# BOLLETTINO DELLA SOCIETÀ DI LINGUISTICA ITALIANA

SLI

XXIV (2006) 1



Bollettino della Società di Linguistica Italiana (SLI), periodico stampato presso la Artigiana Multistampa (via Luca Valerio, 65 - 00146 Roma) per conto della SLI.
**Anno XXIV (2006), primo semestre (gennaio-giugno)**
Responsabile: Stefano Gensini. Reg. del Tribunale di Roma n. 312 dell'11 VII 1994. Sped. in abb. post. Legge 662/96 Art. 2 Comma 20/c Filiale di Roma

## SOCIETÀ DI LINGUISTICA ITALIANA

**Presidente:** Leonardo Savoia (fino al 2007, non rieleggibile)
e-mail: lsavoia@unifi.it

**Vicepresidente:** Martin Maiden (fino al 2006, non rieleggibile)
e-mail: martin.maiden@modern-languages.oxford.ac.uk

**Segretario:** Elisabetta Jezek (fino al 2008, rieleggibile)
Dipartimento di Linguistica Teorica e Applicata, Università degli Studi di Pavia, Strada Nuova 65, 27100 Pavia
Fax: 0382-984487; e-mail: jezek@unipv.it

**Tesoriere:** Monica Palmerini (fino al 2008, rieleggibile)
e-mail: monpalm@tin.it

**Comitato Esecutivo:**
Michela Cennamo (fino al 2006) <micennam@unina.it>; Mari D'Agostino (fino al 2008) <mdago@unipa.it>; Edoardo Lombardi Vallauri (fino al 2007) <lombardi@uniroma3.it>; Giovanna Massariello Merzagora (fino al 2008) <merzago@libero.it>; Davide Ricca (fino al 2007) <davide.ricca@unito.it>; Miriam Voghera (fino al 2006) <voghera@unisa.it>; Adriano Colombo, Segr. GISCEL <pof.6973@iperbole.bologna.it>; Federico Albano Leoni, Responsabile GSCP <federico.albanoleoni@uniroma1.it>; Gabriele Iannàccaro, Resp. GSPL <gabriele.iannaccaro@unimib.it> Giuliano Merz, Curatore del sito SLI <giuliano.merz@uibk.ac.at>

**Comitato per le Nomine:**
Rosanna Sornicola (fino al 2006) <sornicol@unina.it>; Teresa Poggi Salani (fino al 2007) <poggisalani@crusca.fi.it>; Giuliano Bernini (fino al 2008) <gbernini@unibg.it>

**Quote di iscrizione:**
quota ordinaria: euro 38,00 (+ 10,00 di immatricolazione per chi si iscrive per la prima volta);
quota studenti: euro 18,00 (+ 5,00 di immatricolazione);
quota per Istituti universitari: euro 73,00 (+ 21,00 di immatricolazione);
quota per Enti culturali, Biblioteche, ecc.: euro 110,00 (+ 31,00 di immatricolazione).
Le quote di associazione per i soci appartenenti ai paesi che non figurano nell'elenco riportato all'ultima pagina di questo bollettino sono ridotte alla metà.

**Modalità di iscrizione:**
mediante pagamento sul conto corrente postale n. 15986003, intestato a: Società di Linguistica Italiana, presso il Dipartimento di Studi Filologici, Linguistici e Letterari, Università La Sapienza, P.le Aldo Moro 5, 00185 Roma; oppure mediante pagamento con carta di credito e spedizione del modulo riportato in fondo al bollettino a: Società di Linguistica Italiana, Casella postale 2476, Roma 158.

☞ Per informazioni sulla propria situazione sociale o per segnalare variazioni di indirizzo o disguidi postali scrivere a: Monica Palmerini, monpalm@tin.it

BOLLETTINO DELLA
SOCIETÀ DI LINGUISTICA ITALIANA
(SLI)
XXIV (2006), 1

a cura di Elisabetta Jezek

www.societadilinguisticaitaliana.org

# CIRCOLARE n. 195 DEL PRESIDENTE

Il 21-23 settembre 2006 presso l'Università di Vercelli si terrà il XL Congresso Internazionale di Studi della Società. Come chiarisce il Comitato organizzatore, il Congresso, dedicato a *Linguistica e modelli tecnologici di ricerca*, vuole 'fare il punto sull'interazione della linguistica con i suoi strumenti di ricerca e [...] evidenziare le tendenze e le linee di evoluzione teorica che ne derivano'. In effetti nuovi strumenti di ricerca a partire dai primi decenni del '900 hanno potenziato l'indagine linguistica fornendo prove e modelli sperimentali via via più sofisticati e più certi, nei diversi campi della sociolinguistica, della linguistica computazionale, nella costruzione di data-base e di corpora. Un campo tradizionale di analisi strumentale è stato quello fonetico, che fin dalle ricerche spettrografiche di Jakobson, Fant e Halle ha influito sulla maniera di pensare i suoni delle lingue, confermando il tenore teorico e universalistico degli approcci linguistici. Le recenti tecniche di brain-imaging connettono in maniera interessante e suggestiva le ipotesi relative all'organizzazione della conoscenza linguistica con i corrispettivi neurocerebrali, delineando a loro volta uno stretto legame fra modelli sperimentali e teoria del linguaggio. Il Congresso promette quindi di affrontare questioni di grande attualità e in alcuni casi di avanguardia, la cui qualità e rilevanza scientifica sono garantite da un comitato scientifico e da un comitato organizzatore di alto profilo.

Il Congresso di Pescara, del 2007, *Alloglossie e comunità alloglotte nell'Italia contemporanea. Teorie, applicazioni e descrizioni, prospettive*, ci riporta all'altra componente della ricerca linguistica, ugualmente prevista e incoraggiata, direi coltivata, dalla SLI, che fin dal suo costituirsi ha avuto una particolare attenzione alle questioni di educazione linguistica e di politica linguistica, come abbiamo avuto modo di discutere in occasione del trentennale delle 10 tesi per un'educazione linguistica democratica. Il Congresso pescarese propone, quasi incalzato dalla situazione sociale italiana e in genere europea, un'occasione di riflessione di 'livello teorico e descrittivo in merito ad una delle componenti essenziali della nuova situazione plurilinguistica del paese'. Le alloglossie, antiche e recenti, e il sorgere di italiani L2 di contatto e di apprendimento configurano una comunità linguistica sempre più complessa e diversificata. In essa le leggi di tutela e le istituzioni, dalla scuola alla pubblica amministrazione, giocano un ruolo importante ma certamente non esauriscono il campo delle prospettive rilevanti. Emergono infatti dinamiche sociali e linguistiche per certi aspetti nuove rispetto a quelle delle alloglossie e delle differenze linguistiche tradizionali, che mettono in gioco l'identità delle persone, i flussi della globalizzazione, i nuovi processi e strumenti di comunicazione, su cui il comitato scientifico e quello organizzatore del XLI Congresso ci invitano a riflettere. Uno dei temi oggi più dibattuti riguarda appunto il destino delle diverse lingue in relazione alla loro distribuzione e diffusione sul territorio e, soprattutto, all'uso generalizzato dell'inglese americano in molti campi della comunicazione. In questo senso, la globalizzazione mette in gioco la salvaguardia di interessi collettivi primari, quali lo specifico patrimonio culturale e linguistico dei diversi paesi o gruppi sociali, mettendo in luce il contrasto fra le differenze linguistiche

e culturali e le esigenze dei poteri economici e politici, come discusso nella giornata di studio GSPL-AItLA del 31 marzo a Milano sulle politiche linguistiche in Europa. È importante aver presente a questo proposito che le diverse lingue parlate in Europa e nel mondo costituiscono un patrimonio comune di tutti i popoli che le parlano. Non a caso, dal 1999 l'UNESCO promuove per il 21 febbraio la Giornata Internazionale della Lingua Madre sullo stato delle lingue, con 'l'auspicio di una politica linguistica mondiale basata sul multilinguismo e garantita dall'accesso universale alle tecnologie informatiche', celebrata in Italia nel 2006 presso le comunità alloglotte della Calabria. I punti essenziali della situazione linguistica del pianeta (relativi al 2004) sono così sintetizzati dall'UNESCO:

> Auspicando la creazione di una politica linguistica mondiale basata sul multilinguismo per tutti, l'Unesco propone di celebrare ogni anno la lingua come strumento di conservazione del patrimonio culturale di ogni popolo. I dati sono infatti preoccupanti:
> - più del 50% delle 6000 lingue mondiali è in pericolo;
> - il 96% delle 6000 lingue mondiali è parlato dal 4% della popolazione mondiale;
> - il 90% delle lingue mondiali non è rappresentato su Internet;
> - una lingua scompare mediamente ogni 2 settimane;
> - l'80% delle lingue africane non ha l'ortografia;
> - la metà di tutte le lingue mondiali risiede in solo 8 paesi: Papua Nuova Guinea (832), Indonesia (731), Nigeria (515), India (400), Messico (295), Camerun (286), Australia (268) e Brasile (234);
> - i contenuti presenti sulla rete Internet sono per il 68.4% in inglese, seguito dal giapponese (5.9%), dal tedesco (5.8%) e dal cinese (3.9%).
> [...]
> Tra queste lingue [a rischio] troviamo lo Scots Gaelic (in Scozia), lo Saami (Svezia), l'Haida (Canada), il Kadazandusun (in Sabah, Malesia), l'Ainu (in Hokkaido, Giappone), il Sharda (in Srinagar, India), l'Idu Mishmi (in Arunachal Pradesh, India), il Cucapa (Messico) e il Tobas (in Argentina).

Le istituzioni scolastiche dei paesi europei si trovano quindi davanti al compito di rafforzare o introdurre le forme di bilinguismo / multilinguismo che potranno garantire non solo la ricchezza del patrimonio delle lingue attuali ma renderlo utilizzabile da un maggior numero di cittadini, in linea con l'esigenza di un'educazione all'accoglienza e alla tolleranza e insieme di un'educazione alla differenza e alla pluralità delle nostre identità. A questo proposito, ha notevole interesse, sia per la concezione del multilinguismo che presenta, sia in quanto suggerisce almeno alcuni degli orientamenti operativi dell'Unione Europea, la recente Comunicazione della Commissione Europea sul multilinguismo, nel quale il bi(/multi)linguismo è collegato ad una cultura più aperta e tollerante e alla valorizzazione delle nostre capacità cognitive:

**I.1 Multilinguismo e valori europei**

L'Unione europea è fondata sull' 'unità nella diversità': diversità di culture, usi, costumi e credenze – e di lingue. Oltre alle 20 lingue ufficiali dell'Unione *(21 con l'irlandese a partire dal 2007; 23 quando si aggiungeranno il bulgaro e il romeno)*, esistono più di 60 lingue autoctone e dozzine di lingue non autoctone parlate da comunità di migranti. È proprio questa diversità a fare dell'Unione europea quello che è: non un 'melting pot' in cui le differenze si fondono, bensì una casa comune in cui la diversità viene celebrata e le nostre numerose lingue materne rappresentano una fonte di ricchezza e fungono da ponte verso una solidarietà e una comprensione reciproca maggiori. La lingua è l'espressione più diretta della cultura, è quello che ci rende umani e conferisce a ognuno di noi un senso d'identità. L'articolo 22 della Carta dei diritti fondamentali dell'Unione europea precisa che l'Unione rispetta la diversità culturale, religiosa e linguistica. L'articolo 21 vieta qualsiasi forma di discriminazione fondata su numerosi motivi, compresa la lingua. Assieme al rispetto per l'individuo, all'apertura alle altre culture, alla tolleranza e all'accettazione dell'altro, il rispetto per le diversità linguistiche costituisce un valore fondamentale dell'Unione europea. L'iniziativa dell'Unione europea e degli Stati membri volta a sostenere il multilinguismo ha quindi un impatto diretto sulla vita di tutti i cittadini.

L'intervento legislativo, per quanto presentato dai suoi detrattori come una sorta di forzatura, concorre positivamente alla considerazione e alla stima del parlante nei confronti della propria varietà linguistica e rafforza il ricorso o il mantenimento del bilinguismo. Per quanto riguarda la situazione italiana, offrono interessanti elementi di valutazione i risultati raggiunti dall'applicazione delle leggi regionali e nazionali sulla tutela delle lingue minoritarie. In questo quadro, la linguistica, la sociologica e l'antropologia sono alla base di un'impostazione teorica corretta delle questioni relative al rapporto tra lingua e società e all'educazione linguistica, e favoriscono linee interpretative autonome e sufficientemente critiche rispetto alla usuale pianificazione linguistica basata su interessi economici e politici. Del resto la Risoluzione 12 della Conferenza generale dell'UNESCO del 1999, 'Attuazione di una politica linguistica mondiale fondata sul plurilinguismo', fissa i presupposti etici, culturali e scientifici di tale politica rinviando esplicitamente ai progressi delle scienze del linguaggio, '[...] notevoli progressi sono stati compiuti negli ultimi decenni dalle scienze del linguaggio [...]'. In particolare individua nella differenziazione delle lingue un valore da salvaguardare e un principio di tolleranza e di mutuo rispetto tra culture e popoli. Le possibilità di uso produttivo di più lingue sono molte e sta alla riflessione teorica e alla sensibilità democratica delle istituzioni trovare i modi e le basi concettuali per sostenere una visione non impositiva e conformata dell'uso linguistico delle persone. Questa attenzione per la diversità linguistica risponde quindi a fattori di ordine sociolinguistico e giuridico, nel senso che accettare la diversità linguistica e riconoscerne l'importanza significa in primo luogo promuovere i diritti linguistici in quanto parte fondamentale dei diritti di libertà universalmente ascritti agli esseri umani. Occorre inoltre tener

presente che vi è un aspetto sostanziale, relativo al valore intrinseco della diversità linguistica. A un livello più profondo infatti l'importanza della diversità linguistica risiede nel fatto che le diverse lingue corrispondono a sistemi possibili, le diverse grammatiche mentali, ammessi dalla nostra facoltà di linguaggio e sono espressione dei meccanismi cognitivi che regolano il funzionamento del linguaggio nella mente degli esseri umani. In questo senso la diversità linguistica è patrimonio dell'umanità.

# CIRCOLARE n. 194 DEL SEGRETARIO

**Candidature alle cariche sociali**

Cari Soci,

l'Assemblea annuale della SLI, che sarà convocata nel corso dei lavori del XL Congresso di Studi (Vercelli, 21-23 IX 2006), dovrà provvedere al rinnovo, a norma statutaria, di alcune cariche sociali della SLI.

Sono infatti giunti al termine del loro mandato il Vicepresidente Martin Maiden (non rieleggibile), il Presidente del Comitato Nomine Rosanna Sornicola (non rieleggibile) e i componenti del Comitato Esecutivo Michela Cennamo (non rieleggibile) e Miriam Voghera (non rieleggibile).

Ai sensi dell'articolo 18 dello Statuto, il Comitato Nomine mi ha comunicato le seguenti designazioni:

Vicepresidente: Max Pfister
Membro del Comitato Nomine: Giovanni Ruffino
Membri del Comitato Esecutivo: Annalisa Nesi e Giuliana Fiorentino

Ai sensi dell'articolo 18 dello Statuto sono possibili candidature alternative, che dovranno essere proposte al Segretario almeno da sei soci e almeno tre settimane prima della XL Assemblea.

Con un cordiale saluto

Elisabetta Jezek

# VERBALE DEL COMITATO ESECUTIVO DELLA SLI

Firenze, 21 aprile 2006

Venerdì 21 IV 2006, nella sala riunioni della Presidenza, presso la Facoltà di Lettere e Filosofia dell'Università di Firenze, in piazza Brunelleschi 4, alle ore 13.00 si riunisce il CE della SLI.

Sono presenti: Leonardo Savoia, Presidente; Elisabetta Jezek, Segretaria; Monica Palmerini, Tesoriera; Edoardo Lombardi Vallauri, Davide Ricca, membri del CE; Giacomo Ferrari e Silvia Dal Negro, rappresentanti del Comitato Organizzatore del XL Congresso SLI, Federico Albano Leoni, responsabile del GSCP, Adriano Colombo, Segretario nazionale Giscel.

Sono assenti giustificati: Martin Maiden, Vicepresidente; Rosanna Sornicola, Presidente del Comitato Nomine, Michela Cennamo, Mari D'Agostino, Giovanna Massariello Merzagora, Miriam Voghera membri del CE; Giuliano Merz, curatore del sito SLI; Gabriele Iannaccaro, responsabile del GSPL.

L'ordine del giorno è il seguente:

1. Comunicazioni del Presidente
2. Resoconto dell'attività dei gruppi SLI
3. XL Congresso Internazionale di Studi: Vercelli, 21-23 settembre 2006
4. Prossimi Congressi e Convegni
5. Ratifica bilancio societario relativo all'anno 2005
6. Definizione dell'O.d.g. della XL Assemblea dei Soci
7. Pubblicazioni e iniziative non congressuali
8. Proposta di impiego delle risorse SLI
9. Varie ed eventuali

## 1) Comunicazioni del Presidente

Non essendovi comunicazioni da parte del presidente, la parola è data ai rappresentanti dei gruppi della SLI per un resoconto delle attività.

## 2) Resoconto dell'attività dei gruppi SLI

Adriano Colombo, Segretario nazionale GISCEL, informa del buon successo dell'iniziativa congressuale tenutasi presso l'Università di Siena Stranieri lo scorso 6-8 aprile dal titolo "Lessico e apprendimenti". Vi è stata un'alta affluenza di pubblico, un'alta qualità degli interventi, anche se il programma affolato non ha facilitato la possibilità di dibattito. Informa che il prossimo convegno si terrà nel 2008 e che vi è una proposta di organizzazione da parte del GISCEL Lombardia. Riporta la situazione stagnante di alcuni gruppi regionali, ma anche la programmazione di nuove future sedi come Trento e Campobasso. Quanto alle pubblicazioni, si sta recuperando il ritardo accumulato nel periodo del cambio di casa editrice. Sono in uscita gli atti del convegno sulla formazione degli insegnanti tenutosi a Pescara nel 2001 (Questioni linguistiche

e formazione degli insegnanti, a cura di Domenico Russo) e i due volumi del convegno GISCEL di Lecce (2004). La cura del volume sul trentennale delle dieci tesi è problematica, anche per l'alto numero dei curatori. Riporta infine che la verifica delle iscrizioni dei soci GISCEL alla SLI non è avvenuta nel 2005 e dovrà essere quindi effettuata nel 2006.

Interviene quindi Federico Albano Leoni in qualità di responsabile del GSCP. Informa che sono pronti gli atti del convegno tenutosi nel 2004 a Padova, pubblicati in formato elettronico dalla casa editrice Liguori di Napoli nella collana "Quaderni di comunicazione parlata". Il convegno di Napoli del 2006 si è svolto alla presenza di molte figure internazionali: anche di questo convegno si prevede la pubblicazione degli atti in formato elettronico. Annuncia l'attivazione del sito www.comunicazioneparlata.org. e la programmazione di un prossimo convegno del gruppo da tenersi nel 2009, in una sede non ancora stabilita. Il gruppo allo stato attuale conta 137 aderenti.

La segretaria riporta ai presenti il contenuto del documento inviatole da Gabriele Iannaccaro, responsabile del GSPL, in cui è descritto lo svolgimento della giornata di studio del 31 marzo 2006 dal titolo "Quali politiche linguistiche per l'Europa e l'Italia?". La giornata di studio si è tenuta presso l'Università Milano Bicocca ed è stata organizzata dal GSPL in collaborazione con l'Associazione Italiana di Linguistica Applicata. Il presidente della SLI che ha partecipato all'iniziativa interviene commentandone il contenuto e sottolineando la presenza di un funzionario della Commissione Europea.

**3) XL Congresso SLI**

Giacomo Ferrari, a nome del Comitato Organizzatore e del Comitato Scientifico del XL Congresso SLI, comunica che i lavori preparatori procedono bene. Al Comitato Scientifico sono pervenute 101 proposte di comunicazione, il che rappresenta un buon risultato ma solleva anche un problema organizzativo. Si prevede di accettare 24 relazioni e 10 poster: entrambi troveranno spazio di pubblicazione negli atti. Il Congresso prevede anche 3 relazioni su invito: Andrea Moro, Roberto Busa SJ, Tullio de Mauro.

Durante le giornate del Congresso, è previsto uno spazio per l'Assemblea dei soci, venerdì pomeriggio, 22 settembre.

Tutte le informazioni saranno aggiornate sul sito del Congresso appositamente predisposto: www.lett.unipmn.it/sli2006/

A proposito dell'elevato numero di proposte, interviene Federico Albano Leoni sottolineando come la politica della SLI sia sempre stata quella di favorire la più ampia partecipazione possibile. Lombardi Vallauri ritiene che la presenza di (al massimo) due sessioni parallele possa essere tollerata. Anche Ricca ritiene che si possa pensare di avere una parte del congresso con sessioni parallele. Giacomo Ferrari dichiara di essere aperto a diverse soluzioni dopo aver sentito il papere del Comitato Scientifico. La segretaria fa presente agli organizzatori la scadenza per l'invio dei riassunti delle comunciazioni e dei poster da inserire nel

bollettino, fissata per il 15 maggio. Ferrari comunica che la sede congressuale potrà ospitare circa 150 congressisti.

**4) Prossimi Congressi e Convegni**

Il presidente dà lettura della lettera inviatagli da Carlo Consani a nome del comitato organizzatore del prossimo congresso SLI che si terrà a Pescara nel settembre 2007. La lettera contiene la bozza del temario, articolata come segue:

TITOLO:

Alloglossie e comunità alloglotte nell'Italia contemporanea.
Teorie, applicazioni e descrizioni, prospettive.

PRINCIPALI CONTENUTI:

A (linguistica)
Descrizione delle varietà alloglotte ai vari livelli di analisi;
Analisi constrastiva delle alloglossie in riferimento alle rispettive 'norme' standard;
Dinamiche linguistiche interne alle alloglossie dovute al contatto e/o alla condizione di isolamento.

B (sociolinguistica)
Dinamiche sociali e sociolinguistiche delle alloglossie;
Relativismo linguistico, lingue e culture, interculturalità;
Gruppi/comunità di parlanti, interazioni, rapporti con le istituzioni.

C
Italiano L2

Per quanto riguarda le modifiche in corso delle modalità organizzative dei Congressi SLI, il Comitato Organizzatore dichiara inoltre di non prevedere la nuova formula congressuale, e di preferire, per ragioni organizzative e al fine della migliore riuscita del congresso, il ricorso alla formula consolidata degli anni passati.

Ricca ritiene che sia ragionevole, oltre che opportuno in considerazione dell'orientamento dei colleghi che organizzeranno il congresso, pensare che il convegno 2007 mantenga ancora l'attuale struttura congressuale: la proposta concreta di modifica sarà infatti varata dall'assemblea di Vercelli soltanto nel prossimo autunno. Lombardi Vallauri osserva fra l'altro che diversamente si tratterebbe di una retrodatazione della proposta. I componenti del CE approvano l'articolazione del temario, che è stata modificata in sintonia con le osservazioni emerse nel corso dell'assemblea dei soci tenutati nel

settembre 2005 a Milano. Approvano inoltre, per le considerazioni appena riportate, l'impianto organizzativo del congresso proposto dal Comitato Organizzatore.

Albano Leoni sottolinea come il meccanismo organizzativo attuale dei congressi renda casuale la scelta del tema, che è lasciata all'autonomia delle sedi. Nota come vi sia stato negli ultimi anni una particolare attenzione all'aspetto della politica linguistica, il che può creare un rischio di disinteresse di una parte dei soci. Il presidente sottolinea come il tema della politica linguistica sia oggi molto sentito, in rapporto ad una complessa situazione sociolinguistica, e come ciò emerga dagli stessi interessi espressi dalle sedi. Albano Leoni ricorda le molte altre tematiche attuali, legate o attigue al linguaggio come biologia, psicologia, antropologia linguistica, neurolinguistica, non toccate di recente nei congressi. Ricca ricorda a questo proposito come l'assemblea di Milano abbia dato in sostanza esplicito mandato al CE di redigere una proposta di riorganizzazione congressuale, che preveda non soltanto un ambito tematico scelto dalla sede, ma anche un giorno e mezzo dedicato ai livelli di analisi linguistica e alle aree disciplinari, con una rotazione. In questo quadro, il CS in sede si occuperà della parte tematica, mentre per quella non tematica si ricorrerà a un gruppo di esperti scelti dal comitato nomine della SLI. Questa nuova struttura potrà eventualmente prevedere sessioni parallele (non più di due) e workshops. In questo modo si attua un riequilibrio tra impegno del CE e impegno della sede locale. Lombardi Vallauri osserva come in una struttura di questo tipo il CE possa mettere in atto una programmazione a lungo termine; Ferrari osserva come le grosse associazioni internazionali si muovano similmente. Colombo osserva come in questo modo cada il criterio della pertinenza al tema come strumento per contenere il numero delle proposte nei convegni, anche se, come osserva Ricca, negli ultimi tempi in realtà si sia registrata un'affluenza più lieve ai convegni. Albano Leoni ricorda come la SLI abbia organizzato nel passato, accanto ai convegni annuali, anche convegni interannuali su temi specifici, e che una struttura congressuale articolata in tema generale e tematizzazioni del tema generale attraverso attività satelliti (workshops ecc.) sia un modello molto praticato in grandi congressi, che però sono primariamente pensati come uno strumento per finanziare la società e darle visibilità. In questo caso non vi sono tempi così rigidi, i convegni durano più giorni ma a volte il livello scientifico ne risulta abbassato. Vallauri osserva come non sia necessario andare in questa direzione, e Albano Leoni sottolinea come la SLI abbia avuto fin dall'inizio una vocazione promozionale, che va mantenuta nella nuova struttura congressuale. Ricca consiglia una gradualità nell'operazione, anche per tenere conto del fatto, ricordato da Savoia, che alcuni soci sono contrari al cambiamento. Colombo osserva come la struttura congressuale attuale consenta alla SLI di fare il punto ogni anno di uno specifico ambito. Ricca vede la SLI come un necessario luogo di innovazione: è importante riproporre al centro dell'attenzione i temi classici della linguistica, individuare un comitato di esperti in grado di operare la selezione, che in sé non deve essere interpretata come un aspetto da evitare. Lombardi osserva come sia essenziale che nel documento di proposta della nuova formula congressuale siano

reintrodotte le aree tematiche, mentre altre questioni sono minori o secondarie. Per quanto riguarda la parte attuativa, Ricca caldeggia l'anonimato degli abstract. Viene quindi data lettura della proposta di modifica preparata, su richiesta dell'intero CE, da Michela Cennamo, Edorardo Lombardi Vallauri, Davide Ricca e Miriam Voghera (allegato 1 al presente verbale), già sottoposta per il giudizio in via telematica a un numero campione di soci. Si decide affinché tale proposta sia messa in approvazione all'assemblea di Vercelli.

