multilivello del Contenuto Linguistico di Dialoghi Spontanei: Dalle Espressioni al Significato in Formato XML, in Cosi P. E.M.Caldognetto, A.Zamboni(eds)(2003), Studi in Onore di Franco Ferrero, UNIPRESS, Padova, pp.117-134.
STRUTTURE SINTATTICHE DALL'ANALISI COMPUTAZIONALE DI CORPORA DI ITALIANO, in Anna Cardinaletti(a cura di), "Intorno all'Italiano Contemporaneo", Franco Angeli, Milano, pp.187-220.

2004

Evaluating Students' Summaries with GETARUNS, Proc.INSTIL/ICALL2004, Unipress, Padova, 91-98.
con P. Cosi, S. Biscetti, R. A. Cole, B. Pellom, S. van Vuren, ITALIAN LITERACY TUTOR: tools and technologies for individuals with cognitive disabilities, in R.Delmonte & S.Tonelli(eds), Proc.INSTIL/ICALL2004, Venezia, 207-215.
Text Understanding with GETARUNS for Q/A and Summarization, Proc. ACL 2004 - 2nd Workshop on Text Meaning & Interpretation, Barcelona, Columbia University, pp.97-104.
Parsing Arguments and Adjuncts, Proc. Interfaces Conference, IEEE - ICEIS (the International Conference on Enterprise Information Systems), Pescara, 1-21.
Evaluating GETARUNS Parser with GREVAL Test Suite, Proc. ROMAND - 20th International Conference on Computational Linguistics - COLING, University of Geneva, 32-41.
con Antonella Bristot, Luminita Chiran, Ciprian Bacalu, Sara Tonelli, PARSING THE ORAL CORPUS AVIP/API (Progetto AVIP/API - Unità di Ricerca dell'Università "Ca' Foscari" di Venezia), Albano Leoni A., Cutugno F., Pettorino M., Savy R.(a cura di), Atti del Convegno "Il Parlato Italiano", M.D'Auria Editore, N08, 1-19.

2005

con Dario Bianchi, Learning Domain Ontologies from Text Analysis: an application for Question Answering, Proceedings of Workshop "Meaning 2005 - Developing Multilingual Web-Scale Language Technologies", Trento, 49-54.
Parsing Overlaps, in B.Fisseni, H.C.Schmitz, B. Schroeder, P. Wagner (Hrsg.), Sprachtechnologie, mobile Kommunikation und linguistiche Ressourcen, Sprache, Sprechen und Computer, Bd.8, Peter Lang, Frankfurt am Main, ISSN 1435-5736, ISBN 3-631-53874-X, pp.497-512.
Deep & Shallow Linguistically Based Parsing, in A.M.Di Sciullo(ed), UG and External Systems, John Benjamins, Amsterdam/Philadelphia, pp.335-374.
con Sara Tonelli, Marco Aldo Piccolino Boniforti, Antonella Bristot, Emanuele Pianta, VENSES – a Linguistically-Based System for Semantic Evaluation, RTE Challenge Workshop, Southampton, PASCAL - European Network of Excellence, pp. 49-52.
con Antonella Bristot, Marco Aldo Piccolino Boniforti, Sara Tonelli, Modeling Conversational Styles in Italian by means of Overlaps, AISV, CNR, Padova.
TEXT UNDERSTANDING FROM DISCOURSE MODEL AND INFERENTIAL PROCESSES, in C.N.Martinez, M.Moneglia(eds), Atti del Convegno "Computers, Literature and Philology" (CLIPS) 2003, pp.149-200.

2006

con Sara Tonelli, Marco Aldo Piccolino Boniforti, Antonella Bristot and Emanuele Pianta. VENSES - a Linguistically-Based System for Semantic Evaluation. In Quiñonero-Candela, J.; Dagan, I.; Magnini, B.; d'Alché-Buc, F. (Eds.), Machine Learning Challenges.

Lecture Notes in Computer Science, Vol. 3944, pp. 177-190, Springer Verlag.
con Antonella Bristot, Marco Aldo Piccolino Boniforti and Sara Tonelli, Another Evaluation of Anaphora Resolution Algorithms and a Comparison with GETARUNS' Knowledge Rich Approach, ROMAND 2006, 11th EACL, Trento, Association for Computational Linguistics, 3-10.
con Antonella Bristot, Marco Aldo Piccolino Boniforti and Sara Tonelli, Coping with semantic uncertainty with VENSES, in Bernardo Magnini, Ido Dagan(eds.), Proceedings of the Challanges Workshop - The 2nd PASCAL Recognizing Textual Entailment Challenge, 86-91, Università Ca' Foscari, Venezia.
Building Domain Ontologies from Text Analysis: an application for Question Answering, in Bernadette Sharp (ed.), Proceedings of the 3rd International Workshop for Natural Language Understanding & Cognitive Science, Cyprus, ICEIS, INSTICC Press, Portugal, 3-16.
Hybrid Systems for Information Extraction and Question Answering, Proceedings of CLIIR Workshop - How Can Computational Linguistics Improve Information Retrieval? -, COLING/ACL2006, Sydney, 1-8.