**5) Ratifica del bilancio societario relativo all'anno 2004**

Viene presentato da Monica Palmerini il bilancio della SLI relativo all'anno 2005, che vede una voce attiva di 23.516 euro. Tale disponibilità permette di pensare ad un utile impiego del fondo disponibile. Al riguardo, si rinvia la discussione su possibili iniziative al punto 8 all'odg.

**6) Proposta dell'Ordine del giorno della XL Assemblea dei Soci**

Viene formulato il seguente Ordine del giorno per la XL Assemblea dei Soci:

1. Comunicazioni del Presidente.
2. Comunicazioni dei Rappresentanti dei Gruppi (Giscel, GSCP, GSPL).
3. Proposta di modifica dell'organizzazione del congresso annuale.
4. Prossimi Congressi e Convegni.
5. Elezione alle cariche sociali.
6. Ratifica del bilancio societario dell'anno 2005.
7. Pubblicazioni e iniziative non congressuali.
8. Impiego delle risorse SLI.
9. Varie ed eventuali.

**7) Pubblicazioni e iniziative non congressuali.**

Elisabetta Jezek comunica che le bozze degli Atti del XXXIX Congresso Internazionale SLI "Lo spazio linguistico italiano e le lingue esotiche: rapporti e reciproci influssi", a cura di Emanuele Banfi e Gabriele Iannaccaro, saranno inviate entro la fine di maggio all'editore Bulzoni.

Il primo volume della collana del premio per giovani studiosi intitolato a Monica Berretta, "Problemi di fraseologia dialettale", autrice Monica Cini, è uscito presso l'editore Bulzoni.

Adriano Colombo avanza una richiesta del Comitato scientifico delle pubblicazioni Giscel. Nel volume che nascerà dalla Giornata per il trentennale delle Dieci tesi, si desidera ripubblicare le rassegne bibliografiche apparse nei volumi SLI n. 31 (La linguistica italiana degli anni 1976-1986 - rassegna di C. Lavinio) e n. 44 (La linguistica italiana alle soglie del 2000 - rassegne di S. Ferreri e M. Vedovelli). La SLI, titolare della proprietà letteraria, dà la necessaria autorizzazione.

**8) Proposta di impiego delle risorse SLI**

Il CE affronta un tema di cui si è già discusso in più occasioni: nell'autunno 2004 attraverso una consultazione telematica, nel CE tenutosi a Firenze nella primavera del 2005 e nell'assemblea tenutasi a Milano nell'autunno 2005. Negli ultimi anni il bilancio della SLI presenta un consistente attivo, che sarebbe opportuno impiegare per utili iniziative scientifiche della società. Le proposte avanzate sono state le seguenti: borse di studio per giovani, una scuola estiva di linguistica, una rivista della Società.

Quanto alla scuola estiva, Savoia osserva che i problemi più rilevanti sono l'individuazione della sede e la molteplicità di iniziative di questo tipo. Una possibilità suggerita del presidente è quella di potenziare il bollettino dedicando un numero all'anno a interventi di carattere scientifico. Ricca osserva che si potrebbe utilizzare il bollettino per gli atti del congresso, anche in vista della nuova proposta congressuale. Albano Leoni avanza la proposta di promuovere la preparazione di un DVD che contenga tutti gli atti dei congressi SLI (o, per iniziare, una parte degli atti), con un motore di ricerca che consenta di effettuare ricerche mirate. A questo proposito il presidente si impegna a contattare l'editore Bulzoni per un preventivo, che potrebbe riguardare sia l'intera operazione, sia i costi di stampa e distribuzione su base di master. Lombardi Vallauri osserva che si possono prevedere sia premi per tesi di dottorato sia premi per tesi di laurea specialistica: eventuali tesi con dignità di stampa potrebbero anche essere pubblicate sul sito della società. Si rimanda la decisione definitiva all'assemblea di Vercelli in presenza del preventivo per il DVD.

**9) Varie ed eventuali**

Non vi sono varie ed eventuali.

Alle ore 17.30, esauriti gli argomenti all'odg, la riunione si conclude.

**Allegato 1**

**Proposta di modifica nell'organizzazione dei congressi SLI**

**1. Periodo e durata del convegno:**

Due giorni e mezzo, eventualmente estendibili a tre; tra settembre e ottobre.

**2. Articolazione generale del convegno:**

Il periodo (nel caso più tipico) di due giorni e mezzo sarà suddiviso nel modo seguente:

• **un giorno** (o due mezze giornate separate, se preferito dalla sede) **dedicato** esclusivamente al **tema definito** dalla **sede organizzatrice**, in accordo con il Comitato esecutivo e approvato dall'Assemblea, secondo la prassi finora seguita.

• **un giorno e mezzo dedicato** a **comunicazioni su temi diversi** da quello specifico proposto dalla sede, che tocchino **ambiti più ampi** di natura teorica e descrittiva (vedi punto **3**).

**3. Articolazione del giorno e mezzo non dedicato al tema definito dalla sede:**

Le **comunicazioni** verteranno **su uno o due ambiti disciplinari di grande ampiezza** (per es. **livelli di analisi**: fonetica, morfologia, semantica, sintassi, pragmatica, lessico, e/o **sottodiscipline** come sociolinguistica, acquisizione, dialettologia, psicolinguistica ecc.), indicati anno per anno dal Comitato esecutivo sulla base di un criterio di rotazione da un anno all'altro e di complementarità rispetto al tema più specifico proposto dalla sede. La selezione delle comunicazioni avverrà sulla base di *abstracts* sottoposti a gruppi di esperti settoriali (vedi punto **4**).

**4. Selezione degli *abstracts*:**

Il Comitato nomine indicherà un **gruppo di «esperti»** (tre–quattro nomi, rinnovabile con ricambi parziali a rotazione sul modello delle cariche sociali) **per ciascuno dei livelli di analisi / sottodiscipline** di cui al punto **3.** Ogni anno verranno richiesti di valutare le comunicazioni relative al tema "ampio" gli esperti degli ambiti disciplinari coinvolti. Per la valutazione degli *abstracts* relativi al tema definito dalla sede, continueranno a valere le procedure finora seguite.

**Eventuali *workshops* (ipotesi non vincolante)**

All'interno del giorno e mezzo non dedicato al tema definito dalla sede, potrebbero trovare posto uno o più ***workshops***: si tratterebbe di sessioni tematiche parallele al convegno principale e della durata di una mezza giornata - a invito o anche con contributi da scegliere attraverso un *call for papers* - proposte da un socio o da un gruppo di soci, d'intesa con la sede organizzatrice, al Comitato esecutivo entro il mese di marzo precedente il congresso. La proposta dovrebbe già contenere una definizione di massima del numero e degli argomenti dei singoli contributi. I soci proponenti avrebbero la responsabilità di organizzare e condurre tali sessioni, qualora approvate dal Comitato esecutivo, nonché di scegliere i contributi su invito e di selezionare gli *abstracts*. In questo quadro potrebbero anche rientrare eventuali iniziative dei gruppi.

# XL CONGRESSO INTERNAZIONALE DI STUDI DELLA SOCIETA' DI LINGUISTICA ITALIANA

Vercelli, 21 - 23 settembre 2006

Linguistica e Modelli tecnologici di ricerca
Sede dei lavori congressuali:

Complesso Conventuale S.Andrea
Università del Piemonte Orientale
Facoltà di Lettere e Filosofia
Via G. Ferraris, 116
13100 Vercelli

## GIOVEDI' 21 Settembre

8,30 – 9,30 **Registrazione**

9,30 **Indirizzi di saluto**

10,00 **Comunicazione su invito:**
A. Moro: Correlati neurali delle lingue possibili

**Pausa caffè**

11,30 V. Bambini: Le metodiche delle neuroscienze nello studio dei fenomeni pragmalinguistici. Dati fMRI ed ERPs sulla comprensione della metafora.
12,00 F. Carota: Dinamiche spazio-temporali dell'attività cerebrale durante la produzione e la percezione del linguaggio: nuove prospettive aperte dalla magnetoencefalografia.
12,30 F. Rosi: Le reti neurali come simulazione del processo di apprendimento della seconda lingua: il caso della morfologia tempo-aspettuale.

**Pranzo**

14,30 S. Montemagni: Aree fonetiche e lessicali toscane a confronto: prime elaborazioni computazionali dei dati dell'Atlante Lessicale Toscano.
15,00 S. Marzo – S. Vanvolsem: Tecniche statistiche per lo studio variazionale: l'italiano parlato nelle Fiandre (Belgio).

15,30 **Poster e caffè**

16,30 L. Battezzato: Coerenza morfologica e database elettronici: forme fantasma, problemi di morfologia, problemi di accento nel greco antico.
17,00 C. Bosco: Linguistic knowledge extraction from corpus parallel annotation.

17,30 R. De Felice - S. G. Pulman: Using clustering to improve adjective selection in English adjective-noun pairs.
18,00 M. Baroni - E. Guevara - V. Pirrelli: Induzione della struttura interna dei composti da corpora linguistici.
18,30 M.A. Piccolino Boniforti, M.Hadersbeck, R.Delmonte: The symbolic/statistical dichotomy: a new evaluation of parsing systems.

**VENERDI' 22 Settembre**

9,00 **Comunicazione su invito**
Roberto Busa SJ: Concludendo 60 anni di ricerche linguistiche con metodologie informatiche.

10,00 M. Nissim - J. Bos: Using the Web as a corpus in Natural Language Processing.
10,00 F. Cutugno - L. D'Anna: Limiti e complessità del recupero delle informazioni da tree-bank sintattiche.

11,00 **Pausa caffè**

11,30 A. Lenci - S. Zarcone: Un modello stocastico della classificazione azionale.
12,00 M. Mosca: Le espressioni spaziali in dialoghi task-oriented: un approccio cognitivo basato su corpora di parlato italiano.

**Pranzo**

14,00 L. Ribaldo - J. Gerbrandy - L. Lesmo: Quantifiers in Dependency Tree Semantics.
14,30 A. Mazzei - V. Lombardo: Toward a dynamic constituency model of syntax.
15,00 C. Gianollo - C. Guardiano - G. Longobardi - G. Rigon: Genealogie linguistiche e tassonomie computazionali. Per una 'storia e geografia della sintassi umana'.
15,30 P. M. Bertinetto - A. Lenci - S. Noccetti: Metodi quantitativi nell'analisi dell'acquisizione delle strutture tempo-aspettuali.
16,00 A. M. Di Sciullo: Asymmetry, the grammar, and the parser.

16,30 **Pausa caffè**

17 - 19,30 **Assemblea soci**

20,30 **Cena sociale**

**SABATO 23 Settembre**

9,00 F. Tamburini: Analisi automatica della prominenza frasale nella lingua parlata: un approccio acustico.

9,30 E. Magno Caldognetto: La partitura del parlato implementata in ANVIL: una metodologia per l'analisi della comunicazione multimodale faccia a faccia.

10,00 C. Avesani - M. Vayra: Spiegazioni fonetiche in fonologia: la sfida della Articulatory Phonology.

10,30 G. Marotta - E. Sardelli: Parametri prosodici per un modello semi-automatico di riconoscimento del parlante.

11, 00 **Pausa caffè**

11,30 **Comunicazione su invito**

T. De Mauro: Consuntivo dei lavori.

12,30 **Chiusura lavori**

**Elenco Poster**

| | |
|---|---|
| Allora A. | Presentazione del motore di ricerca EnTeR (Engine for Textual Reserachers) |
| Calderone B., Bertinetto, P.M. | La sillaba come stabilizzatore di forze fonotattiche. Per una modellizzazione |
| Delmonte R. | VIT – Venice Italian Treebank: caratteristiche sintattico-semantiche e quantitative |
| Dell'Orletta F., Lenci A., Marchi S., Montemagni S., Pirrelli V. | Text-2-Knowledge: uno strumento linguistico-computazionale per l'estrazione di conoscenza da testi |
| Giorgolo G. | Un modello per l'integrazione multimodale basato sulle grammatiche categoriali |
| Grimaldi M. | Cosa possono offrire le neuroscienze alla teoria linguistica? |
| Panunzi A., Fabbri, M. | Estrazione automatica di parole chiave da documenti in ambiente multilingue. Una infrastruttura linguistica nel progetto IST AXMEDIS. |
| Sansò A., Da Milano F., Mauri C. | La variazione linguistica in Europa attraverso un database tipologico. L'esperienza del Pavia Typological Database |
| Savy R., Castagneto M. | Funzioni Comunicative e Categorie d'Analisi Pragmatica: Dal Testo Dialogico allo Schema XML e Viceversa |
| Tini Brunozzi F., Quazza S., Zovato E. | Atti illocutivi e segnali discorsivi. Un contributo linguistico a un sistema TTS verso la sintesi vocale espressiva |
| Tomatis M. | SMORFIA: un analizzatore della morfologia verbale dell'italiano moderno per gli apprendenti di lingua italiana |
| Tonelli S., Delmonte R. | Knowledge-poor and knowledge-rich approach in anaphora resolution algorithms: a comparison |
| Velardi A., Plebe A. | Problemi teorici dei modelli: l'interazione tra approccio matematico e approccio cognitivo allo studio delle categorie semantiche |

# RIASSUNTI DELLE COMUNICAZIONI

**Cinzia Avesani e Mario Vayra**
Istituto di Scienze e Tecnologie della Congizione - CNR, Università di Bologna

## Spiegazioni fonetiche in fonologia: la sfida della Articulatory Phonology

Benché il parlato sia fondamentalmente "a set of movements made audible" piuttosto che "a set of sounds produced by movement" (Stetson, 1928), le scienze fonetiche sono state caratterizzate, fino alla metà degli anni Ottanta del secolo scorso, da una esplosione negli studi acustici e da una assai minore espansione in quelli articolatori. La disparità è riportabile, da una parte, alla facilità nel reperire strumentazioni e software per l'analisi acustica, dall'altra, agli alti costi, alla difficoltà di reperimento, e alla difficoltà d'uso delle strumentazioni articolatorie. Nonostante tali difficoltà, le analisi articolatorie e, più in generale, fisiologiche del parlato sono progredite negli ultimi anni ad una velocità straordinaria. Le tecniche di "imaging" (raggi x, ultrasuoni, tomografia, risonanza magnetica funzionale) hanno rivoluzionato il modo di osservare il tratto vocale, perché permettono di registrare movimenti interni al tratto senza imporre la presenza di sensori sulle strutture articolatorie. Hanno il grandissimo vantaggio di cogliere dati importanti del comportamento articolatorio che sono spesso difficili da ottenere, come, ad esempio, l'interazione di articolatori diversi e l'attività della lingua nell'area faringale. I sistemi a tracciamento ottico di trasduttori collocati sulla superficie degli articolatori visibili (labbra e mandibola) e interni (lingua) del tratto vocale hanno avuto, ad esempio, un forte impatto sulla nostra concezione di coarticolazione, rivelando relazioni tra articolatori che potevano in precedenza solo essere ipotizzate. Infine, anche tecniche che hanno una lunga tradizione di uso nelle scienze del parlato, come l'elettropalatografia e l'elettromiografia, sono state migliorate per fornire dati sempre più affidabili. Il massimo vantaggio nello studio delle basi fonetiche dei fenomeni linguistici naturalmente si trae quando si possa avere accesso ad una analisi sincronizzata di dati articolatori e acustici. Studi di questo tipo sono ancora rari nel panorama italiano. In questa sede desideriamo presentare i risultati di due nostri lavori condotti nel quadro della Articulatory Phonology. Uno degli assunti fondamentali di questa teoria è che le strutture fonologiche siano meglio intese come andamenti di "gesti articolatori": dove il gesto è insieme unità d'azione (fonetica) e d'informazione (fonologica). Abbiamo esaminato i processi che governano l'ordinamento seriale di vocali e consonanti in italiano e gli effetti della prosodia sulle proprietà segmentali. Usando i dati cinematici relativi ai "gesti articolatori "di apertura e chiusura del tratto vocale associati a labbra e mandibola, abbiamo analizzato, nel primo studio, gli effetti della struttura sillabica (sillaba aperta vs. chiusa) sulla coordinazione vocale-consonante; nel secondo, gli effetti della prominenza accentuale sulle proprietà acustiche e articolatorie della sillaba ai diversi livelli della gerarchia prosodica. I dati articolatori sono stati ottenuti con un sistema optoelettronico per la rilevazione tridimensionale di dati cinematici (ELITE), che permette anche la registrazione simulatanea del segnale acustico.

**Valentina Bambini**
Scuola Normale Superiore di Pisa, Laboratorio di Linguistica

## Le metodiche delle neuroscienze nello studio dei fenomeni pragmalinguistici. Dati fMRI ed ERPs sulla comprensione della metafora

INTRODUZIONE. Questa ricerca si muove al confine tra linguistica pragmatica e neuroscienze cognitive, indagando la comprensione della metafora attraverso metodiche di indagine proprie delle neuroscienze. Saranno presentati i risultati di due esperimenti complementari, realizzati testando i medesimi materiali con due diverse metodiche: risonanza magnetica funzionale e registrazione dei potenziali evento-correlati. Il fine è la descrizione nello spazio (in termini di aree cerebrali coinvolte) e nel tempo (in termini di onde cerebrali generate) dei processi cognitivi che presiedono all'elaborazione delle metafore, fornendo un supporto biologico alla modellizzazione linguistico-teorica dei fenomeni metaforici, e dei fenomeni pragmatici in generale. Fino a pochi decenni fa, le osservazioni cliniche rappresentavano l'unica via per indagare le basi neurali del linguaggio. Recentemente, lo sviluppo di tecniche non invasive per l'indagine in vivo del cervello ha rivoluzionato lo studio della mente. Le metodiche di neuroimmagine, soprattutto risonanza magnetica funzionale (fMRI) e tomografia ad emissione di positroni (PET), caratterizzate da elevata risoluzione spaziale, consentono di visualizzare le regioni cerebrali attivate durante lo svolgimento di determinati compiti. La registrazione elettroencefalografica, caratterizzata da altissima risoluzione temporale, consente di studiare con precisione l'andamento dell'attività elettrica a livello dello scalpo, identificando "onde" cerebrali legate a specifici compiti cognitivi (potenziali evento-correlati, ERPs). L'applicazione delle suddette tecniche allo studio del linguaggio ha aperto nuovi orizzonti di ricerca, rivelando che i processi linguistici coinvolgono diffuse reti neurali, ben oltre le regioni individuate dagli studi sugli afasici, e si snodano nel tempo attraverso precise fasi riflesse in precise componenti elettrofisiologiche [1, 2]. Tuttavia, molti degli studi finora realizzati hanno considerato processi linguistici piuttosto semplici (ad esempio, la lettura o la ripetizione di parole isolate), che non esauriscono il funzionamento della facoltà di linguaggio, in grado di operare simultaneamente a gradi diversi di complessità. Recentemente, si è imposta l'esigenza di indagare i processi linguistici in contesti di elaborazione ecologicamente plausibili, che riproducano il più possibile le situazioni comunicative reali [3]. Le neuroscienze del linguaggio cominciano ad affrontare la dimensione pragmatica dell'uso linguistico, intendendo per pragmatica l'integrazione tra processi di elaborazione della struttura linguistica, a vari livelli, ed ingredienti del contesto linguistico ed extra-linguistico. La presente ricerca ha selezionato un fenomeno che si colloca per necessità a livello pragmatico: la metafora. Lungi dall'essere prerogativa di poeti, la metafora pervade massicciamente la lingua d'uso, e sembra ragionevole supporre l'esistenza di meccanismi cognitivi preposti alla sua interpretazione, in grado di integrare significato letterale e conoscenze di sfondo (conformemente

al modello griceano). Recenti studi psicolinguistici [4], clinici [5], fMRI [6] ed ERPs [7] testimoniano un crescente interesse delle neuroscienze per la metafora, ma non forniscono un quadro unitario alla comprensione delle basi cognitive del fenomeno.

METODI. Sono stati reclutati 9 soggetti (5M/4F) per l'esperimento fMRI e 14 (7M/7F) per lo studio ERPs (destrimani, età 23-27 anni, istruzione universitaria). Il protocollo utilizzato differisce da quelli applicati in precedenza soprattutto per la creazione di un contesto di elaborazione maggiormente naturalistico, nonché per la particolare cura nella costruzione degli stimoli. Il disegno sperimentale si basa sul confronto tra coppie di frasi, di identica complessità sintattica, dove la medesima parola bersaglio (es. "squalo") appare usata in due significati diversi, letterale e metaforico. LETT: "Sai che cos'è quel pesce? Uno squalo." MET: "Sai che cos'è quell'avvocato? Uno squalo." Le frasi, bilanciate per frequenza d'uso, parametrizzate per ricchezza contestuale e familiarità della metafora, opportunamente mescolate a riempitivi, sono state presentate ai soggetti visivamente, con tempi di presentazione calibrati per ciascuna delle metodiche utilizzate. I soggetti, ignari dello scopo dello studio, sono stati istruiti a eseguire un compito di decisione semantica, scegliendo l'aggettivo che meglio si riferiva alla frase precedente (es. "feroce" vs "geografico"). Tale compito è servito a rendere più naturalistica la comprensione metaforica, evitando che i soggetti adottassero un atteggiamento metalinguistico.

RISULTATI fMRI. Rispetto alle frasi letterali, le metafore attivano maggiormente le seguenti regioni cerebrali: - aree di Broca e di Wernicke (giro frontale inferiore e giro temporale medio sinistri), tradizionalmente note per i processi linguistici, specie sintattici e semantici; - omologhi destri delle aree di Broca e Wernicke ed insula bilateralmente, recentemente messi in luce in compiti linguistici di vario tipo; - solco temporale superiore sinistro, precuneo bilateralmente ed altre aree, da riferirsi a processi cognitivi non prettamente linguistici, quali teoria della mente, immaginazione a base visiva, memoria episodica.

RISULTATI ERPs. Le parole con significato metaforico generano un'onda con polarità negativa intorno ai 400 ms (N400), caratterizzata da ampiezza maggiore rispetto alla condizione letterale, riferibile ad uno sforzo di elaborazione semantica. Inoltre, è stata riscontrata una significativa positività intorno ai 600 ms (P600), già osservata in compiti di rielaborazione linguistica a lunga distanza, specie sintattici, ma mai riscontrata finora per la metafora.

CONCLUSIONI. L'uso congiunto delle metodiche fMRI ed ERPs ha permesso di chiarire come la comprensione di una metafora sia un fenomeno cognitivo complesso, che poggia parallelamente su un'intensa elaborazione linguistica a vari livelli (testimoniata dall'attivazione di numerose regioni cerebrali reclutate a vario titolo in compiti linguistici, e riflessa dalla successione di N400 e P600) e su altre funzioni cognitive superiori (deducibili dall'attivazione di centri neurali legati a immaginazione, memoria, teoria della mente etc.). I dati sembrano smentire i modelli psicolinguistici che ipotizzano l'accesso diretto al significato

metaforico e riconfermare piuttosto il modello griceano, secondo cui la metafora, e gli usi non-letterali in genere, attraversano vari stadi di processing e richiedono costi di elaborazione aggiuntivi. Ciò che la linguistica definisce "livello pragmatico" sembra dunque possedere una realtà psicologica, riferibile tuttavia non ad un unico e compatto "modulo cognitivo", bensì a meccanismi di integrazione tra processi linguistici ed altre funzioni della mente.

Riferimenti bibliografici:

[1] Moro A. 2006. I confini di Babele. Longanesi.
[2] De Vincenzi M., R. Di Matteo. 2004. Come il cervello comprende il linguaggio. Laterza.
[3] Small S.L., H.C. Nusbaum. 2004. On the neurobiological investigation of language understanding in context. Brain&Language 89:300-11.
[4] Giora R. 2003. On Our Mind. Oxford University Press.
[5] Rinaldi M.C. et. al. 2004. Metaphor comprehension in right brain-damaged patients with visuo-verbal and verbal material: A dissociation (re)considered. Cortex 40:479-40.
[6] Lee S.S., M. Dapretto. 2005. Metaphorical vs. literal word meanings: fMRI evidence against a selective role of the right hemisphere. NeuroImage 29:536-44.
[7] Coulson S., C. Van Petten. 2002 Conceptual integration and metaphor: An event-related potential study. Memory&Cognition 30:958-68.

**Marco Baroni, Emiliano Guevara, Vito Pirrelli**
Università di Bologna - Dipartimento di Lingue, ILC - CNR, Pisa

**Induzione della struttura interna dei composti da corpora linguistici**

I composti costituiscono un'area centrale della ricerca linguistica teorica, applicata e computazionale. Uno degli aspetti più interessanti della composizione è la natura variabile delle relazioni sintattico-semantiche fra i costituenti dei composti. Da sempre, gli studi sulla composizione (Tollemache 1945, Marchand 1969, Levi 1978, ecc.) basano la classificazione delle parole composte sull'individuazione di relazioni sistematiche e ricorrenti fra i costituenti, approccio spesso tradotto in modelli "trasformazionalisti" in cui ogni struttura compositiva è correlata ad una relazione sintattica esplicita. Bisetto e Scalise (2005) propongono uno schema di classificazione il cui criterio principale è la relazione grammaticale non esplicita tra i costituenti del composto: le relazioni possibili sono Coordinazione, Subordinazione e Attribuzione (da cui i composti CRD, SUB e ATT). I composti SUB, del tipo "porta-ombrelli", presuppongono una relazione sintattica tra la testa e un suo argomento; nei composti CRD, come "attore cantante", i costituenti sono legati da un rapporto di coordinazione; nei composti ATT la testa seleziona una

relazione di modificazione spesso non prevedibile, come in "gentiluomo" o "pesce-spada". Presentiamo qui una serie di esperimenti in cui una classificazione alla Bisetto-Scalise di un campione di composti è indotta a partire dall'evidenza offerta da corpora testuali. La nostra ipotesi di base (collegata all'intuizione che vede nei composti una sorta di proto-sintassi, v. Lees 1960, Marchand 1969, Lieber 1992, Jackendoff 2002) è che, dato un corpus sufficientemente grande, sia possibile trovare corrispondenze tra un composto e le costruzioni sintattiche in cui co-occorrono i suoi costituenti. La nostra sperimentazione si basa su un campione di 90 composti SUB, CRD e ATT di tipo N+N dell'italiano e dell'inglese (ottenuti dal database MorboComp sviluppato al Dipartimento di Lingue dell'Università di Bologna), preclassificati manualmente seguendo Bisetto e Scalise, e sui corpora "la Repubblica" (380M di parole; Baroni et al. 2004) e itWaC (2 miliardi di parole; Baroni e Ueyama 2006) per l'italiano, ed il British National Corpus per l'inglese (100M di parole; Aston e Burnard 1998). Nel primo esperimento abbiamo cercato di stabilire se esiste una relazione fra il tipo di composto e l'ordine dei suoi costituenti quando questi ultimi co-occorrono in strutture sintattiche del testo. Poiché, in italiano come in inglese, i membri di una coordinazione non presentano un ordine fisso, mentre tendono a manifestarsi nell'ordine testa-dipendente nella subordinazione sintattica, ci aspettiamo che i costituenti di un composto SUB tendano a co-selezionarsi sintatticamente secondo un ordine relativamente stabile (testa-dipendente), mentre prevediamo un ordine sintattico meno stabile per i costituenti di un composto CRD. I costituenti di composti ATT, infine, dovrebbero mostrare una minore tendenza a co-occorrere in costruzioni diverse dalla composizione. Abbiamo dunque analizzato tutte le frasi nel BNC e in "la Repubblica" in cui la coppia di costituenti di un composto del nostro campione compare separata da almeno un token (eliminando dunque gli esempi in cui i costituenti sono parte del composto stesso). I risultati confermano, in entrambe le lingue, una minore tendenza dei costituenti dei composti ATT a ricorrere nella medesima frase, e una maggior tendenza dei costituenti dei SUB rispetto ai costituenti dei CRD a rispettare l'ordine testa-dipendente nelle costruzioni sintattiche. Il secondo esperimento intende gettare luce sulla natura degli elementi che si interpongono tra i costituenti di ogni composto nel testo. Sono stati ricercati tutti i "connector pattern", ovvero le sequenze da 1 a 5 tokens che separano i costituenti nel corpus itWaC (limitandoci ai contesti che appaiono all'interno di una frase), e si è analizzata la distribuzione dei connector patterns più frequenti rispetto ai tipi di composti. I risultati ottenuti confermano l'idea di una corrispondenza sistematica tra la classificazione SUB-CRD-ATT e le strutture sintattiche in cui si co-selezionano i costituenti dei composti. Nei due terzi dei casi i costituenti dei composti CRD hanno come connector pattern più tipici "e" o "," (virgola); nella metà dei casi i costituenti dei composti SUB sono connessi da "di", "del", ecc. (e in molti altri casi da altre marche tipiche di subordinazione, quali "per il"); i composti ATT, infine, non sono caratterizzati da nessun connector pattern tipico per una proporzione significativa della classe. Il terzo esperimento, infine, utilizza le medesime idee per attribuire una struttura

semantica a composti nuovi, anche non attestati nel corpus. L'idea è che se una coppia di nomi ricorre con connector pattern simili a quelli in cui occorrono i costituenti di un composto già conosciuto, e la coppia in questione viene usata per formare un composto, l'interpretazione di questo composto sarà simile a quella del composto conosciuto, secondo un plausibile meccanismo analogico-proporzionale. L'analisi, attualmente in corso d'ampliamento, indica una situazione di sostanziale coerenza fra l'impostazione classificatoria di Bisetto e Scalise (2005) e la distribuzione dei costituenti nel testo. I risultati della nostra ricerca, per quanto preliminari, mostrano come un modello di classificazione sviluppato sulla base degli strumenti d'indagine tradizionali della linguistica teorica e una metodologia basata sul data-mining di pattern estratti da corpora, lungi dall'essere in opposizione, possano complementarsi: il modello teorico ci ha suggerito una classificazione che è risultata motivata anche dal punto di vista distribuzionale, e l'analisi empirica automatizzata promette di fornire nuovi dati rilevanti per la riflessione teorica.

Riferimenti bibliografici:

Aston, G. e L. Burnard. 1998. The BNC Handbook. Edinburgh University Press.
Baroni, M. e M. Ueyama. 2006. Building general- and special-purpose corpora by Web crawling. NIJL International Workshop on Language Corpora.
Baroni, M., S. Bernardini, F. Comastri, L. Piccioni, A. Volpi, G. Aston e M. Mazzoleni. 2004. Introducing the "la Repubblica" corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian. LREC 2004.
Bisetto, A. e S. Scalise. 2005. The classification of compounds. Lingue e Linguaggio IV.2, 319-332.
Jackendoff, R. 2002. Foundations of Language. Oxford University Press.
Lees, R.B. 1960. The Grammar of English Nominalizations. Indiana University Press.
Levi, J. 1978. The Syntax and Semantics of Complex Nominals. Academic Press, New York.
Lieber, R. 1992. Deconstructing Morphology. University of Chicago Press
Marchand, H. 1969. The Categories and Types of Present-Day English Word-Formation. C. H. Beck.

**Luigi Battezzato**
Università del Piemonte Orientale

**Coerenza morfologica e database elettronici: forme fantasma, problemi di morfologia, problemi di accento nel greco antico.**

La relazione discuterà l'impiego di un database elettronico per risolvere due problemi di morfologia e di accentazione nel greco antico. Questo tipo di indagine non sarebbe stato possibile senza la presenza di strumenti elettronici

(TLG versione E). La sezione finale dell'intervento tratterà dell'interazione tra strumenti elettronici e strumenti filologici tradizionali. Nel greco antico, le parole non enclitiche hanno un peso prosodico minimo: non esistono parole monosillabiche non enclitiche terminati in vocale breve e anche le parole monosillabiche con sillaba breve e terminanti in consonante sono 'marginali' (Herodianus vol. I p. 403 Lentz; Devine e Stephens 1994). Questo significa che alcune forme verbali teoricamente possibili vengono evitate per non creare forme eccessivamente brevi. Il caso più notevole è quello dell'imperativo aoristo, molto comune, del verbo echo 'avere': sches, non *sche (cfr. invece enepo 'narrare': sia enispe che enispes, 'narra tu!', sono attestati). La norma in età classica agì sicuramente anche sui composti (ad es. parecho 'fornire'), come appare chiaro da esempi in poesia, garantiti dalla metrica. Lo studio sistematico di questo fenomeno non era possibile fino alla creazione del TLG (Thesaurus Linguae Graecae). Il TLG comprende tutti i principali testi letterari greci da Omero fino all'età imperiale, assieme a molti testi di età bizantina. Un'analisi mostra che il corpus compreso nel TLG comprende alcuni casi erronei ereditati da edizioni ottocentesche e novecentesche poco corrette, o poco attente a questa norma. La norma viene meno in età imperiale e le forme scorrette infiltrano la tradizione manoscritta bizantina di alcuni autori. La pratica editoriale registrata nel TLG varia, con casi di normalizzazione e altri di forme 'scorrette'. Verranno prese in esame anche le variazioni di accento comportate dalle due forme in concorrenza (quella 'regolare' paràsches, e quella che non rispetta la norma sul peso prosodico minimo, pàrasche, modellata sul presente). Un secondo campo d'indagine è quello dell'accentazione di monosillabi: una norma attestata dai grammatici antichi dice che la vocale finale 'alta' (iota o hypsilon) lunga di un nome terminante in consonante doppia viene abbreviata (ad es. kerux 'araldo' [nominativo], con u lunga, nonostante kerukos 'dell'araldo' [genitivo], con u breve) (Herodianus II 9. 4-33 Lentz; Sommerstein 1973, 177; Steriade 1988). La norma non vale per i monosillabi. La pratica editoriale registrata nel TLG anche in questo caso varia. Uno studio sistematico delle parole attestate nel corpus e delle variazioni editoriali soggiacenti permette di smentire alcune spiegazioni offerte (Schwyzer 1939, 391) e una pratica ortografica recentemente raccomandata (West 1990). La relazione si conclude con una discussione sulla relazione tra l'autorevolezza del testo elettronico e la ricerca linguistica e filologica. Il testo elettronico del TLG per molti soppianta il testo a stampa: poche biblioteche possiedono una collezione di testi così completa come quelli inseriti nel TLG (spesso da edizioni ottocentesche, o in tirature limitate). Questo strumento di lavoro ha cambiato le prospettive di ricerca linguistica e filologica, permettendo controlli sistematici sull'uso morfologico, sintattico e lessicale degli autori. Il TLG ha permesso di eliminare una serie di attestazioni fantasma o di forme errate che si erano introdotte nei vocabolari o nelle opere di consultazione generale realizzate nell'ottocento e nel primo novecento, ancora alla base della ricerca linguistica per il greco antico. D'altra parte, il processo di revisione dei testi favorito dal TLG porta a cambiare il TLG stesso, che, nelle sue più recenti versioni, ha abbandonato alcune vecchie

edizioni di autori importanti per adottarne di nuove, basate anche sulle nuove ricerche rese possibili dal TLG stesso.

Riferimenti bibliografici:

Devine, and Stephens 1994: A. M. Devine, and L. D. Stephens, The Prosody of Greek Speech. New York-Oxford, 1994
Lentz 1867: Herodiani Technici Reliquiae, collegit ... A. Lentz, Tomus I (Grammatici Graeci III.i)
Lentz 1868: Herodiani Technici Reliquiae, collegit ... A. Lentz, Tomus II (Grammatici Graeci III.ii. 1 et 2)
Schwyzer 1939: E. Schwyzer, Griechische Grammatik, München vol. I 1939
Steriade 1988: D. Steriade, "Greek Accent: A Case for Preserving Structure", Linguistic Inquiry 19.2 (1988) 271-314
Sommerstein 1973: A. H. Sommerstein, The Sound Pattern of Ancient Greek, Oxford 1973 TLG (# E)
West 1990: Aeschyli tragoediae cum incerti poetae Prometheo.

**Pier Marco Bertinetto, Alessandro Lenci, Sabrina Noccetti**
Scuola Normale Superiore di Pisa, Università di Pisa

**Metodi quantitativi nell'analisi dell'acquisizione delle strutture tempo-aspettuali.**

L'esistenza di una forte correlazione tra l'apprendimento delle categorie tempo-aspettuale e le proprietà semantico-azionali dei verbi è nota da tempo e ha ricevuto numerose conferme empiriche in varie lingue, tra cui anche l'italiano (Antinucci & Miller 1976, Li & Shirai 2000). Al tempo stesso i tratti azionali si intrecciano strettamente con le proprietà sintattico-distribuzionali e argomentali dei verbi. In particolare il ruolo della correlazione tra transitività e telicità, oltre che sul piano dei dati linguistico-tipologici, è stato messo in risalto anche a livello acquisizionale (Wagner 2005). Quello che emerge è dunque l'esistenza di un'interazione tra molteplici categorie linguistiche che agiscono su piani diversi (morfologico, sintattico e semantico), realizzando dinamiche complesse nei processi di acquisizione del sistema verbale. Tale quadro ripropone interrogativi importanti per il dibattito sull'ontogenesi di tali categorie, in modo particolare sullo status di primitivi cognitivi delle categorie azionali, e sulla misura in cui queste possano invece essere modellate come proprietà emergenti, risultanti da dinamiche di auto-organizzazione del sistema verbale. Allo scopo di investigare questo complessa rete categoriale, è stata realizzata una banca dati integrata, a partire dalle trascrizioni di tre serie di conversazioni tra bambini e adulti, due delle quali tratte dal corpus italiano del progetto CHILDES, ed una elaborata personalmente dal terzo autore (SN) secondo il sistema di codifica CLAN. Le produzioni linguistiche sono state

annotate su tre livelli di analisi: i.) morfologia e semantica verbale, ii.) morfologia nominale, iii.) sintassi. Questo lavoro rientra in un progetto di più ampio respiro, mirante a correlare lo sviluppo linguistico degli apprendenti di L1 sui tre livelli indicati, colmando una lacuna frequentemente osservata nei lavori sull'acquisizione delle strutture tempo-aspettuali, in cui si tende per lo più ad esaminare il problema dal solo punto di vista della semantica e morfologia verbale, estrapolando questo singolo aspetto dal contesto più ampio dell'acquisizione della competenza grammaticale. In questo lavoro, ci concentriamo in modo particolare sui risultati concernenti la semantica e morfologia verbale. I dati etichettati sono stati sottoposti a due tipi di analisi:
1. un'analisi estensiva di carattere statistico, mirante a far emergere le correlazioni tra le diverse categorie semantiche (sul piano del tempo, dell'aspetto e dell'azionalità) e le categorie morfologiche via via acquisite dai tre apprendenti, ovviamente in rapporto al comportamento degli adulti interagenti.
2. un'analisi computazionale condotta sui verbi più frequenti per ciascuna classe azionale attraverso l'applicazione di Multi-dimensional Scaling (MDS) e Self-Organizing Maps (SOMs) (Kohonen 1997). Attraverso il MDS è possibile ottenere una visione sinottica delle modalità secondo le quali verbi appartenenti a classi azionali diverse si dispongono in uno spazio multidimensionale di variabili quantitative, che registrano distribuzioni di tratti morfologici e sintattici. Le SOMs sono un modello neurale non-supervisionato, che consente invece di simulare il processo dinamico di formazione di proto-classi azionali che emergono dalle correlazioni statistiche presenti nell'input linguistico. Le due analisi sono complementari, in quanto vertono su aspetti diversi. La prima considera le categorie (morfologiche e semantiche) e le loro interazioni, cumulando il comportamento di tutti i verbi rappresentati nel corpus, indipendentemente dalla loro frequenza. La priorità è affidata, in questo caso, alle categorie di classificazione. La seconda analisi parte invece dalla selezione di un insieme di rappresentanti prototipici delle categorie azionali (in accezione vendleriana), per valutarne in concreto i rapporti di attrazione o repulsione, in relazione alle restanti categorie, morfologiche e semantiche, che attraverso di essi si manifestano. La priorità è affidata, in questo caso, a specifiche e individuali manifestazioni delle categorie azionali, con due importanti conseguenze. Da un lato, il rischio che i singoli verbi selezionati per l'analisi facciano riverberare le proprie idiosincrasie comportamentali; dall'altro lato, e per converso, la possibilità di osservare in un concreto campione testuale lo sgranarsi delle categorie azionali, la cui assoluta omogeneità di comportamento appare verosimilmente come una postulazione aprioristica, ragionevole e utile al fine di estrarre regolarità di fondo, ma indubbiamente bisognosa di verifica empirica. Il confronto tra i due tipi di analisi consente di mettere a fuoco una tematica cruciale per la ricerca linguistica, riferibile alla dialettica tra generalizzazione compiuta a partire da categorie analitiche postulate come omogeneamente distribuite, e emergenza delle medesime sulla base degli effettivi comportamenti linguistici, molto più variegati e sfaccettati, di quanto per tradizione non si sia portati a credere.

Riferimenti bibliografici:

Antinucci, F. e Miller, R. (1976), "How children talk about what happened", Journal of Child Language, III: 167-189.
Kohonen, T. (1997), Self-organizing Maps, New York, Springer.
Li, P. e Shirai, Y. (2000), The Acquisition of Lexical and Grammatical Aspect, New York, Mouton.
Wagner, L. (2005), "Aspectual bootstrapping in language acquisition: telicity and transitivity", Language Learning and Development, II: 51-76.

**Cristina Bosco**
Università di Torino, Dipartimento di Informatica

**Linguistic knowledge extraction from corpus parallel annotations.**

Natural Language Processing (NLP) and linguistics share the corpus-based approach, which consists in referring to large sets of naturally occurring linguistic data (i.e. corpora) previously annotated. Syntactically annotated corpora, aka treebanks, are currently considered as crucial resources for linguistics and NLP. For instance, by searching the Susanne corpus Sampson demonstrated that does not exist an absolute limit on the global left-branching factor of parse trees in English (Sampson, 2003). Instead, the major existing treebank, i.e. the English Penn Treebank (Marcus et al., 1993), has allowed for the development of state-of-the-art parsing techniques (e.g. Collins, 1999). Treebank annotations can differ in various aspects, making explicit different kinds of linguistic knowledge; for instance, they differ in the paradigms of representation, i.e. constituency or dependency, and refer to theoretical frameworks that exploit different kinds of syntactic units (phrases or words), and allow syntactic units to combine in different ways (as sub-unities or dependents). Several evidences experimentally supports the irreproducibility of results obtained on Penn Treebank on other treebanks or languages (e.g. on Czech (Collins et al. 1999), German (Dubey and Keller 2003), or Chinese (Levy and Manning 2003)). In order to investigate the causes of this, the comparison among annotations is currently a crucial issue. The paper presents an Italian treebank, i.e. the Turin University Treebank (TUT), where the comparison among formats is made explicit by parallel annotations. After a description of the treebank and its parallel formats, the paper focuses on cases of linguistic knowledge extraction that are both crucial for NLP and centred on specific issues on which the TUT formats are meaningfully comparable. 1.An Italian treebank project. The project started with the development of an Italian treebank, the Turin University Treebank (TUT, http://www.di.unito.it/~tutreeb/), by semi-automatic annotation process. Currently, TUT includes 1,800 sentences, i.e. 52,700 tokens, and 200 different dependency relations. TUT features a dependency-based annotation following the dependency grammar

(Hudson, 1984) and centred on a notion of morpho-syntactic-semantic grammatical relation which aims at represent the syntax-semantics interface by means of the Augmented Relational Structure (http://www.di.unito.it/~bosco/phdthesis.zip, Bosco and Lombardo 2003/2004). Moreover, TUT implements a trace-filler mechanism for pro-drop, equi phenomena, extractions and long distance dependencies, thus allowing for the recovery of predicate-argument structures also where deletions or movements happen. (1) is the representation of "il Governo di Berisha appare in difficoltà" (The Government of Berisha appears in trouble) in TUT format: (1) 1 Il (IL ART DEF M SING) [5;VERB-SUBJ] 2 Governo (GOVERNO NOUN COMMON M SING) [1;DET+DEF-ARG] 3 di (DI PREP MONO) [2;PREP-RMOD] 4 Berisha (IBerishaI NOUN PROPER) [3;PREP-ARG] 5 appare (APPARIRE VERB MAIN IND PRES INTRANS 3 SING) [0;TOP-VERB] 6 in (IN PREP MONO) [5;VERB-PREDCOMPL+SUBJ] 7 difficolta' (IdifficoltaI NOUN COMMON F ALLVAL) [6;PREP-ARG] 8 . (#\. PUNCT) [5;END] (The line pattern is: "P W (L PoS) [F;R]", where P = position of the word in the linear order of the sentence, W = word, L = lemmatized W, PoS = part of speech of W, F = position in the the sentence of the head from which W depends, R = dependency relation linking W to F.) 2.Comparison among formats by linguistic knowledge extraction We developed an automatic conversion process from TUT to Penn. Moreover, we devised two intermediate formats, which describe different layers of similarity/variation with respect to TUT and Penn in terms of richness/poorness of grammatical relations and constituency structure. The result is a set of four parallel treebanks, i.e. the same corpus annotated in four formats. Parallel annotations for the same corpus may serve as a suitable infrastructure for comparisons among different linguistic frameworks. The definition of a conversion process from A to B is in itself a comparison between A and B, since it involves a virtually complete and correct mapping which translate every analysis in A into the corresponding analysis in the linguistic framework B (Musillo and Sima'an, 2002). The experimental comparison among the four TUT formats is devoted to pinpoint representation problems and test the adequacy of information encoding. As case study we see the extraction of verbal sub-categorization frames from TUT formats. The experiment involves 830 lemmas and around 3,700 verb forms; as a gold standard for the evaluation of the extracted data, we assume the sub-categorizations stored in an Italian dictionary, i.e. DISC (Dizionario Italiano Sabatini Coletti). The best results (94% of matching sub-categorization frames) is reported for the extraction from the functionally richer annotation, regardless of the paradigm adopted (i.e. dependency or constituency). Other knowledge extraction concern to detection of the behaviour of adjuncts of verbs and nouns (7,700 around), e.g. their position with respect to the head and with respect to other adjuncts and complements, which is allowed in formats where the distinction complement/adjunct is both functionally and structurally drawn.

References:

Bosco C., Lombardo V. (2003) A relation-schema for treebank annotation. In A. Cappelli, F. Turini (eds.) Advances in Artificial Intelligence, LNCS 2829 –
Bosco C., Lombardo V. (2004) Dependency and relational structure in treebank annotation. In Proceedings of Workshop on Recent Advances in Dependency Grammar at COLING'04.
Collins M.J., Hajic J., Ramshaw L., Tillmann C. (1999) A Statistical Parser of Czech. In Proceedings of ACL 1999
Collins M.J. (1999) Head-driven statistical models for natural language parsing. Ph.D. thesis, University of Pennsylvania.
Dubey A., Keller F. (2003) Probabilistic parsing for German using sister-head dependencies. In Proceedings of ACL'03.
Hudson R. (1984) Word Grammar, Blackwell.
Levy R, Manning C. (2003) Is it harder to parse Chinese, or the Chinese treebank?. In Proceedings of ACL'03.
Marcus M., Santorini B., Marcinkiewicz M. (1993) Building a large annotated corpus of English: the Penn Treebank. In Computational Linguistics, 19:313-330.
Musillo G., Sima'an K. (2002) Towards comparing parsers from different linguistic frameworks. An information theoretic approach. In Proceedings of Workshop Beyond PARSEVAL - Towards improved evaluation measures for parsing systems at LREC'02.
Sampson G. (2003) Thoughts on two decades of drawing trees. In Abeillé A. (ed.) Building and using syntactically annotated corpora, Kluwer.

**Francesca Carota**
Institute for Cognitive Sciences

**Dinamiche spazio-temporali dell'attività cerebrale durante la produzione e la percezione del linguaggio: nuove prospettive aperte dalla magnetoencefalografia.**

Il costante avanzamento delle recenti tecniche di "neuroimaging" ha consentito l'esplorazione non-invasiva ed in-vivo delle funzionalità del cervello umano al lavoro nella molteplicità dei suoi livelli di organizzazione, che si articolano essenzialmente lungo gli assi dimensionali spaziale e temporale. Le principali tecniche attualmente in uso per l'indagine delle basi neurologiche del linguaggio si fondano sulla registrazione, al livello della corteccia cerebrale, dei segnali neurofisiologici associati a tali dimensioni. In particolare, il "neuroimaging" basato sulla risonanza magnetica funzionale (fMRI, functional magnetic resonance imaging) sfrutta i cambiamenti del livello del flusso sanguigno cerebrale connessi alle attività cognitive, privilegiando la dimensione spaziale dell'organizzazione cerebrale. Ne risulta l'acquisizione di neuroimmagini caratterizzate da un'elevata risoluzione spaziale, ma da una

bassa risoluzione temporale. In tal modo, le aree corticali recrutate da specifiche funzioni linguistiche possono essere localizzate con estrema accuratezza spaziale, come dimostra il vastissimo e tuttora crescente numero di lavori tesi a cartografare e dissociare i circuiti neuroanatomici che sottendono il linguaggio. Dall'insieme di tali studi -che si pongono ben al di là dell'assunto classico a favore della dicotomia tra produzione linguistica/area di Broca e comprensione linguistica/area di Wernicke- emerge un ampio panorama di "aree del linguaggio" tra loro funzionalmente interconnesse. Tali regioni includono il giro frontale inferiore sinistro (identificato tradizionalmente come area di Broca), il giro temporale posteriore superiore (o area di Wernicke), il lobo temporale anteriore, il lobo frontale mediano superiore sinistro, l'insula anteriore, il cervelletto, e la giunzione occipitale temporale inferiore sinistra (per uno stato dell'arte: Stowe et al., 2005). Ciascuna di tali aree appare potenzialmente deputata a supportare funzioni diversificate, sia linguistiche che non linguistiche. Per esempio, l'area di Broca risulta co-attivata da un mosaico di funzioni linguistiche legate sia alla produzione che alla percezione del linguaggio parlato (e.g. giudizi fonologici: Heim et al., 2003, semantici: Friederici et al., 2003a, e sintattici: Friederici et al., 2003b), e, più in generale, da funzioni comunicative sia verbali che non verbali. In particolare, l'area di Broca sembra contenere dei "mirror neurons" attivi sia nell'esecuzione che nell'osservazione dell'azione, simili a quelli riscontrati nell'area frontale F5 delle scimmia (Gallese e al., 1996). Inoltre, come evidenziano le attuali teorie motorie del linguaggio, l'area di Broca risulta essenziale nella comprensione dell'azione (Rizzolatti e Arbib, 1998), nell'imitazione dell'azione, consentendo l'accesso a modelli di rappresentazione interna del movimento (Meltzoff et Decety, 2003). Un ruolo particolare sarebbe giocato da tale area anche nell'imitazione di azioni orientate ad uno scopo (Koski et al., 2002). Inoltre, essa supporta la produzione e l'osservazione dei gesti manuali e orofacciali, cosa che suggerisce il collegamento tra linguaggio, processi sensoriali e movimento (e.g.: Gentilucci et al., 2000; Buccino et al., 2005). Dati sperimentali preliminari confermano l'interazione significativa tra linguaggio e azione motoria durante l'esecuzione di compiti comunicativi di deissi multimodale, o pointing verbale (manuale e oculare), e verbale o linguistico (focus prosodico e estrazione sintattica), sotteso dallo stesso fine pragmatico di portare un'informazione di rilievo all'attenzione dell'interlocutore (Carota et al., in stampa). Il quadro appena schematizzato, pur rivelando i loci delle attivazioni legate alle funzioni linguistiche e comunicative, si rivela meramente statico, in quanto lascia in disparte la componente temporale che è invece intrinseca all'attività cerebrale. L'inclusione della dimensione temporale costituisce il punto di forza delle tecniche che si basano sulla registrazione del segnale elettromagnetico derivante dall'attività elettrica delle cellule neuronali. Al di là dell'elettroencefalografia (EEG), che si basa sulla misura del potenziale elettrico conseguente all'attività post-sinaptica dei neuroni al livello dello scalpo ed è tradizionalmente impiegata nello studio della cronologia relativa degli eventi cerebrali, la magnetoencefalografia (MEG), che registra le variazioni nel campo magnetico al livello dello scalpo, si impone

attualmente come tecnica di interesse -sinora poco sfruttata- per l'approccio neuroscientifico al linguaggio. Essa permette di ottenere neuroimmagini caratterizzate da una soddisfacente risoluzione spaziale, e da una risoluzione temporale estremamente elevata, contribuendo a cogliere la dinamica delle attivazioni corticali lungo l'asse dello spazio e del tempo. L'integrazione dell'informazione spaziale e temporale è un requisito essenziale per la caratterizzazione e la comprensione del susseguirsi delle tappe che sottendono l'accesso e l'elaborazione del messaggio linguistico. La dinamica spazio-temporale apre prospettive nuove, da un lato, nella concezione dei moduli tradizionali del sistema linguaggio (dalla fonologia alla morfologia, dalla sintassi alla semantica passando attraverso il lessico, fino alla pragmatica), come è mostrato dagli studi recenti (Nishitani et al., 2004; Pulvermüller, 2005; Schurmann et al., in stampa). e delle loro interfacce, e, dall'altro, nella determinazione dei meccanismi di interazione tra i loci delle attivazioni corticali associate a specifiche funzioni linguistiche, introducendo un principio dinamico di causa-effetto nella valutazione degli eventi cerebrali. L'impatto della MEG sullo studio neuroscientifico della produzione e della percezione del linguaggio sarà discusso nel corso della comunicazione adducendo vari esempi tratti dal contesto di esperimenti attualmente in corso.