2007
con G. Nicolae, S. Harabagiu, C.Nicolae, A Linguistically-based Approach to Discourse Relations Recognition, in B.Sharp & M.Zock(eds.), Natural Language Processing and Cognitive Science, Proc. 4th NLPCS, Funchal, Portugal, Insticc Press, pp. 81-91.
con A. Bristot, M.A.Piccolino Boniforti, S.Tonelli, Entailment and Anaphora Resolution in RTE3, in Proc. ACL Workshop on Text Entailment and Paraphrasing, Prague, ACL Madison, USA, pp. 48-53.
con A. Bristot, S. Tonelli, Overlaps in AVIP/IPAR, the Italian Treebank of Spontaneous Speech, in Manuel Alcantara Pla & Thierry Declerk(Eds.), Proc. SRSL7 - Semantic Representation of Spoken Language, CAEPIA - Salamanca, pp. 29-38.
con Bristot A., Tonelli S., VIT - Venice Italian Treebank: Syntactic and Quantitative Features, in K. De Smedt, Jan Hajic, Sandra Kübler(Eds.), Proc. Sixth International Workshop on Treebanks and Linguistic Theories, Nealt Proc. Series Vol.1, pp. 43-54.
con G. Nicolae, S. Harabagiu, A Linguistically-based Approach to Detect Causality Relations in Unrestricted Text, in Proc. MICAI-2007, IEEE Publications, 173-185.

2008
con Jaber Suhel, Arabic Morphology Parsing Revisited, Proc. CICLing-2008 - Haifa, Israel, February 17-23, in Computational Linguistics and Intelligent Text Processing, LNCS, Springer Berlin / Heidelberg, 96-105.
Speech Synthesis for Language Tutoring Systems, in V.Melissa Holland & F.Pete Fisher(eds.), (2008), The Path of Speech Technologies in Computer Assisted Language Learning, Routledge - Taylor and Francis Group -, New York, pp. 123-150.
con Marco Aldo Piccolino Boniforti, Reranking GOOGLE with GReG, The 4th Web as Corpus Workshop: Can we do better than Google?, Marrakech, Morocco, LREC 2008, pp. 1-7.
Inducing Frames in the Italian Lexicon, in Rema Rossini Favretti(ed.), Frames, Corpora and Knowledge Representation, Bologna, Bononia University Press, pp.234-258.
curatore con Bos Johan, Semantics in Text Processing (STEP), Research in Computational Semantics, Vol.1, College Publications, London.
con E. Pianta, Answering Why-Questions in Closed domains from a Discourse Model, in Bos & Delmonte (eds.), STEP, pp. 109-114.
Semantic and Pragmatic Computing with GETARUNS, in Bos & Delmonte (eds.), STEP, pp. 287-298.

2009
con E. Pianta, Computing Implicit Entities and Events for Story Understanding, in H.Bunt, V.Petukhova and S.Wubben(eds.), Proc. Eighth International Conference on Computational Semantics IWCS-8, Tilburg University Press, pp. 277-281.
Treebanking in VIT: from Phrase Structure to Dependency Representation, in Sergei Nirenburg (ed.), Language Engineering for Lesser-Studied Languages, IOS Press, Amsterdam, The Netherlands, pp.51-80.
Computing Implicit Entities and Events with Getaruns, in B.Sharp and M.Zock (eds.), Natural Language Processing and Cognitive Science 2009, Insticc Press, Portugal, 23-35.
con A. Bristot, G. Voltolina, V. Pallotta, Scaling up a NLU system from text to dialogue understanding, Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing (SETQA-NLP 2009), NAACL, Boulder(USA), 7-11.
con A.Bristot, S.Tonelli, E.Pianta, English/Veneto Resource Poor Machine Translation with STILVEN, in International Review BULAG – Special Edition, Presses Universitaires de Franche-Comté, Besançon, pp.82-89.
A computational approach to implicit entities and events in text and discourse, in International Journal of Speech Technology (IJST), Springer, pp. 1-14.
con S.Tonelli, R. Tripodi, Semantic Processing for Text Entailment with VENSES, in Proceedings of Text Analysis Conference (TAC) 2009 Workshop - Notebook Papers and Results, NIST, Gaithersburg MA, pp. 453-460.