**Francesco Cutugno e Leandro D'Anna**
Università di Napoli "Federico II" - Dipartimento di Scienze Fisiche - Gruppo NLP, Università di Salerno - Dipartimento di Studi Linguistici e Letterari

**Limiti e complessità del recupero delle informazioni da tree-bank sintattiche.**

L'accesso all'informazione linguistica racchiusa in corpora di testo annotati a livello sintattico presenta difficoltà di natura teorica e, al contempo, di natura applicativa. Sul piano teorico si susseguono proposte per la definizione di categorie standard (metadati) nell'ambito delle teorie sintattiche che consentano la copertura del più vasto repertorio di fenomeni possibile in relazione alla variazione diafasica, alla tipologia di corpus in esame (i.e. scritto vs. parlato), alla capacità di mettere in relazione il livello di analisi sintattica del corpus con altri livelli di analisi linguistica. Allo stesso tempo, le teorie e le tecniche di rappresentazione della conoscenza linguistica, applicata al livello dell'analisi dei costituenti sintattici dell'enunciato, fanno ampio uso di strutture ad albero (binarie e non) con presenza di ricorsioni (spesso computazionalmente non decidibili), fanno ricorso a meccanismi atti ad alterare la rappresentazione della originale linearità del testo in favore di una maggiore potenza descrittiva delle forme assunte dai costituenti stessi. La linguistica del corpus è, però, una disciplina basata su grandi numeri, sul potere descrittivo di comportamenti coerenti osservati in numerose repliche di eventi linguistici, in cui la variazione statistica diventa capacità di previsione. Le problematiche fin qui esposte rendono dunque difficile manipolare grandi quantità di dati in quanto spesso

strutturati in maniera poco funzionale, in riflesso di una corrispondente confusione teorica e metodologica; rendono arduo e talvolta scarsamente ripetibile il lavoro manuale di annotazione da parte di più esperti; con la conseguenza che è difficile la generazione di un processo di annotazione automatico di grandi moli di materiale testuale. Nella pratica quotidiana, dunque rilevanti quantità di materiale linguistico vengono rappresentate con strutture-dati corrispondenti ad alberi di ricerca (tree-bank) e successivamente riversati in database dai quali viene poi estratta informazione tramite delle specifiche interrogazioni. La natura di questi database si è evoluta nel tempo: ai database relazionali (estremamente rigidi e poco adatti a fornire strumenti che aiutassero a superare le difficoltà sopra esposte) si sono progressivamente sostituiti database semistrutturati che hanno trovato ampio sviluppo grazie all'avvento del metalinguaggio standard XML. Interrogare una base di dati formata da alberi vuol dire essere in grado di riconoscere contemporaneamente le relazioni fra i livelli della struttura (sfruttando i concetti di nodo parente –i.e. padre, figlio ...). In questa ottica, uno dei vantaggi di XML è quello di fornire anche un linguaggio associato (Xpath) utile per la realizzazione di interrogazioni basate su operazioni di selezione e di filtraggio che utilizzano appunto il concetto di parentela nell'albero. Tuttavia non tutti i linguisti attualmente ricorrono a questi metodi di rappresentazione a causa di varie difficoltà: se, da un lato, XML non si presenta semplice e di immediato utilizzo da parte degli utenti meno esperti, gli utenti più esigenti denunciano limitazioni nel potere espressivo di alcune interrogazioni, scarsa efficienza dei processi estrattivi dovuta alla presenza di operazioni informatiche di alta complessità. Tutto questo ha portato alcuni studiosi (vedi Bird e altri citati) a proporre o una estensione del linguaggio Xpath o nuovi linguaggi di interrogazione di database semistrutturati rappresentanti tree-bank. In questo modo è dunque ora possibile aumentare l'efficienza delle interrogazioni, sia sul piano espressivo che su quello dell'usabilità, sviluppando apposite interfacce di interrogazione che rendono più facile l'accesso a queste tecniche anche ad utenti meno smaliziati. Questi sforzi, pur ragguardevoli, hanno aggirato i vari problemi, allontanandosi dallo standard definito con Xpath, implementando soluzioni algoritmiche proprietarie e non standardizzate, tradendo in una certa misura la missione del progetto XML e rischiando dunque la perdita di potere di generalizzazione. Le soluzioni adottate hanno di fatto trascurato completamente l'aspetto essenziale dell' Xpath ossia che esso comunque consente la raggiungibilità di ogni nodo in un albero, facendo uso di algoritmi a complessità lineare sul numero di nodi. In questo lavoro presenteremo un nuovo ed innovativo sistema di interrogazione dei tree bank interamente basato sull' XPath. Questa sistema si è mostrato adattarsi molto facilmente a varie tipologie di tree-bank disponibili all'interno del progetto nazionale "Parlare Italiano (PRIN 2004)". Esso è stato testato su tree-bank derivate dalla trascrizione di dialoghi, di parlato radiotelevisivo, di vari testi scritti presenti nel corpus "Penelope", di testi scritti da soggetti apprendenti l'italiano come L2 fornendo sempre la stessa interfaccia di interrogazione in modo tale da facilitare enormemente il compito dello studioso.

Grazie all'uso di questo applicativo è possibile interrogare le tree-bank estraendo informazioni di vario tipo quali: l'ordine dei costituenti nelle sequenze, il livello di ricorsione presente nell'albero, il numero, il tipo e la distribuzione statistica dei costituenti in differenti contesti. Particolare attenzione è stata infine posta nella determinazione della pesantezza dei sintagmi, argomento di rilevante interesse per il gruppo di sintatticisti a cui gli autori fanno riferimento.

**Rachele De Felice, Stephen G. Pulman**
Oxford University, Computational Linguistics Group

**Using clustering to improve adjective selection in English adjective-noun pairs.**

This paper addresses the natural language generation problem of automatically selecting the correct collocation for adjective-noun pairs in English when two or more nearly synonymous alternatives are available but only one is appropriate, e.g. tall(*high) man vs. high(*tall) status. The problem is particularly evident in machine translation between languages which use a different number of adjectives to express the same property, e.g. Italian and English. For example, the property of height is expressed by alto in Italian, but by both tall and high in English: the translation system must choose between the two adjectives, selecting the collocationally appropriate one for the given noun. There is also a more complex instance of this problem, conceptual translation ambiguity: a noun may collocate with different adjectives in the two languages, i.e. adjectives which are not translations of each other. This makes literal translations nonsensical or difficult to understand, e.g. nightmare which is brutto sogno ('ugly dream') in Italian, but bad dream (It. 'cattivo sogno') in English. We propose that Adj+N pairings can be predicted using cluster membership properties, based on the notion that adjective-noun collocations are not random, but follow certain patterns (e.g. tall and short with man, woman, girl; fake and real with diamond, leather, gold). We cluster nouns and assign adjective preferences to clusters rather than to individual nouns; in preliminary tests, our method achieves up to 78% accuracy. Furthermore, in solving the empirical problem we are presented with useful insights into the interaction of various parameters in clustering. The approach so far has been tested on a set of 28 common English adjectives often found in complementary distribution (e.g. strong tea vs. *powerful tea) and therefore susceptible to the translation ambiguities mentioned, such as strong – powerful, fake – false, big – large – great. To develop and test the system, a small corpus of British English was constructed from the British National Corpus. A set of nearly 200, 000 Adj+N pairs was extracted, comprising 2379 distinct nouns and their occurrences with one or more of the adjectives considered. As this method rests on the notion that patterns of adjective collocation exist and are predictable for a given set of nouns, a way of grouping the nouns was needed which would yield sets

exhibiting homogeneous behaviour with respect to adjective choice. Automated clustering was chosen as it avoids human bias while allowing underlying semantic properties to be captured, and enables the analysis of more data than is feasible manually. The intuition underlying the use of clustering is that if a set of nouns is observed to behave similarly in two contexts, in our case co-occurrence with a set of 250 nouns and 250 verbs, we can predict that they will behave similarly in a third context, too, namely co-occurrence with a given set of adjectives. We assigned the nouns to one of 100 clusters and through pointwise mutual information and chi-square tests established that there is a correlation between cluster membership and adjective choice: significant scores for a cluster identify instances where an adjective occurs significantly more or less often than expected with nouns of that cluster, pointing to the nouns' strong preference or dispreference for that adjective. Cluster membership is shown to be a good indicator of adjective choice and on this premise we developed a simple model for the translation of Italian Adj+N pairs into English. We assessed the model's performance against a baseline which used unigram probabilities only, i.e. the most frequent translation of the Italian adjective regardless of the co-occurring noun. The assumption is that with no other information, a translation system would use that value as a guide to lexical choice rather than choose randomly. 300 Italian Adj+N pairs were submitted for translation; we distinguished between Adj+N combinations not seen in our corpus, and those which were. We find that the clustering model significantly outperforms the baseline in two tasks (percentage of correct results): ClusteringModel – 75% Baseline – 63% Seen_data_only: ClusteringModel – 78% Baseline – 63% Unseen_data_only: ClusteringModel – 66.6% Baseline – 64% Despite the non-significant difference in the unseen data task, the results are positive as they show that the model overall is solid, and its two components can be used together without impairing its performance. Our model, unlike the baseline, successfully generates pairs like heavy (*strong) infection, tall (*high) man, big (*great) breath. We discuss the advantages the model has over the baseline in dealing with infrequent collocations, as well as possible causes for the drop in performance on the unseen data task. Our results show that the clustering approach is viable and successful. The 75% success rate confirms the notion that clustering is a useful procedure for NLP applications, even when performed with basic constraints, as done here. These results were obtained using a small cluster set, with some large clusters susceptible to noise. It is hypothesised that more refined clustering – varying parameters such as number of clusters, feature window size, feature number/type, consideration of syntactic dependencies – could lead to a higher success rate. The usefulness of clustering for target-word selection in MT has been previously suggested (e.g. Dagan and Itai, 1994; Kikui, 1999). Our results reiterate these findings, and show their validity specifically for the domain of the translation of Adj+N pairs, which has so far received little attention. The clustering model is a helpful addition to NLG, and offers an improvement that is not only quantitative – as expressed by the significant difference in accuracy – but also qualitative, since, compared to the baseline, it

extends the range of Adj+N pairs that are correctly translated. Lapata et al. (1999) suggested that it would prove difficult to generalise adjective preferences to unseen data, but we show that this is not the case, achieving an encouraging success rate of 66.6% in preliminary tests. Finally, the paper also briefly surveys some recent work in the related field of 'generation-heavy' MT (Habash and Dorr, 2002), where most of the work is done at target language level – an approach found especially appropriate for language pairs with few parallel corpora and other resources used in 'traditional' stochastic MT. Our research supports this claim, showing that a successful translation module can be implemented in the absence of bilingual corpora, given sufficient target language information.

**Anna Maria Di Sciullo**
Université du Québec à Montréal

**Asymmetry, the grammar, and the parser.**

Asymmetry, the grammar, and the parser Anna Maria Di Sciullo, Philippe Gabrini, Calin Batori,Stanca Somesfalean Université du Québec à Montréal UG is the theory of the knowledge of language, and UP is the theory of parsing. UG and UP share objects of inquiry, one of these objects is asymmetry (i.e., the irreversibility of the members of a pair in a set, e.g., precede, dominate, asymmetric c-command in a tree), understood as a central property of the language faculty, of the linguistic expressions, and of the computational model. We take the theory of the knowledge of language and the theory of the recovery of this knowledge to be part of the same scientific paradigm, viz., bio-linguistics. 1. Asymmetry is central to both theoretical and computational linguistics. Strong hypotheses on the asymmetry of linguistic relations have been formulated within generative grammar (Chomsky 1981, 1995, 2001, 2005; Kayne 1984, 1994, 2005; Moro 2000; Di Sciullo 2003, 2005). Computational implementations of asymmetric relations are available (Marcus 1970, Berwick and Weinberg 1984; Berwick, Abney, Tenny 1991, Berwick 1985, 1991; Delmonte 2005; Di Sciullo 2001; Di Sciullo and Fong 2001, Fong 1991, 2005; Harkema 2005; Philips 1995). The questions raised by this paper are the following: How does the parser interpret efficiently the grammar? How does it efficiently parse linguistic expressions? A plausible answer to the first question requires means to restrict the actions of the parser. A plausible answer to the second question requires means to reduce the search space of the parser. We approach these questions in the light of the following hypothesis: (1) UG/UP Hypothesis Universal Grammar (UG) is designed to optimally analyze linguistic expressions in terms of asymmetrical relations. Universal Parser (UP) is designed to optimally recover natural language asymmetries. (Di Sciullo 1999). The UG/UP Hypothesis posits that there is a perfect match between the grammar and the parser enabled by asymmetry. The linguistic expressions derived by the operations of UG, and the recovery of asymmetric relations by the parser is what makes linguistic

expressions interpretable by the external systems. The hypothesis in (1) is both empirically testable and computationally implementable. It implies that points of symmetry lead to non-optimal matches between the grammar and the parser, and thus non-optimal interpretations of linguistic expressions. Assuming the UG/UP hypothesis, we describe the properties of an LL(1) parser, which is a computational implementation of the Asymmetry Theory (Di Sciullo 2005). The architecture of the parser is such that it limits the search space while it incrementally recovers the asymmetric structure from the input. We focus on the parsing of simple affirmatives, passives and questions in order to show that the parser recovers these structures, including the relations between the elements in the functional domain (CP/DP), operators, modifiers and arguments, and their reflexes within the VP/NP. First, we describe the properties of Asymmetry Theory and Asymmetry Grammars. Second, we show how the grammar is used by the parser and discuss three parses in order to illustrate how the parser incrementally recovers asymmetric relations. Finally we summarize the results. 2. Asymmetry Theory is a theory of grammar that extends the Derivation by phase model (Chomsky 2001, 2005; Uriagereka 1999) to a fully parallel model of UG. The generic operations of the grammar (Shift (2), Link, (3)) apply in the derivation ): Given b, a of linguistic expressions under asymmetric Agree, (4). (2) Shift ( projected from ?. (Di d) derives a new object b, a, Shift (b and a two objects a, b and a): Given two objects b, a Sciullo 2005: 29) (3) Link (are b and a), where b, a) derives the object (b, a, Link (b sister-containing featurally related. (Di Sciullo 2005: 29) (4) Agree (j1, j2): Given two sets of features j1 and j2, Agree holds between j1 and j2, iff j1 properly includes j2, and the node dominating j1 sister-contains the node dominating j2. (Di Sciullo 2005: 30) Agree requires that the proper subset relation holds between the constituents that undergo the operations for structure building (Shift), and linking (Link). One consequence of Agree is that each head in a projection chain asymmetrically selects its closest sister-contained head, thus deriving Homogeneous Projections (HP). The semantic relations between the constituents, including semantic scope, are derived by Shift and Link, whereas Flip, (5), contributes to linearization. (5) Flip (T): Given a minimal tree T, Flip (T) is the tree obtained by creating a mirror image of T. (Di Sciullo 2005: 30) In this theory, the generic operations of the grammar are customized to apply in different derivations, yielding different sorts of linguistic objects, syntactic and morphological. Asymmetry grammars are specific actualizations of the generic operations of the Asymmetry Theory. 3. The parser. We describe an implemented parser that computes asymmetric relations and assembles phrase structure, respecting the left-to-right nature of parsing. The incremental parser processes the input of a word at a time, producing one or potentially more partial parses at each step. Each time a word is introduced, it extends the analyses produced so far. We define a strict incremental parser to be one that is monotonic in its output. That is, partial parses may only be augmented (no transformation or deletion is allowed) with the introduction of each succeeding word. The model includes mechanisms to implement efficient parsing, without backtracking or

unnecessary search of the derivational history. The parser interprets the operations of the grammar as applying groups of rules in local domains. The rules of the domains (CP, DP) define the maximal realizations of syntactic projections; the rules of the groups (Grp CP, Grp DP, Grp PP, etc.) exhaustively enumerate the potential realizations of domains. (6) Dom CP Proj CmaxP Spec [getw(word, 'cat') == 'WHadj'] : shift, link(FPP.Spec); H [word.get_cat() == 'Vaux'] : shift, link(FauxP.h); Cmpl shift;#TP ... Grp DP ( PName | Dpron ... The most inclusive domain is the CP domain which is open by default at the beginning of the parse. Conditions, including Agree, apply before an action (shift, link, flip) is taken. After the first word has been matched with a category of the grammar, the next word's category is recovered from the numeration and is matched against the next available positions in the group. If no match is found, no parse is generated. The above procedure is repeated until a word is found that does not belong to the current group. The group is then closed, and together with it any domain it may have triggered. Following the projection path, the last domain to close is the first opened, i.e. the maximal CP domain. Thus, a common noun is recognized as a head and is first shifted with a determiner to form a DP, before this DP is shifted in the specifier of the TP. Subsequent linking of this DP to positions in the more inclusive domains, including the VP domain, derives the effects of movement from left to right. We show that the parser makes an optimal use of the operations of the grammar, while it reduces the search space by providing a parse for simple declaratives, passives, and questions, and it recovers as well the right asymmetric relations between different sorts of adjuncts they may include. 4. The parser efficiently interprets the grammar by restricting the operations of the grammar to apply in local domains, recovering asymmetric relations therein. It efficiently parses linguistic expressions by limiting the search space to the asymmetric relations available under Agree. The processing of linguistic expressions is based on the recovery of the asymmetric relations of UG (see Tsapkini, Jarema and Di Sciullo (2004) for experimental results), the incorporation of these relations in parsing can only improve their efficiency by bringing them closer to human performance.

References:

Berwick, R. and A. Weinberg. 1984. The Grammatical Basis of Linguistic Performance. Cambridge Mass.: The MIT Press.
Chomsky, N. 2005. On phases. To appear in C. Otero et al. (eds.) Foundational Issues in Linguistic Theory. Cambridge, Mass.: The MIT Press.
Chomsky, N. 1995. The Minimalist Program. Cambridge, Mass.: The MIT Press.
Chomsky, N. 2001. Derivation by phase. In M. Kenstowicz (ed.) Ken Hale: A Life in Language, Cambridge, Mass.: The MIT Press. Pp. 1-52.
Chomsky, N. 1981. Lectures on Government and Binding . Dordrecht: Foris Publications.
Delmonte, R. 2005. Deep & shallow linguistically based parsing. In A.M. Di Sciullo (ed.) UG and External Systems. Amsterdam: John Benjamins. Pp. 335-374.

Di Sciullo, A.M. 2005. Asymmetry in Morphology. Cambridge, Mass.: The MIT Press.
Di Sciullo, A.M. 2003. Morphological Relations in Asymmetry Theory. In A.M. Di Sciullo (ed.) Asymmetry in Grammar. Volume 2: Morphology, Phonology and Acquisition. Amsterdam: John Benjamins. Pp. 9-36.
Di Sciullo, A.M. 2001. Strict Asymmetry Theory and Morpho-Conceptual Parsing. World Multiconference on Systemics, Cybernetics and Informatics. Orlando, Florida. Pp. 481-487.
Di Sciullo, A.M. 2000. Parsing Asymmetries. In Natural Language Processing. Springler Computer Science Press. Pp. 24-39.
Di Sciullo, A.M. 1999. An Integrated Competence-Performance Model, A Prototype for Morpho-Conceptual Parsing and Consequences for Information Processing. In Proceedings of VEXTAL. Venice: Università Ca`Foscari. Pp. 369-379.
Di Sciullo, A.M. and S. Fong. 2005. Morpho-syntax parsing. In A.M. Di Sciullo (ed.) UG and External Systems. Amsterdam: John Benjamins. Pp. 247-268.
Di Sciullo, A.M. and S. Fong. 2001. Asymmetry, Zero Morphology and Tractability. The 15th Pacific Asia Conference on Language Information and Computation. University of Hong Kong. Pp. 61-72.
Fong, S. 2005. Computation with probes and goals. In A.M. Di Sciullo (ed.) UG and External Systems. Amsterdam: John Benjamins. Pp. 331-334.
Fong, S. 1991. The computational implementation of principle-based parsing. In R. Berwick, S. Abney and C. Tenny (eds.) Principle-Based Parsing. Dordrecht: Kluwer. Pp.65-82.
Harkema, H. 2005. Minimalist languages and the correct prefix property. In A.M. Di Sciullo (ed.) UG and External Systems. Amsterdam: John Benjamins. Pp. 289-310.
Harkema, H. 2001. Top-down recognition of minimalist grammars with look-ahead. Ms. UCLA.
Kayne, R. 2005. Movement and Silence. New York: Oxford University Press.
Kayne, R. 1994. The Antisymmetry of Syntax. Cambridge, Mass: The MIT Press.
Kayne, R. 1984. Connectedness and Binary Branching. Dordrecht: Foris Publications.
Marcus, M. 1970. A Theory of Syntactic Recognition for Natural Language. Cambridge, Mass.: The MIT Press.
Moro, A. 2000. Dynamic Antisymmetry. Cambridge, Mass.: The MIT Press.
Niyogi S. and R. Berwick. 2005. A Minimalist implementation of Hale-Keyser incorporation theory. In A.M. Di Sciullo (ed.) UG and External Systems. Amsterdam: John Benjamins. Pp. 269-288.
Phillips, C. 1995. Right Association in Parsing and Grammar. In C. Schütze, J. Ganger and K. Broihier (eds) Papers on Language Processing and Acquisition. MITWPL 26, 37-93.
Tsapkini, K., G. Jarema and A.M. Di Sciullo. 2004. The role of configurational asymmetry in the lexical access of prefixed verbs: Evidence from French. Brain and Language 90, 143-150.

Uriagereka, J. 1999. Multiple Spell-Out. In S. D. Epstein and N. Hornstein (eds.) Working Minimalism. Cambridge, Mass: The MIT Press. Pp. 251-282.
Weinberg, A. 1999. A Minimalist Theory of Human Sentence Processing. In S. D. Epstein and N. Hornstein (eds.) Working Minimalism. Cambridge, Mass.: The MIT Press. Pp.283-315.

**Chiara Gianollo, Cristina Guardiano, Giuseppe Longobardi, Gabriele Rigon**
Università di Trieste, Università di Modena e Reggio Emilia, Università di Trieste, Università di Pisa - Università di Trieste

## Genealogie Linguistiche e Tassonomie Computazionali. Per una 'Storia e Geografia della Sintassi Umana'

Nell'ambito degli studi sulle filogenie computazionali sono stati elaborati, a partire dagli anni Sessanta, sofisticati metodi che permettono di rappresentare in forma di grafi ad albero generati meccanicamente relazioni complesse fra entità comparabili. Tali procedimenti, già applicati con successo in biologia al trattamento dei dati relativi alla variazione genetica e alla distribuzione di specie e popolazioni (anche umane [1]), sono stati recentemente impiegati, nell'ambito della linguistica storica, per il trattamento di dati lessicali, allo scopo di ricostruire e rappresentare i rapporti filogenetici fra le lingue [e.g. 2,3]. In questo lavoro intendiamo mostrare che è possibile ottenere risultati interessanti e innovativi nella ricostruzione delle relazioni filogenetiche fra le lingue sfruttando sinergicamente il potenziale offerto dai metodi computazionali, rigorosi e sperimentalmente riproducibili, e le sofisticate teorie parametriche elaborate nel dominio della sintassi formale. I parametri sintattici infatti, in quanto discreti, finiti, profondamente deduttivi rispetto ai dati superficiali e, forse, caratterizzati da maggiore stabilità (tendenziale inerzia al mutamento [4]), appaiono in linea di principio più adatti a misurare realisticamente la distanza storica fra le lingue (e a suggerire modelli generali di distribuzione diacronica e diatopica) di quanto lo siano le somiglianze fonetiche nel lessico. In questo senso, mentre i dati lessicali, su cui si basano i metodi comparativi classici, hanno uno statuto simile ai caratteri morfologici esterni (e.g. antropometrici) utilizzati dalle tassonomie biologiche tradizionali, i dati sintattici, espressi in forma di parametri, sembrano avere statuto più simile ai dati molecolari utilizzati dalla genetica delle popolazioni. Allo scopo di verificare la significatività genealogica della comparazione di dati sintattici elaborati con metodi computazionali è stato raccolto un corpus di dati parametrici provenienti da un totale di 24 lingue antiche e moderne, le cui relazioni sono a priori ben note. Tali dati sono stati poi rappresentati in una "griglia parametrica" nella quale ogni lingua (cioè ogni unità tassonomica) è definita da una stringa di stati, ciascuno corrispondente ad un preciso valore parametrico [5,6,7,8]: essendo i parametri concepiti come caratteri binari, gli stati che possono essere assunti basicamente sono + o -. Il valore di questi stati è puramente oppositivo, e in nessun caso uno

dei due corrisponde a una condizione ancestrale o marcata. Inoltre, a causa delle implicazioni tra valori parametrici, nei casi in cui un parametro subordinato ad un altro (o ad una condizione esterna) non possa essere fissato significativamente, il suo valore è codificato con 0 (o 0+, 0-) e viene considerato come blank nelle stringhe, quindi inutilizzabile ai fini della comparazione. Tali dati sono stati successivamente trattati ed elaborati mediante i nuovi metodi computazionali utilizzati in biologia [9]. Per identificare gli alberi che sembrano migliori in relazione a determinate ipotesi evolutive e all'evidenza empirica, sono stati adottati diversi metodi filogenetici. Ben motivate idealizzazioni e ipotesi sulla natura dei dati portano ad escludere i metodi per i quali gli stati dei caratteri non siano paritari (ad esempio, tra i metodi di parsimonia, Camin-Sokal parsimony, Dollo parsimony, Polymorphism parsimony) o debbano essere necessariamente definiti (Compatibility method), mentre risultano più adatti al trattamento dei dati parametrici i metodi a matrice di distanza. Per calcolare le distanze a partire dalle stringhe di stati sono stati selezionati due criteri che sembrano i più significativi: la distanza frazionaria (FD), e la fuzzy Hamming distance (FHD). Indicando con i il numero di identità, d quello delle differenze e n il numero totale di parametri comparabili per una coppia di lingue, la FD corrisponde a d/(d+i), mentre la seconda distanza, presentata nella sua forma più recente in [10], corrisponde a (d+n-i)/2, e consiste nella media tra una Hamming distance e una defective Hamming distance (che prende in considerazione anche i blanks delle stringhe). Entrambe le distanze sono state calcolate una prima volta in relazione alla lista completa di parametri, ed una seconda volta in relazione ad un loro subset, in cui vengono presi in considerazione solo i parametri la cui fissazione sembra essere diacronicamente più stabile e, pertanto, più significativa a livello genealogico. Quindi le matrici di distanza fornite in input ai programmi sono quattro, e sulla base di ognuna viene ricostruito uno specifico albero filogenetico. Gli algoritmi utilizzati per elaborare le matrici appartengono al pacchetto di programmi PHYLIP, messo a disposizione da Joseph Felsenstein [11], e sono il Fitch-Margoliash, (compreso nel programma Kitsch) e UPGMA (Unweighted Pair-Group Method Using Arithmetic Averages, compreso nel programma Neighbour). Entrambi producono alberi rooted e si basano sull'ipotesi del molecular clock, per il quale la "quantità di cambiamento" in ogni linea evolutiva è costante. Nel caso del Fitch-Margoliash la computazione è finalizzata a minimizzare la somma di quadrati: Sommatoria (i, j): (n(i,j)*((D(i,j) – d(i,j))elevato a 2))*(D(i,j)elevato a p) D(i,j) è la distanza osservata tra le lingue i e j, mentre d(i,j)corrisponde alle distanze attese, cioè alla lunghezza dei rami che separano i e j. n(i,j) e p (power) nel nostro caso corrispondono a 1 e 2 (cf. [12]). L'albero migliore è, appunto, quello in cui la differenza fra le distanze osservate e la lunghezza dei rami ricostruiti risulta minima. UPGMA invece è una procedura tipica della cluster analysis: consiste nel congiungere per prima la coppia di lingue con la distanza minore, calcolare una nuova matrice considerando la coppia come unica unità tassonomica e procedere in questo modo fino al completamento dell'albero. A procedura ultimata, viene prodotto un totale di 8 alberi filogenetici, derivanti

dalle 4 matrici di distanza. Gli 8 alberi sono rielaborati in un unico albero di consenso con il programma Consense, compreso in PHYLIP, in base alla regola di maggioranza estesa. Le relazioni filogenetiche fra lingue che emergono da quest'ultima elaborazione corrispondono in modo significativo alle ipotesi genealogiche tradizionali formulate sulla base del metodo comparativo classico. Tali incoraggianti risultati suggeriscono che il metodo, modellato e calibrato sulla base di relazioni filogenetiche conosciute, possa essere in futuro applicato con successo a lingue le cui relazioni genealogiche sono più misteriose, a causa dell'impossibilità di applicare il metodo comparativo classico e dei risultati difficilmente falsificabili offerti da altre procedure, come la mass comparison (cf. [13], [14]).

Riferimenti bibliografici:

[1] Cavalli Sforza, L. Luca, Paolo Menozzi e Alberto Piazza, 1994. The History and Geography of Human Genes. Princeton NJ: Princeton University Press.
[2] Ringe, Don, Tandy Warnow, Ann Taylor, 2002. Indo-European and Computational Cladistics. Transactions of the Philological Society 100.1, 59-129.
[3] McMahon, April e Robert McMahon, 2003. Finding families:quantitative methods in language classification. Transactions of the Philological Society 101.1, 7-55.
[4] Keenan, Edward, 1994. Creating anaphors. An historical study of the English reflexive pronouns. Ms. UCLA.
[5] Longobardi, Giuseppe, 2003. Methods in Parametric Linguistics and Cognitive History. Linguistic Variation Yearbook, 3, 101-138.
[6] Guardiano, Cristina e Giuseppe Longobardi, 2005. Parametric Comparison and Language Taxonomy. In Grammaticalization and Parametric Variation, ed. by Montserrat Batllori, Maria-Lluïsa Hernanz, Carme Picallo, and Francesc Roca, 149-174. Oxford: Oxford University Press.
[7] Gianollo, Chiara, Cristina Guardiano, Giuseppe Longobardi, 2004. Historical Implications of a Formal Theory of Syntactic Variation, paper presented at DIGS VIII, Yale. In press in Proceedings of DIGS VIII, ed. by Stephen Anderson and Dianne Jonas. Oxford: Oxford University Press.
[8] Longobardi, Giuseppe, 2005. A Minimalist Program for Parametric Linguistics?, in H. Broekhuis, N. Corver, M. Huybregts, U. Kleinhenz, and J. Koster (eds.) Organizing Grammar: Linguistic Studies for Henk van Riemsdijk, Mouton de Gruyter, Berlin/New York, 407-414.
[9] Rigon, Gabriele, 2005. Parametric comparison and computational phylogenies. Tesi di Laurea, Università di Trieste.
[10] Sgarro, Andrea, 2005. A few non-metric "effective" Hamming distances. Ms. Università di Trieste.
[11] Felsenstein, Joseph, 2004. Inferring phylogenies. Sunderland, MA: Sinauer
[12] Fitch, W. M., Margoliash, E., 1967. Construction of phylogenetic trees, Science 155, 279-284.

[13] Greenberg, Joseph, 1987. Language in the Americas. Stanford CA: Stanford University Press.
[14] Greenberg, Joseph, 2000. Indoeuropean and its Closest Relatives: the Eurasiatic Language Family. Stanford CA: Stanford University Press.

**Alessandro Lenci, Alessandra Zarcone**
Università di Pisa, Dipartimento di Linguistica

**Un modello stocastico della classificazione azionale.**

Sebbene l'Aktionsart sia spesso citato con il nome di "aspetto lessicale", i tratti semantici intrinseci di un verbo sono solo uno degli elementi utili a determinare il suo comportamento azionale nei contesti linguistici. La presenza di complementi, la loro definitezza, avverbiali di varia natura (temporali, modali, ecc) e i tratti morfologici stessi del verbo interagisco in maniera complessa per realizzare l'effettivo valore azionale dell'evento descritto da una frase. Ad esempio, se dal punto di vista lessicale leggere esprime un'azione atelica e durativa (Gianni ha letto tutto il giorno), la presenza di un complemento diretto è in grado di trasformare l'evento in telico se il sintagma nominale è singolare definito o indefinito (Gianni ha letto un libro), ma non se è un plurale senza articolo (Gianni ha letto giornali di sport tutto il giorno). Similmente, un verbo telico e puntuale come arrivare (Il treno è arrivato alle 5 in punto) può apparire in contesti in cui l'evento è in realtà durativo, come in Gli ospiti sono arrivati per ore, in cui è il soggetto plurale a indurre l'interpretazione iterativa e durativa della particolare situazione descritta. In generale, i casi di "ibridismo azionale" (Bertinetto 1986) – ovvero il fatto che uno stesso verbo possa avere valori azionali diversi a seconda del suo contesto linguistico - sollevano il problema di come modellare la complessa interazione dei fattori costituivi dell'Aktionsart. Infatti, per nessuna classe azionale sembra possibile selezionare un insieme di tratti la cui presenza in un contesto sia congiuntamente necessaria e sufficiente a garantire che l'evento venga interpretato come appartenente a quella particolare classe. Inoltre, gli stessi tipi azionali non si presentano come entità monolitiche, bensì come categorie che contengono rappresentanti verbali prototipici resistenti a variazioni contestuali, accanto invece a verbi che più facilmente danno luogo a fenomeni di ibridismo azionale. L'ipotesi che vogliamo esplorare in questo lavoro è che l'interpretazione del valore azionale di un verbo in contesto possa essere modellato come il risultato di un complesso processo di integrazione di vincoli linguistici di natura ibrida, che agiscono come "soft constraints" probabilistici. A tale scopo, presentiamo un modello computazionale stocastico della classificazione azionale di verbi italiani basato sul principio della massimizzazione dell'entropia (Maximum Entropy; cf. Berger et al. 1996, Dell'Orletta et al. 2005). Maximum Entropy

è un metodo di apprendimento automatico (machine learning) molto usato in linguistica computazionale e nel trattamento automatico del linguaggio per creare modelli stocastici per la disambiguazione linguistica (morfosintattica o semantica), ma a nostra conoscenza mai applicato al problema specifico dell'interpretazione azionale. Bisogna inoltre aggiungere che, in generale, l'Aktionsart non ha ricevuto molta attenzione in linguistica computazionale; un'eccezione è rappresentata dal lavoro di Siegel & McKeown (2000), che però non affronta il problema dell'ibridismo azionale. In questo lavoro riportiamo alcuni esperimenti relativi alla creazione di un modello stocastico per l'individuazione della classe azionale di un verbo in un particolare contesto. Il metodo della Maximum Entropy richiede la selezione preventiva di un repertorio di "features" linguistiche potenzialmente rilevanti per la classificazione. A partire da un "corpus di addestramento", l'algoritmo di apprendimento automatico stima il peso probabilistico di ciascuna feature e le combina in un modello stocastico integrato per l'assegnazione della classe azionale appropriata in ogni contesto. Per i nostri esperimenti abbiamo selezionato 33 verbi italiani da TreSSI (Montemagni et al. 2003), un corpus dell'italiano contemporaneo annotato a livello morfosintattico e sintattico. Dal momento che la Maximum Entropy è un algoritmo di apprendimento automatico supervisionato, ciascuna delle 3443 occorrenze dei verbi in TreSSI è stata annotata manualmente rispetto al valore azionale espresso dal verbo nella frase. Le classi usate per l'annotazione sono "state", "process", "achievement", "accomplishment" e "gradual_completion". I seguenti sono invece i tipi di features linguistiche selezionate: 1. tratti morfologici della testa verbale; 2. presenza di complementi; 3. tratti morfosintattici e sintattici dei complementi (es. definitezza, numero, ecc.); 4. tratti semantici degli argomenti (es. animatezza); 4. diatesi del verbo; 5. presenza nella frase di avverbiali appartenenti a varie tipologie (es. temporali delimitativi, frequentativi, decorrenziali, ecc.). L'ampia tipologia di tratti usati nella nostra indagine ci consente quindi un'esplorazione ad ampio spettro del processo di interpretazione azionale. In questo lavoro presenteremo i risultati di esperimenti condotti con diverse combinazioni di tratti linguistici, allo scopo di valutare i parametri più rilevanti nella classificazione azionale dei verbi. Per ciascun modello riporteremo sia una valutazione quantitativa – verificando la capacità acquisita dal modello di assegnare la corretta classe azionale a un verbo in un contesto – sia soprattutto una valutazione qualitativa, finalizzata a indagare il ruolo di specifici tratti linguistici nell'interpretazione azionale, e ottenuta ispezionando i pesi probabilistici stimati attraverso la Maximum Entropy. L'uso di metodi stocastici di apprendimento automatico permette di gettare nuova luce sul tema dell'azionalità verbale. In particolare, diventa possibile creare un modello più dinamico dell'Aktionsart, in grado di catturare l'effetto congiunto di fattori multidimensionali nel determinare il comportamento azionale di un verbo, fattori a loro volta rappresentati come vincoli probabilistici radicati nella distribuzione dei dati linguistici.

Riferimenti bibliografici:

Berger, A., Della Pietra, S., Della Pietra, V. (1996), "A maximum entropy approach to natural language processing", Computational Linguistics 22(1): 39-71.
Bertinetto, P.M. (1986), Tempo, Aspetto e Azione nel verbo italiano. Il sistema dell'indicativo, Firenze, Accademia della Crusca.
Dell'Orletta, F., Lenci, A, Montemagni, S., Pirrelli V., (2005), "Climbing the Path to Grammar: a Maximum Entropy Model of Subject/Object Learning", Proceedings of the ACL 2005 Workshop on Psychocomputational Models of Language Acquisition, Ann Arbor, USA.
Montemagni, S. et al. 2003. Building the Italian syntactic-semantic treebank. In Abeillé A. (ed.) Treebanks. Building and Using Parsed Corpora, Kluwer, Dordrecht: 189-210.
Siegel, R., McKeown, K.R. (2000), "Learning methods to combine linguistic indicators: improving aspectual classification and revealing linguistic insights", Computational Linguistics, 26(4): 595-628.

**Emanuela Magno Caldognetto**
Istituto di Scienze e Tecnologie della Congizione - CNR, sezione di Padova

**La partitura del parlato implementata in ANVIL: una metodologia per l'analisi della comunicazione multimodale faccia a faccia.**

In una interazione comunicativa faccia-a-faccia il parlante che produce un enunciato, realizzato oralmente grazie all'attività del sistema pneumo-fono-articolatorio, trasmesso per via acustica e percepito nella modalità uditiva, può produrre contemporaneamente gesti e movimenti trasmessi per via ottica e percepiti nella modalità visiva (Magno Caldognetto e Poggi 2001). Dal punto di vista metodologico, nell'analisi degli scambi comunicativi, lo studioso dovrà affrontare due problemi analitici, connessi tra loro: - la descrizione esaustiva dei sistemi di comunicazione modo-specifici, studiati separatamente e indipendentemente l'uno dall'altro; - la descrizione delle sincronizzazioni dei segnali che sono all'opera nell'interazione multimodale, che non necessariamente sono in rapporti di simultaneità, ma anche di anticipazione o di perseverazione tra le diverse unità nei diversi segnali. Con questa serie di dati si potranno definire le interazioni tra i diversi sistemi all'interno di un atto di comunicazione complesso e formulare le regole di pianificazione, cioè le regole in base alle quali questi diversi sistemi di comunicazione interagiscono gli uni con gli altri e insieme cooperano a costituire il complesso atto finale di comunicazione. Per descrivere la multimodalità della comunicazione, è stato suggerito e applicato un metodo (Magno Caldognetto e Poggi 1994, 2001; Poggi e Magno Caldognetto 1996, Magno Caldognetto et al., in corso di stampa): la Partitura, un sistema per trascrivere e analizzare i segnali

multimodali classificati sia separatamente che nella loro interazione reciproca. Questo metodo permette di individuare gli elementi comunicativi trasmessi simultaneamente da: contenuto verbale del parlato, prosodia, gesto, movimenti di bocca, occhi, sguardo, sopracciglia, testa e postura, poichè ogni segnale in ciascuna modalità viene segmentato ed etichettato su cinque diversi livelli, relativi a: - Descrizione del segnale, che fornirà una descrizione discorsiva o in termini di movimenti o di parametri formazionali (se si analizza un gesto coverbale, vedasi Stockoe 1980; per la descrizione del segnale verbale su base acustica). - Tipologia descrittiva, che classificherà il segnale in base alle categorie individuate (se si analizza, per esempio, un gesto coverbale quale l'indice teso e innalzato che si muove avanti e indietro, si sceglierà tra simbolico, batonico, iconico, ecc.). - Significato, che riporta una parafrasi verbale del significato di quel particolare segnale (es. per il gesto simbolico sopra citato: "Stai/state attento/i" " Fai/fate attenzione!"); con una opportuna sigla di segnalazione (M per "Metaforico") può essere indicato il passaggio da un significato concreto ad uno astratto o metaforico (da "pesare" a "valutare, giudicare"). - Tipologia semantica, che classifica il significato di gesti e movimenti in base ad una tassonomia semantica od ontologica (conoscenze del mondo, conoscenze della mente del parlante, autopresentazione; l'etichettatura del gesto coverbale sopra citato sarà "conoscenza della mente del parlante"). - Funzione, che individua il rapporto semantico fra il segnale considerato e il concomitante segnale verbale. La funzione può essere ripetitiva, quando il segnale analizzato ha lo stesso significato del verbale; aggiuntiva, se vi aggiunge informazioni; sostitutiva, se porta informazioni che l'altro segnale non esplicita; contraddittoria, se ne contraddice il significato; indifferente, se i due segnali fanno parte di piani comunicativi diversi. Oltre a queste valutazioni, che devono essere il più possibile oggettive, cioè aderenti alla struttura del segnale e non alle "ricostruzioni" dello studioso, sarà probabilmente da implementare un livello finale di valutazione, non limitato al solo segmento in analisi, in cui si dovrà utilizzare il contesto per giungere all'individuazione di sovrascopi e significati indiretti. Ad esempio, ad un aggrottamento di sopracciglia fornito dall'interlocutore come segnale di backchannel è possibile attribuire, al di là di un significato letterale di incomprensione ("non capisco"), un significato indiretto di disaccordo ("non sono d'accordo con te"). Nel caso del gesto "mano a borsa", il significato letterale "ma che dici?" trasmette l'affermazione "non è vero quello che dici". In questi casi, a seconda che si stia analizzando il livello letterale o quello indiretto, si dovrebbero forse utilizzare delle etichette diverse per la tipologia semantica e la funzione.Un problema metodologico reale potrebbe porsi quando si dovesse giudicare la funzione non solo dei gesti e dei vari movimenti rispetto al parlato (che nel nostro approccio è stato scelto come segnale di riferimento), ma dei gesti e altri tipi di movimenti tra loro, come nel caso dell'interazione tra espressione facciale e gesto simbolico.

2. LA PARTITURA IMPLEMENTATA IN ANVIL Analisi della comunicazione

multimodale basate sulla Partitura sono state già presentate nel passato: i risultati ottenuti dall'analisi frame by frame, della sequenza della registrazione televisiva, dalla sua visualizzazione contemporanea al segnale vocale grazie a programmi di acquisizione quali Adobe Première, e dall'analisi del segnale acustico, eseguite off line in termini di caratteristiche spettrali, di andamenti di F0 e di intensità, venivano riportati manualmente su grafici (Magno Caldognetto e Poggi 1994, Poggi e Magno Caldognetto 1996). Tali descrizioni potevano generare errori nella valutazione confrontativa di dati elaborati separatamente e a volte con scale temporali diverse a causa della diversità dei dispositivi di analisi. Queste limitazioni metodologiche sono state superate implementando la Partitura su ANVIL (Kipp 2001) per le sue qualità di robustezza, flessibilità e relativa facilità d'uso. E' stato creato innanzi tutto un menu dei sistemi comunicativi che utilizzano i canali uditivo e visivo e per ciascuno di essi sono stati implementati i cinque livelli di analisi. Ove possibile, per esempio nel caso della tipologia del gesto o della sua funzione, sono state elaborate, sulla base di ricerche precedenti (per es. Kendon 1995, McNeill 1992), delle liste di etichette classificatorie inserite in menu a tendina, mentre per gli altri livelli analitici è prevista una descrizione discorsiva o una descrizione simbolica (vedasi Poggi 2001 e per i cheremi Stokoe 1980). Ciò che rende particolarmente utile il sistema ANVIL è la possibilità di memorizzare le segmentazioni relative a ciascun segnale in relazione alla scala temporale scelta dalla visualizzazione della forma d'onda e dello spettrogramma del segnale verbale. Una volta eseguite tutte le valutazioni previste dalla Partitura, è possibile il controllo delle relazioni temporali tra le unità delle varie modalità e la valutazione confrontativa dei significati trasmessi dai vari segnali. Etichettature del segnale restano comunque totalmente affidate allo studioso. A fronte delle potenzialità del sistema descrittivo appena illustrato, vanno ricordati alcuni limiti tecnologici di ANVIL 4.0, soprattutto per quanto riguarda l'analisi acustica, sia a livello segmentale che soprasegmentale. Mentre ANVIL prevede solo la trascrizione ortografica delle parole, è già stato inserito un livello di trascrizione fonetica e sillabica, anche se la segmentazione non è ancora soddisfacente perché viene eseguita sulla base dello spettrogramma importato grazie al plug in Sonogram. Va ricordato che ogni etichettatura e descrizione viene inserita all'interno di caselle definite temporalmente in base ai frame in cui è stato individuato il segnale sottoposto ad analisi. All'interno di ciascuna casella, una volta attivata, è possibile leggere le relative etichette definitorie. Il contenuto delle descrizioni e delle categorizzazioni sono disponibili nella loro completezza e totalità solo nella presentazione multimediale on line: solo in questo caso, infatti grazie al cursore che scorre lungo l'andamento temporale del campione da analizzare, è possibile scoprire come interagiscono, in relazione allo stesso frame le diverse modalità comunicative. Proprio questa importante caratteristica rende particolarmente significativo ricorrere a sistemi informatizzati come ANVIL.

**Giovanna Marotta e Elena Sardelli**
Università di Pisa, Dipartimento di Linguistica

**Parametri prosodici per un modello semi-automatico di riconoscimento del parlante.**

Il riconoscimento automatico del parlante rappresenta da tempo una frontiera importante degli studi sul parlato; le tematiche connesse trovano infatti molteplici applicazioni, dalla sintesi del segnale vocale alla psicoacustica, dalla glottodidattica alla fonetica forense. I tentativi finora realizzati richiedono una serie di informazioni puntuali sulle dinamiche di influenza reciproca tra i diversi parametri acustici, che tuttavia, per loro natura, sembrano essere caratterizzati da valenze olistiche piuttosto che discrete. In questa sede si intendono presentare due tentativi di realizzazione di un modello atto alla gestione ed al riconoscimento dei tratti prosodici coinvolti nella caratterizzazione diatopica del parlante italiano, senza far ricorso alle informazioni provenienti dal livello segmentale. La base empirica del nostro lavoro è costituita dall'analisi prosodica di tre realtà urbane (Roma, Milano e Catanzaro), di cui sono state preliminarmente individuate le marche diatopiche acusticamente rilevanti. Risultano salienti a questo proposito diversi parametri, tra cui gli andamenti melodici delle vocali prominenti e l'escursione frequenziale ad esse associata; la distribuzione delle durate vocaliche e le modalità di ancoraggio degli accenti intonativi con la stringa segmentale. In particolare, il parametro della durata sembra essere responsabile dei rapporti esistenti tra livello segmentale e livello soprasegmentale; infatti, la lunghezza segmentale non solo costituisce la base fisica indispensabile per la piena realizzazione di alcuni movimenti frequenziali, ma influisce anche in maniera incisiva sull'intera organizzazione ritmica dell'enunciato. Per lo sviluppo del primo tentativo di modello semi-automatico di riconoscimento del parlante su base prosodica, abbiamo proceduto innanzitutto alla segmentazione del segnale in unità intonative, classificate secondo la tipologia frasale di appartenenza (essenzialmente, enunciati dichiarativi, interrogativi, sospensivi) in un corpus statisticamente rilevante (dialoghi Map Task e Test delle Differenze, dal Corpus CLIPS). Per ogni tipo di enunciato esaminato, sono stati misurati i seguenti parametri: - escursione melodica dei movimenti distintivi e percettivamente salienti; - range melodico; - valore minimo di F0; - valore finale di F0; - durata della vocale tonica prominente; - durata della vocale tonica nucleare; - durata della vocale atona finale. Si è inoltre proceduto alla trascrizione tonale secondo il sistema ToBI, nonché alla classificazione ed etichettatura dei confini prosodici. I valori medi dei parametri sopra indicati sono stati utilizzati come termine di confronto per enunciati provenienti da altri corpora, al fine di verificare la possibilità di un riconoscimento diatopico automatico del parlante. I test pilota effettuati finora a partire dalle tre varietà italiane suddette sembrano ottenere un buon grado di successo. Nel secondo tentativo di sviluppo di un modello di riconoscimento del parlante su base prosodica, è stata effettuata una parziale etichettatura dei files sonori mediante

un semplice text grid - all'interno del software PRAAT - sul quale sono stati annotati solamente i confini delle porzioni frasali prosodicamente salienti, per ciascun tipo frasale indagato. Per i contesti interrogativi di tipo polare, sono stati manualmente etichettati i confini dell'ultima sillaba tonica e della sillaba atona finale, nonché marcati i margini delle relative vocali. Per i contesti dichiarativi e sospensivi sono stati individuate le sillabe prominenti e le sillabe atone seguenti, nonché le sillabe toniche e atone finali. Grazie alla collaborazione con l'Équipe Prosodie et Représentation Formelle du Langage - Laboratoire Parole et Langage di Aix-en-Provence, è stato quindi realizzato un apposito script per PRAAT, in grado di stabilire la distanza prosodica tra le varietà italiane esaminate. Il programma elaborato procede al calcolo automatico dei seguenti valori: - durata delle sillabe e delle vocali etichettate; - rapporto tra la durata della vocale tonica finale e quella della vocale atona seguente; - valore di FO minimo, massimo, iniziale, finale e medio di ogni porzione selezionata; - escursione melodica espressa in Semitoni dei movimenti tonali all'interno dei confini stabiliti. I primi risultati di questa applicazione hanno fornito risultati incoraggianti per la prosecuzione della ricerca. In particolare, sulla base del solo calcolo del rapporto tra la durata della vocale tonica finale e quella della vocale atona seguente, il programma sembra produrre buoni risultati nel riconoscimento della varietà catanzarese, che risulta marcata da una maggiore distanza prosodica rispetto a Roma e Milano, e dunque più facilmente identificabile. D'altra parte, il parlato dei due più grandi centri urbani italiani sembra invece essere riconosciuto su base prosodica soprattutto in rapporto alle caratteristiche dei movimenti melodici che si realizzano sulla sillaba tonica prominente e sull'atona seguente. Infatti, una consistente escursione melodica in senso discendente a partire dalla vocale prominente verso la vocale atona seguente sembra essere marca prosodica diatopica della varietà romana, mentre un mantenimento del livello tonale raggiunto oppure un eventuale andamento ascendente, connotano in senso forte la varietà milanese. I due sistemi di riconoscimento su base prosodica sono intimamente diversi, in quanto l'uno può dirsi quantitativo-statistico, mentre l'altro è di tipo fondamentalmente, ma non esclusivamente, qualitativo: il modello statistico permette infatti di mettere in relazione i valori dei parametri acustici rilevati per una singola frase con le medie precedentemente calcolate per una data varietà; il modello qualitativo (che, ricordiamo, è basato su uno script per PRAAT) mira invece ad individuare le caratteristiche prosodiche in loco, presso alcuni punti strategici dell'enunciato. Il fine ultimo del nostro lavoro sarebbe quello di elaborare un modello di riconoscimento prosodico integrato, in quanto composto di due componenti distinte per calcolo e verifica: da una parte, il confronto specifico tra i singoli valori di alcuni parametri prosodici e i valori medi statistici provenienti dall'analisi acustica preliminare; dall'altra, il calcolo automatico di alcuni parametri acustici, concentrati sui punti salienti dell'enunciato, in rapporto alla tipologia frasale. A tale scopo, intendiamo nel prossimo futuro confrontare sistematicamente i risultati ottenibili dai due modelli, oltre a testare le loro potenzialità su un campione più ampio di parlato e su un numero maggiore di varietà italiane.

**Stefania Marzo, Serge Vanvolsem**
Katholische Universität, Leuven

## Tecniche statistiche per lo studio variazionale: l'italiano parlato nelle Fiandre (Belgio).

In questo contributo intendiamo presentare due tecniche statistiche che ci hanno permesso di quantificare la variazione nell'italiano parlato in due comunità di Italiani nelle Fiandre. Lo scopo di tale studio è stato di verificare la ripetutamente menzionata "eterogeneità" (Bettoni 1993: 441) nella parlata italiana all'estero, cercando una struttura sia nei tratti linguistici che nella miriade di fattori sociali caratterizzanti gli Italiani all'estero. I due metodi applicati in questo studio sono: l'analisi delle componenti principali (Principle Component Analysis, d'ora in poi ACP) e l'analisi di regressione logistica (d'ora in poi Rlog). Obiettivo della presentazione è di dimostrare da un lato l'applicazione dei due metodi e dei risultati che riportano e dall'altro le possibilità che offrono per ulteriori analisi di corpus, dell'italiano parlato e scritto in generale. Nonostante la proliferazione di studi sull'italiano all'estero a partire dagli anni '80, sono poche le ricerche che si sono occupate a dare una struttura alla variazione dell'italiano parlato all'estero, ovvero a determinare in base a quali fattori si possa definire la varietà come un insieme nettamente più omogeneo. Un tale scopo necessita un approccio empirico e quantitativo, che consente di individuare non solo i meccanismi linguistici e extralinguistici che agiscono sulla lingua italiana all'estero ma anche la struttura alla base della variazione. Per desumere la struttura s'intende capire come agiscono i meccanismi (cioè qual è l'effetto di un determinato fattore sociale sullo sviluppo linguistico di un individuo) e pertanto dedurre il comportamento linguistico in base ad una serie di fattori sociali. L'approccio richiede inoltre l'appoggio di tecniche statistiche atte a rivelare e a comprendere la struttura di fondo. Alla luce di questi principi, il presente contributo vuole presentare i suddetti metodi statistici, l'ACP e la Rlog, che consentono di creare una struttura base affidabile per definire la varietà di italiano parlata all'estero. In termini statistici, si cercherà di "prevedere" il tipo di varietà d'italiano parlato, in base ad una serie di variabili sociali. Il campione di riferimento è composto da un gruppo di 48 Italiani nati e cresciuti nelle Fiandre (Belgio); da un punto di vista sociolinguistico si tratta di un campione controllato, proporzionalmente distribuito secondo quartiere di residenza (due quartieri, Lindeman e Zwartberg), sesso (donna–uomo), generazione (second-terza) e età (due gruppi d'età). Per la selezione e la descrizione dei fenomeni linguistici ci si è basati su un corpus di parlato semi-spontaneo, contenente 148.726 parole grafiche. L'analisi linguistica ha previsto una fase di estrazione ed etichettatura delle ricorrenze dei fenomeni linguistici, effettuate con il programma Abundantia Verborum (Speelman 1997), un software linguistico ideato alla K.U.Leuven per automatizzare analisi lessicologiche, morfosintattiche e fonologiche. Tale analisi ci ha consentito di individuare 30 variabili linguistiche, attinenti ai tre influssi

caratterizzanti l'italiano parlato all'estero: variabili regionali, semplificanti e contattuali. Le variabili sociali selezionate sono 12. Per illustrare l'applicazione delle due tecniche statistiche, in questo contributo ci si concentrerà essenzialmente sulla variabile "quartiere" (in cui risiedono gli informatori: Lindeman o Zwarberg), definita come "la rete sociale" e resa operativa secondo i criteri della "social network theory" di Milroy (1987): Lindeman rappresenta la rete sociale "close-knit", Zwartberg quella "loose-knit". Le analisi statistiche sono state effettuate con il software R (2003), versione 2.0.1.(1) Le due tecniche statistiche corrispondono in effetti a due fasi fondamentali nella strutturazione della variazione nel corpus: 1) In una prima fase è stata applicata l'ACP, una tecnica statistica descrittiva ed esplorativa frequentemente impiegata nella prima fase di elaborazione dei dati. La tecnica consente di identificare e di estrarre dal totale di variabili (linguistiche e sociali) una struttura con le variabili più importanti, a volte rimaste latenti. Tecnicamente lo scopo dell'ACP consiste nell'individuare in un gruppo di variabili tra loro correlate x1, x2, ... xq, poche nuove variabili sintetiche - dette componenti principali - non correlate y1, y2, ... yq, di cui ognuna costituisce una combinazione lineare delle variabili originali x, offrendone inoltre una sintesi il più fedele possibile(2). Applicata al presente contesto, l'ACP ha permesso di rivelare la struttura base della variazione nel corpus mediante appena due variabili, senza tuttavia ridurre l'informazione presente nel set di variabili originali. Queste due componenti principali spiegano una parte consistente della variazione e possono essere utilizzate nelle seguenti analisi statistiche (di regressione): le variabili sono "presenza di varianti regionali" e "presenza di varianti semplificanti". Il vantaggio di queste componenti consiste non solo nell'offrire una struttura base della variazione presente nel corpus, ma anche – sul lato analitico e pratico – nel ridurre un set di 30 variabili ad appena 2 variabili. Tale riduzione facilita l'input e rende più stabile l'analisi di regressione(3). 2) La Rlog consente di quantificare la struttura identificata nell'ACP e di stimare in che misura le due componenti principali permettono di prevedere la rete sociale (Lindeman o Zwartberg) alla quale appartengono gli informanti. In quest'analisi la variabile "quartiere" verrà inserita come "variabile risposta" (o dipendente) e le due componenti principali come "variabili predittive" (o indipendenti). Queste due tecniche ci hanno consentito di stimare e di prevedere il tipo di rete sociale in base alle due componenti principali, che riassumono due importanti caratteristiche della parlata italiana all'estero (regionali, semplificanti). L'analisi ha rivelato come la comunità a forte coesione interna tende a conservare la varietà parlata regionale e come, invece, quella più aperta, tende a sviluppare una parlata migratoria più tradizionale, vale a dire, caratterizzata innanzitutto da fenomeni di semplificazione. L'interesse dell'approccio presentato risiede non solo nei risultutati che sono stati ottenuti per questa ricerca e per l'italiano all'estero in generale, ma anche e soprattutto per le possibilità che offre nell'ambito della ricerca variazionale. L'ACP consente infatti di capire e di interpretare un ampio corpus di lingua, alle quali sono correlati numerosi fattori sociali, sintetizzando il tutto in poche componenti facilmente interpretabili. Le

componenti principali sono inoltre direttamenti utilizzabili come variabli predittive nelle regressioni logistiche.

Riferimenti bibliografici:

Bettoni, C. (1993), "L'Italiano Fuori D'Italia." Introduzione All'Italiano Contemporaneo. La Variazione e Gli Usi. A. A. Sobrero (ed.). Bari: Laterza.
Maggino, F. (2005), L'analisi dei dati nell'indagine statistica. Vol. 2. L'esplorazione dei dati e la validazione dei risultati. Firenze University Press.
Milroy, L. (1987), Language and Social Networks. New York: Basil Blackwell.
Speelman, D. (1997), Abundantia Verborum. A computer tool for carrying out corpus-based linguistic case studies. K.U.Leuven.
http://wwwling.arts.kuleuven.ac.be/genling/abundant/

**Alessandro Mazzei, Vincenzo Lombardo**
Università di Torino, Dipartimento di Informatica

**Toward a dynamic constituency model of syntax.**

Incrementality is a basic feature of the human language processor. There is a considerable amount of objective data that humans build-up interpretations of the sentences before perceiving the end of the sentence [Marslen-Wilson, 1973, Kamide et al., 2003], and there is also evidence to support the idea that such incremental interpretation is based on fast, incremental construction of syntactic relations[1] (see, for example Kamide et al. [2003], Sturt and Lombardo [2005]) and which components of a syntactic parser account for incrementality. Steedman proposed a general anatomy to classify a syntactic parser [Steedman, 2000]. In this schema we can distinguish three main components: 1. a grammar, 2. a parsing algorithm and 3. an oracle. The grammar is a static object, usually defined in a declarative form, which contains competence-related information about the grammaticality of the sentences. The parsing algorithm can be divided into two parts: a parsing strategy specifying the order of application of the grammar rules (e.g. left-to-right, bottom-up, top-down), and a memory organization strategy, which speci es how the parser memorizes the partial results (e.g. depth-first, back-tracking, dynamic programming). The oracle takes into account the ambiguity of syntactic analysis and ranks the possible alternatives for the parsing algorithm through a rule-based or probabilistic strategy. In most incremental parsing models, the competence grammar is usually a standard generative formalism, e.g. context-free grammar [Roark, 2001] or a categorial formalism, e.g. Combinatory Categorial Grammar [Steedman, 2000]. As a consequence, the parsing strategy is the component that totally specifies the incremental nature of the syntactic process (i.e. it is not part of the competence knowledge). However there exist a few approaches that incorporate the parsing strategy, i.e. the order of application

of the rules of the grammar, into the grammar component. Phillips proposed a model of grammar in which the parsing strategy is part of the competence grammar [Phillips, 1996]. His fundamental hypothesis (PiG, parser is grammar) is that a grammar is a formal generative device, that specifies what are the grammatical structures (i.e. the legal constructions) in terms of their computation. He has pointed out that in this model the notion of constituency is variable during the analysis of the sentence: a fragment of the sentence that is a constituent at some point in the derivation, may not be a constituent after the processing of a subsequent word [Phillips, 2003]. Phillips' fundamental hypothesis ties the notion of constituency to the parser computation. Constituents have a variable status during the analysis of the sentence and depend on the parsing state. The only difference between a parser and grammar is that the former has to take into account the finiteness of the resources, while the latter does not. An example of a categorial grammar oriented approach that follows PiG hypothesis is proposed by Milward in the context of dynamic grammars, in which the syntactic analysis is viewed as a dynamic process [Milward, 1994]. In this model the syntactic process is a sequence of transitions between adjacent syntactic states S(i) and S(i+1), while moving from left to right in the input string. The syntactic state contains all the syntactic information about the fragment already processed. Similarly to Phillips' model, the parsing strategy is a part of the dynamic competence grammar. Both Phillips and Milward can give a competence account for non-constituency coordination: two fragments x and y of the sentence are grammatical conjuncts if they continue in the same way from the same starting state. Assuming that the competence grammar specifies the parsing strategy, there are several strategies that model the incremental nature of the syntactic process. It has been argued that the most parsimonious account of psycholinguistic experimental data is to assume that each new word is connected to the syntactic structure as soon as the word is perceived [Kamide et al., 2003, Sturt and Lombardo, 2005]. A formalized version of this property of the syntactic processor, that we call strong connectivity, has been proposed by Stabler: people typically incorporate each (overt) word into a single, totally connected syntactic structure before any following words; incremental interpretation is achieved through the interpretation of this single syntactic structure [Stabler, 1994]. In this paper we present a grammatical formalism such that the parsing strategy is totally determined by the competence grammar and that fulfills the strong connectivity hypothesis for simplifying the syntax-semantic interface[2]. In particularly it is a dynamic version of Tree Adjoining Grammars (TAG, Joshi and Schabes [1997]), called DVTAG. TAG is a family of generative formalism that uses tree elementary structures rather than string elementary structures, as in context-free grammars. Moreover, TAG uses two distinct rewriting operations. A substitution operation, that is the attachment of an elementary trees on a node of the fringe of another elementary tree. An adjoining operation, that grasp an elementary tree into an internal node of an another elementary tree. This latter operation increases the generative power of TAG with respect to context-free grammars. Indeed DVTAG

generates the class of the mildly context-sensitive languages, that is supposed to be the class where the human languages fall in the Chomsky hierarchy [Vijay-Shanker et al., 1990].

*[1] We stress our discussion about parsing. Similar arguments have been made for the generation process [2] It should be pointed out that strong incrementality at the syntactic level is not necessarily required for word-by-word incremental interpretation, as pointed out by [Shieber and Johnson, 1993]. However, the alternative requires the semantic interpreter to have access to the parser internal state.*

**Simonetta Montemagni**
Istituto di Linguistica Computazionale - CNR, Pisa

**Aree fonetiche e lessicali toscane a confronto: prime elaborazioni computazionali dei dati dell'Atlante Lessicale Toscano.**

In campo dialettologico e geolinguistico l'uso di strumenti e metodi informatici può essere ricondotto ai seguenti indirizzi: 1) costruzione di risorse dialettologiche quali dizionari dialettali o atlanti linguistici; 2) rappresentazione cartografica dei dati raccolti o dei risultati delle analisi effettuate; 3) comparazione e classificazione dei materiali dialettali raccolti. In 1) e 2) il computer gioca un ruolo conservativo, ovvero di rafforzamento degli apparati metodologici tradizionali della disciplina. Diverso è il caso 3) in cui lo strumento informatico orienta e ridefinisce le possibilità di elaborazione di cui i dati si rendono suscettibili con un impatto innovativo sul versante più strettamente metodologico. In questo caso il computer si presenta come strumento per l'esplorazione dei dati, in particolare per l'individuazione, all'interno della massa dei materiali raccolti, di correlazioni sulla base delle quali formulare generalizzazioni relative alla variazione diatopica non immediatamente percepibili all'osservazione diretta dei singoli dati. Il passaggio dalla descrizione della distribuzione geografica di specifici tratti linguistici (fonetici, morfologici, lessicali) a un livello di descrizione più astratto volto a identificare confini e continua dialettali è oggi reso possibile dall'uso combinato di tecnologie linguistico-computazionali con tecniche di analisi statistica multivariata. L'uso di tecniche statistiche per il calcolo della distanza linguistica proposto agli inizi degli anni '70 da Seguy (1971) e ulteriormente raffinato da Goebl (1984) ha costituito un passo fondamentale in questa direzione. Tuttavia, questo tipo di misure si basavano esclusivamente su distinzioni categoriali. Kessler (1995) segna una svolta in questo filone di studi, applicando misure di similarità tra stringhe di caratteri ( "Levenshtein distance") direttamente sul dato dialettale in trascrizione fonetica. Nerbonne et al. (1999), Heeringa (2004) hanno esteso l'uso dell'algoritmo della Levenshtein distance a diversi tipi di rappresentazioni linguistiche (ad esempio, rappresentazioni a tratti). Ad oggi, l'uso combinato di questi due tipi di tecniche sembra particolarmente promettente nello studio

della variazione linguistica di diverse lingue (olandese, inglese, irlandese, sardo, norvegese, bulgaro). Oggetto della presente comunicazione è l'esplorazione dei materiali dialettali dell'Atlante Lessicale Toscano (ALT) disponibili nel sito di ALT-Web (http://serverdbt.ilc.cnr.it/altweb) per la ricostruzione di visioni "sintetiche" della variazione linguistica in Toscana. A tal fine è stato usato il software Rug/L04 per analisi dialettometriche e cartografazione dei risultati messo a disposizione nel sito http://www.let.rug.nl/~kleiweg/L04/. In particolare, il presente studio si è focalizzato sull'identificazione di aree fonetiche e lessicali toscane e sull'esplorazione dei tipi di correlazioni esistenti tra i due tipi di aree dialettali identificate. La definizione delle aree fonetiche e lessicali toscane si è basata sui diversi tipi di rappresentazione presenti in ALT-Web dove, ad ogni attestazione dialettale sono associati diversi tipi di codifica: 1) rappresentazione in trascrizione fonetica per la quale è stato adottato uno schema di codifica complesso che affianca rappresentazioni composizionali e rappresentazioni atomiche; 2) rappresentazioni in ortografia italiana suddivise in: 2a) traslitterazione in ortografia italiana dell'attestazione dialettale; 2b.) normalizzazione di primo livello, che neutralizza tratti specifici della realizzazione fonetica del dato (tipicamente, variazioni fonetiche produttive in Toscana) senza fare astrazione da variazioni morfologiche. Le rappresentazioni di livello 1 sono alla base della definizione delle aree fonetiche toscane, mentre per le aree lessicali è stato preso come punto di partenza il livello di rappresentazione 2b. Per la definizione delle aree fonetiche toscane sono stati selezionati 592 tipi lessicali: tale selezione è stata guidata da criteri di varietà di realizzazione e di frequenza di attestazione. La distanza fonetica tra le diverse varietà linguistiche locali è stata calcolata sulle rappresentazioni fonetiche atomiche con la misura di Levenshtein. Le aree fonetiche risultanti, identificate attraverso tecniche di clustering applicate sulla matrice delle distanze fonetiche, si presentano come aree concentriche che vedono un nucleo centrale che dall'area fiorentina si spinge: verso sud all'area senese; verso ovest fino al confine lucchese e a all'area pisano-livornese (con sporadici sbocchi costieri). Attorno a questo nucleo centrale si osserva un'area di contorno all'interno della quale si distinguono nettamente i dialetti romagnoli, lunigianesi e l'aretino. Per la definizione delle aree lessicali sono state selezionate le risposte normalizzate a 165 domande onomasiologiche che presentavano un ambito di variabilità compreso tra 5 e 50. In considerazione del fatto che questo livello di normalizzazione non fa astrazione da variazioni morfologiche si è ritenuto opportuno ricorrere alla misura di Levenshtein anche in questo caso. La morfologia delle aree lessicali risultanti è alquanto diversa da quella delle aree fonetiche descritta sopra. Si distinguono chiaramente i dialetti non toscani (romagnolo e lunigianese), e quelli toscani ripartiti in fiorentino, pistoiese, lucchese, pisano-livornese, senese, aretino e elbano-grossetano. A conclusione di queste prime esplorazioni computazionali nell'ALT è interessante notare come la distribuzione delle aree lessicali che emerge coincida con quanto descritto in Giacomelli (1975) in relazione al lessico, e in buona parte anche con la classificazione dei dialetti toscani di Giannelli (2000) basata su diversità

morfosintattiche, fonetiche, fonologiche e lessicali. Quest'ultima classificazione, però, non sembra trovare conferma al livello delle aree fonetiche rilevate. Infatti, a un'analisi attenta emerge una discrepanza tra le aree fonetiche e lessicali identificate. Ciò non meraviglia se si considera quanto sostengono Chambers e Trudgill (1998:97) secondo i quali aree lessicali e aree fonetiche non necessariamente coincidono, in quanto le prime sono più soggette a influenze esterne delle seconde. Alla luce di questi risultati la correlazione tra la variazione fonetica e lessicale in Toscana appare un fertile terreno di ricerca dove l'apporto di tecniche informatiche è essenziale: sono in corso nuove analisi per misurare la correlazione tra gli schemi di variazione fonetica e di variazione lessicale in Toscana sulla base delle attestazioni dell'ALT.

Riferimenti bibliografici:

Chambers, J., Trudgill, P. 1998. Dialectology. Cambridge University Press.
Goebl, H. 1984. Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. Max Niemeyer, Tübingen.
Giacomelli, G. 1975. Aree lessicali toscane, «La ricerca dialettale», I, pp. 115-152.
Giannelli, L. 2000. Toscana. Pacini Editore, Pisa.
Heeringa, W. 2004. Computational Comparison and Classification of Dialects. Ph.D. thesis, University of Gröningen.
Kessler, B. 1995. Computational Dialectology in Irish Gaelic. In Proceedings of EACL, pp. 60-67.
Nerbonne, J., Heeringa, W., Kleiweg, P. 1999. Edit distance and dialect proximity. In D. Sankoff, J. Kruskal Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, CSLI, Stanford, CA.
Séguy, J. 1971. La relation entre la distance spatiale et la distance lexicale. «Revue de LinguistiqueRomane», 35:335–357.

**Monica Mosca**
Università di Pisa, Università UPO Vercelli - Laboratorio Linguistica Computazionale e applicata

**Le espressioni spaziali in dialoghi task-oriented: un approccio cognitivo basato su corpora di parlato italiano.**

Lo spazio e le espressioni spaziali sono di primaria importanza nello studio delle lingue storico-naturali per molti aspetti. In particolare lo studio della deissi, in una prospettiva tipologica, richiede in ogni caso la formulazione di un modello dello spazio e dei suoi punti di riferimento; la linguistica cognitiva ha affrontato in maniera diversa i problemi di percezione dello spazio che vengono evidenziati

in diversi tipi di espressioni linguistiche in relazione alle concettualizzazioni sottostanti. La deissi spaziale si esprime attraverso elementi linguistici che stabiliscono un collegamento tra il parlante e lo spazio in cui pronuncia il suo enunciato. A fianco alle descrizioni grammaticali del comportamento degli elementi deittici, molti studi hanno tentato di focalizzare il fenomeno in una prospettiva tipologica, nel tentativo di identificare un insieme di tratti adatti a classificare in maniera uniforme il più alto numero di lingue. Nella descrizione delle lingue europee i tratti di prossimità al parlante o all'ascoltatore si intersecano con l'opposizione person oriented /distance oriented (cfr. Benedetti & Ricca, 2002; Da Milano 2005). Questi tratti oltre a essere proprietà che classificano un sistema linguistico rinviano probabilmente a un modello dello spazio in cui parlante e ascoltatore si muovono. L'identificazione classica del centro deittico è quella di "hier, jetzt und ich " di Buehler (1934), non sempre di facile applicazione nella descrizione linguistica. La nozione di origo è utilizzata anche per caratterizzare i principali verbi di movimento deittico tra itivi e ventivi (cfr. Ricca 1993). Recentemente Jungbluth (2003) ha proposto e individuato la nozione di "diade conversazionale" applicata ai pronomi dimostrativi dello spagnolo in cui l'evento comunicativo ha senso con la complementarietà di parlante e ascoltatore. Queste proposte teoriche, a mio avviso, introducono una prospettiva cognitiva, anche se non esplicitamente, in quanto la nozione di centro deittico è in ogni caso legata a concettualizzazioni dello spazio di parlante e ascoltatore . Gli studi di tipo cognitivo non dedicano interesse specifico alla deissi, ma piuttosto alla generalità delle espressioni spaziali, ricercando apertamente un modello dello spazio, sia esso in una prospettiva di iconicità o di relativismo linguistico (Slobin 2005). Ad esempio in Levinson (1996, 2003) si afferma che la diversità di spatial frame of reference è una diversità di sistemi di coordinate e non di oggetti che ne fanno uso. Non esiste un solo "quadro di riferimento" ma diversi tipi di frame of reference che alla fine si riducono a tre, intrinseco, assoluto e relativo. Studiosi come Talmy (1985, 2000) e Slobin (2003, 2004) analizzano le frasi che esprimono eventi di movimento in termini di componenti semantici (MOTION, PATH, MANNER, FIGURE, GROUND). Il diverso modo di lessicalizzare questi elementi dà luogo ad una tipologia linguistica che distingue lingue verb-framed, che esprimono nel verbo principale MOTION+PATH (es. lingue romanze), e lingue satellite-framed, che distribuiscono l'informazione tra il verbo principale ed i suoi satelliti. Lo studio della lessicalizzazione e della distribuzione sintattica dei componenti semantici delle espressioni spaziali porta Slobin (2005) ad approfondire la caratterizzazione tipologica di alcune famiglie linguistiche e a sostenere una nuova interpretazione della nozione di iconicità. Questo lavoro propone un'indagine sulle espressioni spaziali in italiano parlato a partire da corpora elicitati (una ventina di dialoghi la cui durata media è di 7 minuti circa) raccolti con metodologia Map Task, per la cui descrizione rinvio a Anderson et alii (1991). Questi tipi di dialogo si caratterizzano per una relativa spontaneità ed una negoziazione ricca di riferimenti spaziali in quanto il compito in cui sono coinvolti parlante e ascoltatore è di orientamento su una mappa. Concordo

quindi sia con la definizione di Lyons (1977) quando parla della condizione in cui parlante e ascoltatore sono face to face (contesto appropriato per la grammaticalizzazione e la lessicalizzazione della deissi e delle espressioni spaziali), sia con il pensiero di Halliday (1990) che definisce la "concezione normale" del dialogo il processo di transazione e rielaborazione che porta a un accordo tra le parti. Il corpus è stato trascritto e successivamente sottoposto alle operazioni di tokenizzazione e lemmatizzazione. Ho scelto la via dell'implementazione di un tokenizzatore specifico che segmenti la variabilità linguistica del testo e gli standard di trascrizione AVIP. Per il riconoscimento degli elementi da tokenizzare si sono utilizzati pattern di espressioni regolari. Segue il processo di lemmatizzazione, fase che risulta utile nel collegamento tra espressioni spaziali e categorie grammaticali, per poi sottoporre il tutto ad analisi statistico-distributiva. L'obbiettivo consiste nell'identificare le espressioni spaziali, normalizzate in forma di espressioni regolari. L' operazione porta all'identificazione delle principali concettualizzazioni, in termini di cluster. Questo è il requisito minimo per condurre indagini statistiche non solo descrittive, ma predittive relativamente alla correlazione tra componenti di significato "sintetizzati", per esempio, nei verbi di moto e la loro proiezione sulla sintassi della frase (cfr. Jackendoff 1983, 1990).

**Malvina Nissim, Johan Bos**
ISTC-CNR - Laboratory for Applied Ontology, Roma - Università "La Sapienza", Roma

**Using the Web as a corpus in Natural Language Processing.**

1 Introduction Research in Natural Language Processing (NLP) has in recent years benefited from the enor- mous amount of raw linguistic data available on the World Wide Web. The presence of standard search engines have made this data accessible to computational linguists as a corpus of a size that had never existed before. Although the amount of readily available data sounds attractive, the Web as a corpus has the inherent disadvantage of being unstructured, uncontrollable, unprocessed, constantly changing, and of heterogeneous quality. These conflicting aspects force NLP researchers to develop new and creative techniques to exploit it as a lin- guistic resource. In this paper we illustrate how the Web can support NLP tasks, focussing on two of them: acquiring semantic knowledge, and validating linguistic hypotheses. As the largest corpus available, the Web can be used as a source for extracting lexical knowledge by means of specifically task-tailored syntactic patterns. One general example is the extraction of hyponymic relations by means of ISA-like patterns. Example applications that benefit from this are named entity recognition, ontology building/editing, and anaphora resolution. The Web can also be used as a testbed for linguistic hypothesis and/or evaluation of system outputs. For example, the frequency of different PP-attachments produced by a parser can be compared, or the frequency of different

spelling variants. To access and retrieve information from the Web, NLP researchers often use off-the-shelf interfaces provided by ma jor search engines such as Google, Yahoo, and AltaVista. We will illustrate the general approach by giving two case studies based on our own recent work: one describing how to use the Web for finding antecedents of anaphoric expressions (Section 2), and one that uses the Web in an answer re-ranking task within a question answering system (Section 3). In both of these studies we used the Google API to retrieve information from the Web. 2 Candidate Antecedent Selection for Anaphora Resolution Other-anaphora is an example of lexical anaphora, where the modifiers other or another provide a set-complement to an entity already evoked in the discourse model. In Example (1), the NP "other, far-reaching repercussions" refers to a set of repercussions excluding increasing costs, and can be paraphrased as "other (far-reaching) repercussions than (increasing) costs". (1) In addition to increasing costs as a result of greater financial exposure for members, these measures could have other, far-reaching repercussions. (Wall Street Journal) For interpreting other, far-reaching repercussions as "repercussions other than costs" a resolution system needs the knowledge that costs are or can be seen as repercussions. Most systems that handle this and similar anaphoric phenomena rely on handcrafted resources of lexico-semantic knowledge, such as the WordNet lexical hierarchy [Fel98]. The alternative method that has been suggested by [MN05] is to use the Web. The implicit hyponymy relation between an anaphor (Y) and its antecedent (X) is made explicit via a specific pattern, such as Y(s) and other Xs, which indeed indicates that Y is a hyponym of X. All base noun phrases occurring in an N-sentence window given the anaphor are tested in the pattern, and all resulting phrases are submitted as quoted queries to Google. (In the example queries below "OR" is the boolean operator and has scope on the previous term only). The most frequent phrase is chosen as the one containing the correct antecedent. For Example 1, the available noun phrases (considering just a one-sentence window for the sake of clarity) are "increasing costs", "greater financial exposure", "members", "these measures". Discarding modification and determination, the following phrases are created and searched on Google: cost OR costs and other repercussions (30) exposure OR exposures and other repercussions (0) member OR members and other repercussions (5) measure OR measures and other repercussions (0) The figures in brackets are the number of hits returned by Google for each phrase. In this case, "costs" is correctly selected as the antecedent for "repercussions". [MN05] show that on this task, as well as on a full NP coreference resolution task, this method works even better than using hyponymy relations in WordNet. 3 Answer Reranking in Question Answering One of the problems in automatic Question Answering is to choose the most likely candidate of a set of possible answers. The open domain QED system [ABC+05] outputs a set of plausible answers, ranked using a simple scoring method. However, in many cases this ranking can be considerably improved. An example is the question from the TREC-2005 evaluation campaign "Where was Woody Guthrie born?", for which QED initially produced the following ranked answer

candidates from the Acquaint newswire corpus: 1. Britain 2. Okemah, Okla. 3. Newport 4. Oklahoma 5. New York Although in this N-best list answers 2 and 4 are correct, in the TREC evaluation systems must return one answer only, and as a consequence, this would lead to an incorrect response (Britain) to the question. To deal with this problem, we experimented with additional external knowledge obtained from the Web in order to generate a possibly more accurate ranking. This technique is also known as "answer validation" or "sanity checking", and can be seen as a tie-breaker between the top-N answers. In more detail, this method works as follows. For each of the N-best answer candidates, we take the semantic representation of the question (see [ABC+05]) and for each answer candidate. From this we generate a set of declarative sentences (covering all morphologi- cal variations). The generated sentences are submitted as strict (within quotes) queries for Google. Any information from the given question topic which is not included in the gener- ated sentence (for this example "Woody") is added as a query term in order to constrain the search space. The queries and number of hits returned for each of the queries (in brackets) are shown below. 1. Woody "Guthrie born in Britain" (0) Woody "Guthrie are OR is OR was OR were born in Britain" (0) 2. Woody "Guthrie born in Okemah, Okla." (1) Woody "Guthrie are OR is OR was OR were born in Okemah, Okla." (10) 3. Woody "Guthrie born in Newport" (0) Woody "Guthrie are OR is OR was OR were born in Newport" (0) 4. Woody "Guthrie born in Oklahoma" (7) Woody "Guthrie are OR is OR was OR were born in Oklahoma" (42) 5. Woody "Guthrie born in New York" (0) Woody "Guthrie are OR is OR was OR were born in New York" (2) The returned Google-counts are used as the deciding factors to rerank the N-best answers. Note that we generate several queries for each answer candidate and we sum the returned hits. In this example, the answers would be reranked as follows: 1. Oklahoma (7 + 42) 2. Okemah, Okla. (1 + 10) 3. New York (0 + 2) 4. Britain (0 + 0) 5. Newport (0 + 0) For this example, reranking correctly promoted "Oklahoma" to best answer. 4 Discussion Although the two examples of applications illustrate successful usages of the Web for NLP tasks, yet there are several obstacles. There are only a limited number of search engines available (Google, Yahoo, and Altavista), so the NLP researchers who want to exploit the web as a corpus are dependent on them. Some search engines offer more expressive power in queries ("near"-operator, AltaVista), others a larger range of documents (Google). Searching on exact strings and punctuation is restricted, and also the number of queries is limited to a couple of thousand a day. Further, it will be important to distinguish between useful webpages and those that are not trustworthy (from a linguistic as well as content point of view) – of course not everything that is written is correct. Finally, it is virtually impossible to reproduce exact experiments, since the Web is constantly changing.

References:

[ABC+05] K. Ahn, J.. Bos, J.R. Curran, D. Kor, M. Nissim, and B. Webber. Question Answering with QED TREC-2005. In Voorhees and Buckland, editors,

The Four- teenth Text REtrieval Conference, TREC 2005, Gaithersburg, MD, 2005.
[Fel98] Christiane Fellbaum, editor. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Mass., 1998.
[MN05] K. Markert and M. Nissim. Comparing knowledge sources for nominal anaphora resolution. Computational Linguistics, 31(3), 2005.

**Marco Aldo Piccolino Boniforti, Rodolfo Delmonte / Max Hadersbeck**
Università "Ca' Foscari" Venezia - Laboratorio di Linguistica Computazionale, Universität München - Centrum für Informations- und Sprachverarbeitung

**The Symbolic/Statistical Dichotomy: A New Evaluation of Parsing Systems.**

Natural Language Processing relies notoriously on two quite different approaches: symbolic and statistical. Roughly speaking, the former is carried on by (computational) linguists and the latter by engineers and computer scientists. Statistical approaches largerly benefit from the recent availability of quite huge annotated corpora. They are generally preferred in many NLP applications since the planning of statistical-based systems is not so time-consuming as is the case with symbolic approaches. Nonetheless, their intrinsic limits are also evident. Regarding syntactic parsing, the success rate of such probabilistic systems is strongly conditioned a.o. by the composition of the training corpus (size and genres) and by the smoothing techniques and values more or less arbitrarily adopted for language modeling. In this paper we present a new evaluation of three well performing parsing systems - some of which freely available - for the English language: the Link Grammar Parser by Sleator and Temperley; the GETARUNS parser by Delmonte; the Stanford Parser by Klein and Manning. Two of them, although very different in the kind of approach adopted, are mainly based on hand-written rules and can be thus classified as symbolic, whereas the other one is a probabilistic parser. For system evaluation we run the three parsers on the same group of sentences (about 500) of unannotated English text. Such texts are taken from well established and widely used language corpora. A functional, dependency-based manual annotation of the sentences is then used to evaluate the output of the three parsers. We chose this representation since it can be quite easily obtained from different types of syntactic structures and is also able to make the necessary generalizations to allow for comparison. For this purpose we also developed some specific output conversion tools. Grammatical relations (like subject, object, xcomp) are manually checked, local and global scores of precision/recall are computed and results are discussed. With this work we want to test the pros and cons of both approaches, with the aim to find perhaps an integration of linguistic rules and frequency data, on the wave of what still seems to be a desideratum of many researchers working in the field. We try to give an answer to the following central questions: -Do some symbolic systems perform better than statistical ones? -Is linguistic knowledge really

necessary to achieve better results in terms of success rate? -If so, where and how does it seem to play a central role? -How could we integrate the benefits of both approaches?

**Fabiana Rosi**
Università di Pavia, Dipartimento di Linguistica Teorica e Applicata

**Le reti neurali come simulazione del processo di apprendimento della seconda lingua: il caso della morfologia tempo-aspettuale.**

Il contributo si propone di offrire dati innovativi al dibattito teorico sull'apprendimento delle categorie tempo-aspettuali tramite il confronto fra il percorso acquisizionale di apprendenti umani di Italiano L2 e il percorso acquisizionale di modelli simulativi computazionali, le Self-Organizing Maps (Kohonen 2001), un particolare tipo di rete neurale non supervisionata. Il confronto mira ad approfondire la conoscenza sul ruolo dei fattori di frequenza e di salienza degli elementi nell'input per mettere in luce i meccanismi cognitivi che guidano l'apprendimento delle strutture linguistiche nella seconda lingua. Gli studi sull'apprendimento delle categorie tempo-aspettuali sia in italiano L1 (Antinucci & Miller 1976) sia in italiano L2 (Banfi & Bernini 2003) hanno evidenziato che le forme morfologiche tempo-aspettuali sono, inizialmente, associate a specifiche classi semantiche di predicati (Bertinetto 1986) e la diffusione delle forme segue un percorso governato da scale implicazionali che procedono dall'associazione non marcata alla più marcata secondo l'incremento della ricchezza e complessità dell'input che gli apprendenti ricevono. All'interno del paradigma emergentista, recenti studi (Li & Shirai 2000) hanno interpretato l'evoluzione da una fase acquisizionale alla successiva come il risultato dell'analisi da parte degli apprendenti del comportamento distribuzionale delle forme morfologiche rispetto ai valori semantici dei predicati nell'input della lingua target. Secondo tale approccio, rivolto prevalentemente all'acquisizione di L1, gli apprendenti estraggono dall'input le frequenze statistiche delle combinazioni fra forme tempo-aspettuali e valori semantico-azionali e tendono a produrre le associazioni più frequenti, finché l'esposizione prolungata all'input e l'incremento qualitativo e quantitativo dei dati della lingua target riducono la differenza statistica fra le combinazioni più frequenti e le più rare. L'apprendente, quindi, inizia ad estendere la morfologia aspettuale a predicati appartenenti alle differenti classi semantiche. Lo studio di Li & Shirai (2000) ha verificato tale ipotesi, che rielabora la salienza cognitiva in termini di frequenza statistica delle strutture da apprendere, per mezzo dello strumento simulativo delle reti neurali auto-organizzative (o Self-Organizing Maps: SOMs). Le SOMs (Kohonen 2001) sono dispositivi computazionali di ispirazione neurale che classificano dati di input non supervisionati traducendo relazioni di similarità in relazioni topologiche di prossimità. Una mappa è una rete (generalmente bidimensionale) di "nodi recettori" che si attivano ogni qual volta viene

somministrato loro un dato. Attraverso un'esposizione incrementale a un numero crescente di dati, i recettori si organizzano topologicamente sulla rete in modo tale che recettori contigui tendono a riconoscere classi omogenee di dati. Durante l'apprendimento, con la crescita qualitativa e quantitativa dell'input fornito alla SOM, la forza delle associazioni più frequenti e prototipiche fra morfologia tempo-aspettuale e valore semantico-azionale tende a "rilassarsi" gradualmente ed affianco alle associazioni iniziali vengono apprese le altre associazioni possibili. Le SOMs si dimostrano uno strumento efficace della simulazione dell'apprendimento del linguaggio in quanto si ispirano al funzionamento neuronale del cervello umano e sono, dunque, biologicamente plausibili. Inoltre, hanno il vantaggio di poter simulare processi di apprendimento basati sulle rappresentazioni prototipiche che emergono dall'analisi della frequenza statistica delle co-occorrenze dei fenomeni nell'input linguistico e di descrivere i meccanismi dinamici di acquisizione piuttosto che la sola rappresentazione statica, dal momento che le rappresentazioni prototipiche variano durante il corso dell'apprendimento. Per la nostra ricerca si sono costituiti tre corpora da confrontare fra loro. Il primo corpus è composto dalle narrazioni scritte e orali di tre sequenze del film "Modern Times (Charlie Chaplin 1936), prodotte da 24 studenti universitari di madre lingua italiana. Tali dati rappresentano, da un lato, il campione di controllo e di confronto rispetto alle narrazioni delle stesse sequenze prodotte dagli apprendenti, dall'altro, con l'aggiunta di altri corpora di italiano, il corpus di addestramento delle SOMs. Il secondo corpus è costituito dalle interlingue prodotte da 24 apprendenti di Italiano L2, 12 ispanofoni e 12 tedescofoni studenti universitari Socrates presso l'Università di Pisa, intervistati tre volte a distanza di due mesi, per esaminare tre fasi consecutive del loro percorso acquisizionale. Agli apprendenti è stato richiesto di produrre narrazioni scritte e orali in Italiano L2 e orali nella L1 delle stesse sequenze narrate dagli italofoni nativi. Il terzo corpus comprende i risultati delle diverse fasi dell'apprendimento dello strumento simulativo, che viene addestrato tramite una serie di training set di input della lingua target, forniti alle SOMs in modo incrementale. La simulazione del processo di acquisizione è resa possibile dalla rappresentazione della morfologia tempo-aspettuale e della semantica verbale all'interno delle SOMs. La rappresentazione morfologica è fornita alla rete tramite la specificazione dei suffissi verbali regolari, perfettivo –to, imperfettivo -va, e irregolari, mentre la rappresentazione semantica è più complessa e problematica. Diverse metodologie di rappresentazione semantica vengono sperimentate per ottenere una classificazione da parte della rete neurale dei predicati in gruppi omogenei, che si avvicinino alle classi azionali. Particolare attenzione è rivolta all'aspetto innovativo della simulazione dell'apprendimento della seconda, e non della prima lingua, su cui si sono finora concentrati gli studi. Negli esperimenti si propongono varie strategie per rispondere all'esigenza di simulare la conoscenza linguistica della prima lingua presente negli apprendenti adulti di una seconda lingua, tenendo conto del ruolo che la distanza tipologica fra la L1 e la L2 ricopre nel percorso acquisizionale. L'applicazione di strumenti tecnologici

all'apprendimento delle categorie tempo-aspettuali nell'italiano L2 offre, così, una metodologia innovativa per analizzare il processo di acquisizione della seconda lingua, il ruolo svolto dalla madre lingua degli apprendenti e i meccanismi cognitivi che operano nell'evoluzione delle strutture linguistiche all'interno di un sistema in formazione.

Riferimenti bibliografici:

Banfi, E. & Bernini, G. (2003), "Il verbo", In Giacalone Ramat, A. (ed), Verso l'italiano, Roma, Carocci, 70-115.
Bertinetto, P.M. (1986), Tempo, aspetto e azione nel verbo italiano. Il sistema dell'indicativo. Firenze, Accademia della Crusca.
Kohonen, T., (2001), Self-Organizing Maps, 3° ediz., Springer&Verlag. - Li, P. & Shirai, Y. (2000), The Acquisition of Lexical and Grammatical Aspect, Berlin, Mouton de Gruyter.

**Fabio Tamburini**
Università di Bologna, Dipartimento di Studi Linguistici e Orientali

**Analasi Automatica della Prominenza Frasale nella Lingua Parlata: Un Approccio Acustico.**

Le tematiche relative all'analisi della lingua parlata in una situazione comunicativa autentica costituiscono un punto cruciale sia nell'ambito dell'analisi linguistica sia nell'ambito di quelle discipline che si occupano della progettazione e della costruzione di strumenti tecnologici che trattano tali informazioni. L'integrazione di questi differenti aspetti d'indagine ha ricevuto un interesse preminente negli ultimi anni e sembra essere un punto di vista privilegiato per l'analisi dei fenomeni che coinvolgono i vari aspetti della lingua parlata. Lo studio del parlato ha storicamente intessuto rapporti molto stretti con le tecnologie, da un punto di vista sia metodologico sia applicativo, per la natura stessa delle analisi compiute sugli oggetti dello studio e per la complessità dei fenomeni che sono coinvolti in analisi linguistiche di questo tipo. Il presente contributo si inserisce in questa prospettiva di indagine, sia linguistica che modellistico-tecnologica, e presenterà uno studio che si inserisce nel filone delle indagini nell'ambito dei fenomeni prosodici e della loro identificazione con metodi automatici, proponendosi di esplorare le complesse relazioni che intercorrono, a diversi livelli, tra il fenomeno percettivo della prominenza prosodica e i fenomeni di tipo acustico che possono essere desunti dalla componente fonetico/acustica degli enunciati. La prominenza risulta essere uno dei fenomeni prosodici maggiormente interessanti e anche uno di quelli più studiati, proprio per l'importanza che riveste nell'ambito dei processi comunicativi umani. Nonostante la notevole quantità di studi nel settore, il problema di un'accurata identificazione del fenomeno con metodi automatici

risulta ancora di notevole interesse a causa della complessità dei fenomeni coinvolti e della mancanza di adeguati strumenti di indagine che consentano un'elaborazione il più possibile automatica di queste informazioni. Nonostante i grandi sviluppi tecnologici ai quali abbiamo assistito in questi anni restano ancora numerosi problemi da affrontare di rilevanza cruciale, sia scientifica che tecnologica, per gestire correttamente i fatti prosodici della lingua parlata. Questo lavoro, fortemente basato su informazioni di tipo fonetico/acustico, tenta di costruire, visti anche i lavori sviluppati finora in questa direzione, un modello che consenta l'identificazione della prominenza, permettendo la creazione di sistemi automatici di etichettatura del fenomeno. Si è ritenuto opportuno che tale indagine utilizzasse come uniche informazioni i parametri derivabili direttamente, anche se in modo estremamente articolato, dall'espressione sonora dell'enunciato. Il modello proposto e l'algoritmo che implementerà il riconoscimento della prominenza non si basano quindi su fonti di informazioni alternative, quali trascrizioni degli enunciati, sia ortografiche che fonetiche, risorse linguistiche etichettate dal punto di vista fonetico, fonologico, prosodico, e nemmeno risorse che contengono informazioni di tipo segmentale sugli enunciati. L'unica informazione fornita all'algoritmo di annotazione è la digitalizzazione dell'enunciato (waveform). Numerosi studi indicano che il fenomeno percettivo della prominenza frasale è supportato, dal punto di vista prosodico, da due fenomeni: il pitch accent e lo stress. Dal punto di vista fonetico/acustico i pitch accent risultano essere connessi con movimenti nel profilo della frequenza fondamentale (F0) e con l'energia globale all'interno del nucleo sillabico, mentre lo stress, inteso come l'indebolimento o il rinforzo degli stress lessicali a livello di frase, presenta una forte correlazione con la durata temporale del nucleo sillabico e la quantità di energia in specifiche bande di frequenza (spectral emphasis). Con questo lavoro intendiamo mostrare come con un'attenta misurazione di questi parametri e l'identificazione di opportune relazioni tra essi e i fenomeni prosodici oggetto di questo studio, sia possibile costruire un sistema automatico capace di identificare le sillabe prominenti nel parlato continuo in modo affidabile, utilizzando unicamente informazioni ricavabili dalla waveform dell'enunciato. Di estremo interesse sembra essere la prospettiva interlinguistica che si presenta: partendo da una analisi concentrata prevalentemente sulla lingua inglese, è stato possibile introdurre una nuova metodologia di analisi fonetica basata sulla manipolazione di opportuni parametri acustici al fine dell'individuazione del fenomeno della prominenza in una prospettiva di comparazione interlinguistica e di classificazione tipologica. Gli esperimenti che descriveremo prendono in esame il problema da questo punto di vista: partendo da risultati ottenuti sull'inglese, relativi sia a parlato continuo sia spontaneo, le metodologie proposte sono state applicate ad altre lingue europee consentendo l'identificazione delle sillabe prominenti con un accordo con annotatori umani confrontabile con quello ottenibile tra esperti del settore. I risultati ottenuti sulle lingue esaminate, ancorché preliminari, possono far supporre un certo livello di regolarità tra le lingue tradizionalmente indicate come stress accented. Lo studio di estensioni del metodo proposto in questa

prospettiva potrebbe portare a interessanti conclusioni sulla natura della prominenza come fenomeno universale, certamente per questo gruppo di lingue, ma, in un'ottica più ampia, anche per lingue tradizionalmente classificate differentemente. Questo risulterebbe particolarmente significativo in quanto la prominenza, nella sua manifestazione acustica, assume differenti connotazioni a seconda della lingua in esame e, benché alcune lingue sembrino mostrare tratti comuni, non risulta finora possibile definire una matrice costante tra di esse.

# RIASSUNTI DEI POSTER

**Adriano Allora**
Università di Torino

## Presentazione del motore di ricerca EnTeR (Engine for Textual Reserachers)

0.INTRODUZIONE EnTeR è un algoritmo di indicizzazione, ricerca ed interazione utente-corpus e calcolatore-corpus. Nel presente testo verranno: - fornite informazioni sul contesto nel quale è stato sviluppato e sulle sue finalità; - descritti due problemi centrali del suo sviluppo; - presentate le caratteristiche fondamentali dell'algoritmo.

1.CONTESTO E SCOPI EnTeR è stato pensato e sviluppato collateralmente ad un progetto di linguistica testuale che ha tra i propri scopi il trattamento e la pubblicazione di un corpus di italiano scritto nei NewsGroups ed un corpus di e per apprendenti di italiano L2. Gli interventi nei NewsGroups e i testi degli apprendenti, e le finalità del progetto, richiedono molto metatesto ed una certa varietà di etichette per fenomeni linguistici specifici (come le autocorrezioni degli apprendenti o le fantasiose firme degli utenti dei NGs). La grande quantità di metatesto è spesso un problema anche per i migliori motori di ricerca testuale (come l'IMS Corpus Wokbench, tutt'ora impiegato nel nostro corpus). E' stato quindi pensato un programma che potesse innestarsi nell'attuale flusso di lavoro per proporsi come alternativa nella fase finale del trattamento dei testi. Tali testi sono dunque etichettati (manualmente) con alcune differenti DTD compatibili con lo standard TEI per il metatesto e con il Treetagger sviluppato da H. Schmidt (Schmidt 1994) per il POS-tagging e la lemmatizzazione. Anche se le caratteristiche di EnTeR saranno descritte nel seguito, una sua specificità va anticipata: doveva poter essere modificato non solo da professionisti della programmazione - come per esempio accade per l'ottimo Lucine (Gospodnetic/Hatcher 2005) -, ma anche da quanti avessero una discreta competenza attiva di quella lingua per la programmazione, il Perl, che tutti i linguisti computazionali conoscono.

2.DUE PROBLEMI 2.1.IL PROBLEMA DELL'INDICIZZAZIONE Il termine indicizzazione può avere due differenti scopes: può essere riferito ai testi (a partire dalla norma ISO 5963/1985 per passare da Feldman 2000, Sutton & McEnery 1992 e Spinelli 2005 per una rassegna in ambito bibliotecario) oppure alle singole parole di un testo (Kientzle 1999). Nel primo caso l'indicizzazione consiste nell'attribuzione di etichette semantico-informative al testo finalizzate ad una ricerca ragionata. Nel secondo caso l'indicizzazione consiste nell'elaborazione di strategie di reperimento veloce di parole o gruppi di parole in un corpus. L'accezione di riferimento è la seconda. Quello dell'indicizzazione è stato il primo problema da affrontare: non solo il linguaggio accessibile che descrivesse l'algoritmo doveva essere accessibile, ma l'algoritmo stesso doveva essere formulato in modo da renderne evidente il funzionamento. 2.2.IL PROBLEMA DEL METATESTO Il tipo di ricerca che EnTeR doveva rendere facile è quello della linguistica testuale (cfr. Beaugrande, Dressler [(1981)1994]) nel quale non ha senso proporre come output dei frammenti di testo perché il testo

nella sua completezza interessa il linguista. Se il risultato di una ricerca con EnTeR doveva quindi essere una selezione di testi completi, si potevano presentare nella stessa sede le relative informazioni metatestuali. Il secondo problema è dunque stato distinguere, all'interno del testo, ciò che era metatesto da ciò che non lo era, ed indicizzare tutto.

3.CARATTERISTICHE DEL SOFTWARE 3.1.OBIETTIVI RAGGIUNGIBILI L'idea originale prevedeva un motore di ricerca che: - permettesse ricerche per parte del discorso ma anche per parametri metatestuali; - mostrasse i testi nella loro interezza; - sapesse scremare risultati identici (p.e. lo stesso testo ospitante due diverse occorrenze del medesimo fenomeno); - potesse essere modificato alle condizioni sopra menzionate; - sapesse trattare le novità che un nuovo corpus o frammento di corpus comportano ed importano in termini di nuove etichette. Fin dall'inizio si è trattato, evidentemente, di uno strumento non programmaticamente innovativo nei risultati, e neppure nei metodi, quanto nelle forme in cui questi metodi dovevano esprimersi. Tali forme hanno tre caratteristiche fondamentali: accessibilità, modularità e automazione. 3.2.ACCESSIBILITA' Perché EnTeR fosse accessibile è bastato provare a tradurre in un rudimentale Perl la descrizione di un procedimento cercando una via di mezzo tra comprensibilità ed efficacia/efficienza. In particolare, ad ogni gruppo forma + POS + lemma è stato assegnata una serie di indirizzi in termini di file, frase e numero di parola, con una struttura molto chiara e insolita. 3.3.MODULARITA' Disporre di uno strumento modulare ha richiesto che il singolo processo sopra menzionato venisse segmentato in una serie di differenti processi: distinzione testo-metatesto; indicizzazione; elaborazione delle interfacce di interrogazione; esecuzione delle ricerche. 3.3.AUTOMATICITA' Infine, intendendo per automazione la possibilità di lasciare al calcolatore l'assolvimento di alcune funzioni, sono state scritte semplici routines che si occupassero della generazione dell'interfaccia e delle verifiche sul corpus. Ad esempio, il modulo di indicizzazione crea una lista di tutte le parti del discorso che possono essere oggetto di interrogazione (perché presenti nel corpus) che il modulo di generazione delle interfacce userà per creare dei menù a tendina. 3.4.CONCLUSIONI Effetti macroscopici delle tre caratteristiche di EnTeR sono: leggibilità e personalizzabilità estrema a chiunque desideri copiarlo sul proprio calcolatore, infatti non richiede installazioni aggiuntive alla distibuzione base di Perl. Dunque, la ricerca di uno specifico tipo di forma ha implicato il ricorso a metodi non sempre ortodossi (come l'indicizzazione per frase, che permette la distinzione tra ricerche intra ed extrafrasali). E nonostante l'assenza di una volontà innovativa nei risultati, la pratica di metodi eterodossi ha avuto come naturale conseguenza un certo grado di innovazione anche nei risultati.

Riferimenti bibliografici:

Beaugrande, R.-A. / Dressler, W.U. (1981), Einfuehrung in die Textlinguistik, Tubinga, Niemeyer. Trad. it. di S. Muscas (1994), Bologna, Il Mulino.
Feldman, Susan (2000), "The answer machine", in Searcher: the magazine for

database professional, vol 8. n1.
Gospodnetic, O. / Hatcher, E (2005), Lucene in action, Greenwich (CT - USA) Manning. Kientzle, T. (1999), Full-Text Searching in Perl, in Dr.Dobb's Journal, Gennaio 1999. Reperibile in:
http://www.ddj.com/documents/s=907/ddj9901c/9901c.htm?temp=+pXvfrN94m
Schmidt, H. (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees. Reperibile presso:
http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html
Spinelli, S. (2005), Introduzione all'indicizzazione. Reperibile in:
http://biocfarm.unibo.it/~spinelli/indicizzazione/
Sutton S. / McEnery A.M. (1992) 'Information retrieval and corpora', New Dimensions in Corpus Linguistics, Berlin, Mouton, pp 152-165.

**Basilio Calderone, Pier Marco Bertinetto**
Scuola Normale Superiore di Pisa, Laboratorio di Linguistica

**La sillaba come stabilizzatore di forze fonotattiche. Per una modellizzazione**

Lo studio tenta di formalizzare, all'interno del framework connessionista, un processo fonologico classico: la sillabazione. In una prospettiva computazionale-simbolica, l'individuazione di possibili confini sillabici si riduce ad una semplice operazione di parsing, ottenuta per mezzo di qualche algoritmo dedicato rule-based. In un approccio sub-simbolico [3, 4, 5, 7], la sillabazione è ricondotta alla capacità di generalizzazione che il modello neurale, ad apprendimento ultimato, è in grado di esibire. In altri termini, i principi che governano il processo di sillabazione e l'organizzazione interna della sillaba non sono definiti a priori da parte dello sperimentatore, ma piuttosto appaiono come il risultato finale dell'attività simultanea di più unità elementari operanti in parallelo (nodi). In tale prospettiva, la nostra modellizzazione conferisce natura 'emergentista' alla sillaba e al processo di sillabazione, entrambi considerati come configurazioni di attività emergenti (o pattern) derivati da effetti di tipo 'collettivo', generati dalle interazioni tra le singole unità mediante eccitamento o inibizione delle loro connessioni. Per le simulazioni condotte, è stata implementata una rete neurale (NN) di tipo feedforward a due strati. Diversi corpora di addestramento sono stati confezionati in italiano, spagnolo ed inglese. Ogni corpus è stato congegnato per esprimere una rappresentatività minima dei vari tipi sillabici di ogni lingua. Nel dettaglio, il corpus dell'italiano è formato da 51 tipi sillabici, quello spagnolo da 36 tipi sillabici, l'inglese da 78 tipi sillabici. Durante la fase di addestramento, i coefficienti di connessione del sistema neurale sono inizializzati in maniera pseudo-random e gradualmente modificati, per minimizzare il valore d'errore, mediante l'algoritmo back-propagation. Ogni segmento di input è stato codificato utilizzando una rappresentazione distribuita di tipo binaria. Sette classi naturali sono state adottate, come prima approssimazione in questa fase iniziale della nostra ricerca, per specificare la

natura di ogni segmento: V (vocale), G (approssimante), L (liquida), N (nasale), F (fricativa), O (occlusiva) e A (affricata). I vettori di input per il protocollo di addestramento sono stati creati utilizzando una codifica 'a finestra': ogni segmento è stato codificato tenendo conto del contesto fonotattico sia a sinistra sia a destra. Tale accorgimento, in fase di training, fa sì che ogni segmento sia elaborato in corrispondenza del suo intorno fonotattico infralessicale. Ad un input di segmenti fonologici considerati all'interno del loro contesto fonotattico, il sistema fornisce come risposta di output un contesto sillabico all'interno del quale compaiono gli stessi segmenti dell'input iniziale, ma il cui legame definisce ora unità sillabiche. Teoricamente, quindi, il sistema neurale funge da filtro, passando da un contesto fonotattico (con finestra costante: il segmento è computato con gli altri due segmenti che lo accompagnano a sinistra e a destra) ad un contesto sillabico (con finestra variabile: il segmento viene restituito assieme ai segmenti costituenti l'unità sillabica). Nella nostra modellizzazione, quindi, la singola unità sillabica è concepita come un attrattore di forze fonotattiche, caratterizzato da valori di repulsione ed attrazione con i diversi segmenti fonologici. Diversi tipi sillabici attraggono differenti contesti fonotattici con intensità variabile. Una struttura sillabica è definita dalle mutue attrazioni fonotattiche che i vari segmenti intrattengono al suo interno. L'accettabilità o meno di una sillaba è decisa dall'equilibrio di attrazione che i vari segmenti stabiliscono gli uni con gli altri. Ad apprendimento ultimato (25000 cicli e learning rate 0.4) la percentuale di sillabazioni corrette, per le forme presenti nei corpora, è stata del 100%. Testato su forme inedite, non presenti cioè durante l'addestramento, il sistema mostra una robusta capacità di generalizzazione riuscendo, sulla base della conoscenza appresa, a riconoscere tipi sillabici nuovi e a sillabarli correttamente. Nuovi tipi sillabici sono quindi scoperti dal sistema per tutti i corpora interessati (italiano, spagnolo, inglese). Il nostro modello neurale definisce i suoi valori di uscita all'interno di un continuum numerico reale tra 0 e 1. Ciò significa che un'uscita di sillabazione non sarà mai pienamente 0 o 1, ma loro approssimazioni (0.012 e 0.987, ad esempio). Tale comportamento consente analisi quantitative raffinate, miranti a ritagliare i possibili confini sillabici. È stato in particolare indagato il comportamento del nesso consonantico /s/ + C [1, 2, 6]. Tale nesso non compariva mai all'interno delle parole dei tre corpora usati per il training, ma è stato volutamente lasciato come forma inedita per testarne la capacità di elaborazione da parte del sistema. Il sistema, grazie alla proprietà di generalizzazione, ha fornito un'evidenza eterosillabica per il nesso /s/ + C in spagnolo e in inglese. Una classificazione diversa è emersa invece per i dati dell'italiano. La rete neurale ha collocato il nesso /s/ + C in una posizione a metà strada tra l'eterosillabicità e la tautosillabicità; quasi una zona di equilibrio (caratterizzata da valori sfumati: 0.543 o 0.496, ad esempio). Sono in fase di sperimentazione modelli che sintetizzano al loro interno informazioni sulla natura fonotattica del dato e informazioni sulla componente sonora (scala di sonorità) espressa da ogni segmento, spacchettando le classi fonetiche inizialmente assunte come prima approssimazione (occlusiva, liquida etc.) nei singoli foni che le costituiscono.

Riferimenti bibliografici:

[1] Bertinetto, Pier Marco. 1996. Psycholinguistic evidence for syllable geometry: Italian and beyond. In Rennison & Kähnhammer (eds.), Phonologica 1996. Syllables!? Holland Academic Graphics:1-28.
[2] Bertinetto, Pier Marco. 2004. On the undecidable syllabification of /sC/ clusters in Italian: Converging experimental evidence. Italian Journal of Linguistics / Rivista di Linguistica, 16:349-372.
[3] Goldsmith, John. 1992. Local Modeling in Phonology. In Steven Davis (ed.), Connectionism: Theory and Practice. Oxford, OUP:229-246.
[4] Joanisse, Marc. 1999. Exploring syllable structure in connectionist networks. In Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, CA.
[5] Laks, Bernard. 1995. A connectionist account of French syllabification. Lingua, 95:349-372.
[6] Marotta, Giovanna. 1995. La sibilante preconsonantica in italiano: Questioni teoriche ed analisi sperimentale. In Ajello & Sani (eds.), Scritti linguistici e filologici in onore di Tristano Bolelli. Pisa, Pacini:393-437.
[7] Stoianov, Ivelin & John Nerbonne. 1998. Modeling the phonotactics of Natural Language Words with SRNs - part two: Exploring phonetic data representations. In CLIN'98: Computational Linguistics in Netherlands. Leuven, Belgium.

**Felice Dell'Orletta, Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli**
Università di Pisa, ILC - CNR Pisa

**Vincoli grammaticali ed indagine inter-linguistica: micro-analisi quantitativa della codifica del soggetto e dell'oggetto diretto in italiano e in ceco.**

I corpora annotati a livello sintattico (o treebank) sono una preziosa fonte di informazione per un ampio spettro di ricerche sul linguaggio: in linguistica teorica, per la individuazione di esempi (o controesempi) a supporto di una particolare analisi; in psicolinguistica, per comparare la distribuzione di frequenza di una costruzione con i dati elicitati dai parlanti o con risultati sperimentali; in linguistica computazionale, per la creazione semi-automatica di repertori lessicali, per l'induzione di grammatiche, ecc. Accanto a questi usi più consolidati, le treebank offrono anche l'opportunità di sperimentare nuovi metodi computazionali nel campo dell'indagine grammaticale comparata. Questi si basano sull'applicazione di modelli di apprendimento automatico ai dati di corpora annotati a livello sintattico allo scopo di identificare le dinamiche di variazione dei vincoli grammaticali a livello interlinguistico. Nella misura in cui i metodi computazionali permettono di simulare processi dinamici nel linguaggio - acquisizione, elaborazione e mutamento – il loro uso nell'indagine linguistica consente di sviluppare modelli empiricamente fondati dell'interazione dei vincoli

grammaticali e della loro variazione in lingue diverse, tenendo al tempo stesso in considerazione il loro rapporto sia con le dinamiche nell'apprendimento di una lingua sia con quelle nei processi di mutamento del sistema grammaticale. All'interno di questa più ampia prospettiva di ricerca, in questo lavoro vogliamo illustrare alcuni esperimenti legati all'esplorazione automatica dell'interazione di vincoli grammaticali induttivamente ricavati da corpora per l'italiano e per il ceco. La ricerca più recente nell'acquisizione ed elaborazione del linguaggio naturale conferma l'idea che la competenza grammaticale consista in un sistema di molteplici vincoli paralleli fortemente integrati (Seidenberg & MacDonald 1999). Esiste inoltre crescente consenso intorno a due proprietà essenziali dei vincoli grammaticali, ovvero la loro natura i.) probabilistica e ii.) intrinsecamente funzionale, risultato dell'interazione di tipi diversi di informazione linguistica e non-linguistica (sintattica, semantica, pragmatico-informazionale, ecc.). Questi tratti emergono in maniera ancora più netta quando ci focalizziamo su uno degli aspetti centrali della grammatica, ovvero le relazioni sintattiche. Ad esempio, l'evidenza psicolinguistica mostra che i parlanti imparano a identificare i soggetti e gli oggetti delle frasi combinando vari tipi di indicatori funzionali e probabilistici, quali l'ordine delle parole, l'animatezza nominale, la definitezza, il caso, l'accordo ecc. Inoltre, la salienza relativa di ciascuno di questi indicatori può variare considerevolmente da lingua a lingua (Bates et al. 1984). Se i vincoli grammaticali sono intrinsecamente probabilistici (Manning 2003), il percorso di formazione di una grammatica può essere simulato come un processo di costruzione di un modello stocastico a partire dall'input linguistico. In linguistica computazionale, i modelli basati sul principio della massima entropia (o ME) si sono dimostrati algoritmi di apprendimento automatico molto efficaci per modellare vari compiti linguistici. In questo lavoro, presentiamo l'applicazione di un modello ME per l'identificazione automatica di soggetti e oggetti diretti in italiano e in ceco. L'interesse di questa indagine particolare è dato dal fatto che la distribuzione di soggetti e oggetti è notoriamente governata da vari tipi di vincoli grammaticali, qui modellati come vincoli funzionali e probabilistici (cf. Dell'Orletta et al. 2005). Inoltre, la comparazione di queste due lingue è di notevole interesse principalmente per i diversi gradi di variazione che esse mostrano nel modo in cui le relazioni grammaticali sono codificate. In entrambi i casi, infatti, l'ordine delle parole non è un indicatore affidabile per identificare soggetto e oggetto, mentre i tratti semantici, come ad esempio l'animatezza, giocano un ruolo preponderante. Dall'altro lato, in ceco le relazioni grammaticali sono esplicitamente codificate attraverso marche di caso, e l'ordine delle parole è molto più libero che in italiano. Infine, in ceco la struttura informazionale della frase e l'articolazione di topic e focus, condizionano pesantemente la distribuzione degli argomenti del verbo e interagiscono con la nozione di soggetto sintattico. Per questi motivi, il nostro esperimento fornisce un punto di osservazione ottimale per l'indagine su un nucleo di vincoli centrale per il sistema grammaticale, indagato sia nel suo emergere dai dati empirici, sia nelle sue diverse modalità di interazione nelle due lingue. Nell'addestramento del modello ME abbiamo fatto uso di diversi tipi di tratti linguistici, estratti

rispettivamente dalla TreSSI, treebank sintattico-semantica dell'italiano (Montemagni et al., 2003) e dalla Prague Dependency Treebank (PDT, Bohmova et al. 2003). Per i modelli addestrati presenteremo una valutazione sia quantitativa sia in una prospettiva linguistico-qualitativa. La valutazione qualitativa si rivolge ai tipi di vincoli grammaticali acquisiti e alle loro differenti salienze nelle due lingue. I risultati mostrano che sia per l'italiano che per il ceco il sistema è in grado di acquisire da dati testuali un insieme di vincoli grammaticali psicolinguisticamente plausibili e linguisticamente motivati, con prestazioni particolarmente accurate nell'identificazione automatica di soggetti e oggetti.

Riferimenti bibliografici:

Bates E., MacWhinney B., Caselli C., Devescovi A., Natale F., Venza V., 1984. A crosslinguistic study of the development of sentence interpretation strategies. Child Development, 55: 341-354.
Bohmova A., Hajic J., Hajicova E., Hladka B., 2003. The Prague Dependency Treebank: Three-Level Annotation Scenario, in A. Abeille (ed.) Treebanks: Building and Using Syntactically Annotated Corpora, Kluwer Academic Publishers, pp. 103-128.
Dell'Orletta, F., Lenci, A., Montemagni, S., Pirrelli, V., 2005. Climbing the Path to Grammar. A Maximum Entropy Model of Subject/Object Learning, in Proceedings of the ACL 2005 Workshop on Psychocomputational Models of Human Language Acquisition, Ann Arbor MI. Manning C. D., 2003. Probabilistic syntax. In R. Bod, J. Hay, S. Jannedy (eds), Probabilistic Linguistics, MIT Press, Cambridge MA: 289-341.
Montemagni S. et al. 2003. Building the Italian syntactic-semantic treebank. In Abeillé A. (ed.) Treebanks. Building and Using Parsed Corpora, Kluwer, Dordrecht: 189-210.
Seidenberg M. S., MacDonald M. C. 1999. A probabilistic constraints approach to language acquisition and processing. Cognitive Science 23(4): 569-588.

**Rodolfo Delmonte**
Università di Venezia

**VIT – Venice Italian Treebank: caratteristiche sintattico-semantiche e quantitative.**

In questo lavoro descriveremo il VIT, Treebank (Sintattico) dell'Italiano (dell'Università) di Venezia (Venice Italian Treebank) di 320.000 parole, e 10.200 frasi per la porzione di testo scritto, creato dal Laboratorio di Linguistica Computazionale del Dipartimento di Scienze del Linguaggio. Focalizzeremo la nostra attenzione sulle caratteristiche sintattico-semantiche del treebank che sono in parte legate al tagset adottato, in parte sono dovute alla teoria linguistica

di riferimento, e infine sono come ogni treebank legate alla lingua prescelta, l'italiano. Con esempi presi anche da treebank disponibili per altre lingue, mostreremo quali sono le differenze e le motivazioni teoriche e pratiche dietro le scelte fatte. Dedicheremo infine una parte della nostra presentazione alla analisi quantitativa dei dati del nostro treebank confrontandoli con gli altri. Il discorso in generale cercherà di dimostrare come l'apprendimento di una grammatica o di un parser in maniera automatica da un treebank, non possa dare gli stessi risultati passando da un treebank all'altro, e come questo processo sia dipendente da fattori sostanziali come il quadro linguistico di riferimento adottato per la descrizione strutturale nonché in ultima analisi, la lingua descritta. Schematicamente, la nostra annotazione ha adottato la teoria X-barra con una eccezione, non utilizziamo il livello di struttura di VP: abbiamo cioè preferito considerare il gruppo verbale direttamente posizionato a livello di F dove si troverà, se espresso lessicalmente anche il SN soggetto, e la struttura dei complementi, questa specializzata per sottocategorizzazione in diversi tipi. In questo, il VIT ha assunto la stessa scelta del NEGRA, treebank del tedesco, mentre il PennTreebank si distingue per una scelta meno dettagliata e più "skeletal" come hanno commentato gli stessi estensori delle specifiche. Se si considera la nostra scelta nel quadro delle LPCFG (Lexicalized Probabilistic Context-Free Grammars), la testa del VP (IBAR nel nostro caso) è la testa della F e, nel VIT non deve essere estratta da nessuna sottostruttura trovandosi già a livello di F; invece nel PennTreebank, la testa potrebbe essere figlia di diversi nodi VP, quanto sono gli ausiliari o modali che la precedono. In un loro lavoro recente (2004) Corazza et al. utilizzano una porzione del VIT – 90mila tokens prodotti nel progetto SITAL – per verificare la possibilità di addestrare un parser statistico probabilistico sulla base di procedure sperimentate per la lingua inglese con il PennTreebank da Collins e da Bikel. Dal momento che i risultati ottenuti sono molto scarsi, gli autori si domandano se la scarsa performance possa essere dovuta a difficoltà intrinseche nella struttura dell'italiano, alla diversa teoria linguistica adottata (vedi la mancanza del nodo VP), o ancora al diverso tagset adottato, molto più dettagliato. Dai commenti di Bikel al lavoro di Collins, considerato tuttora un landmark per la creazione di parser probabilistici, risulta che il lavoro di costruzione del modello del linguaggio viene preceduto da un nutrito lavoro di preprocessing e che quindi non si lavori sui dati "raw" del treebank per produrre il modello ma su una sua versione addomesticata all'uopo. Le probabilità vengono associate a relazioni strutturali lessicalizzate, in cui è cioè presente la testa del costituente da codificare, che hanno come obiettivo quello di aiutare le decisioni nella scelta di argomenti vs. aggiunti, di livelli di attaccamento di un modificatore e altre simili questioni di difficile risoluzione se basate solo su regole simboliche. A questo scopo è stato necessario intervenire sul treebank marcando i complementi, marcando le frasi prive di soggetto e con il soggetto in posizione inversa, ecc. Il lavoro di preprocessing compiuto da Corazza et al. si limita invece all'utilizzo di lemmi al posto di forme di parola come testa dei costituenti lessicalizzati. Dall'esperimento elaborato sulla base della teoria dell'informazione risulta che la differenza in performance non può

essere ascritta al numero di regole e quindi al tipo di annotazione introdotta, ma alla scarsa prevedibilità delle loro relazione strutturali. Dunque, il cattivo funzionamento di un parser statistico addestrato su un treebank può essere messo in relazione al quadro linguistico di riferimento scelto dagli annotatori e in ultima analisi alla lingua di riferimento. Sembra intuitivo poter affermare che maggiore la regolarità strutturale di una lingua o della sua rappresentazione, maggiore sarà la sua riproducibilità su basi statistiche: al contrario, una lingua che contenga molti casi ricorrenti una volta sola, allora è meno probabile una buona resa statistica del modello – questo viene definito sparseness/sparsity. Dai dati quantitativi globali in nostro possesso si evince che molto più della metà delle frasi italiane non hanno soggetto espresso lessicalmente: questo fatto viene accuratamente corretto nelle procedure di pre-processing. Nel caso del PennTreebank, invece ci sono 4647 frasi classificate con il nodo S-TPC di struttura topicalizzata e le frasi a SINV soggetto inverso sono 2587, questi numeri sono in percentuale molto bassi, attorno al 10% delle strutture di frase complessive. A questo scopo abbiamo analizzato tutte le strutture che possono svolgere ruolo di modificazione o di argomento in italiano, in particolare le strutture aggettivali che possono ricorrere liberamente in posizione postnominale anche lontane dalla testa – quest'ultima non possibile in inglese. I modificatori in posizione non canonica – non adiacenti alla testa costituiscono il 25%. Se si guarda al livello di frase, il dato che salta agli occhi è che le strutture discontinue ricorrono con percentuali dell'ordine del 20/30%, o addirittura del 60% se ci si riferisce alle frasi con soggetto lessicalmente inespresso. Inoltre vanno considerati altri elementi che possono introdurre problemi di discontinuità o di non canonicità: il numero di F3 o frammenti è alquanto elevato rispetto al numero degli enunciati complessivi, 3237 (32%); il numero di frasi parentetiche è molto elevato, 4162. Da qui la necessità di introdurre correttivi per permettere alla fase di apprendimento di distinguere frasi di diverse tipologie (soggetto espresso in posizione preverbale, soggetto espresso in posizione postverbale, soggetto lessicalmente inespresso). Questo nell'esperimemento di Corazza et al. non è avvenuto con la conseguenza di rendere il modello statistico poco informativo: nella parole di Collins, la parametrizzazione non è stata scelta in modo adeguato.

**Gianluca Giorgolo**
Università di Pavia, Dipartimento di Linguistica Teorica e Applicata

**Un modello per l'integrazione multimodale basato sulle grammatiche categoriali.**

La maggiore affidabilità offerta dalle odierne tecniche di riconoscimento vocale e gestuale ha aperto la strada per lo sviluppo di applicazioni in grado di semplificare notevolmente l'interazione fra l'utente e la macchina. La possibilità di comunicare con un agente informatico utilizzando la naturale combinazione di tutte le risorse espressive solitamente impiegate nella comunicazione umana

rappresenta certamente una delle prospettive più interessanti per lo sviluppo di sistemi "intelligenti", o quantomeno maggiormente user-friendly. Di seguito descriverò un modello formale per la gestione di alcuni aspetti della comunicazione multimodale, più specificamente l'integrazione di due canali comunicativi e la ricostruzione di un unica informazione semantica a partire da questi. La complessità di una comunicazione convogliata su più canali (parole, gesti, espressioni facciali...) richiede uno studio attento delle dinamiche che caratterizzano questo fenomeno "in natura". Senza tale analisi preliminare, sarebbe impossibile progettare un modello computabile in grado di gestire l'effettivo svolgimento di una comunicazione multimodale. Perciò, nello studio qui riassunto, si è seguita una metodologia conforme a queste semplici osservazioni. Per prima cosa si è cercato di definire con precisione gli oggetti poi formalizzati nel modello. Il campo d'analisi è stato ridotto al canale verbale ed a quello della gestualità co-verbale. Questa limitazione ha permesso di sfruttare un quantitativo notevole di ricerche già condotte su questo particolare sottoinsieme della comunicazione multimodale. La teoria fondamentale su cui ci si è basati per lo sviluppo effettivo del modello è quella proposta da David McNeill e ripresa da numerosi altri ricercatori. Dal complesso di queste analisi sono stati derivati alcuni concetti teorici poi utilizzati come linee guida per la formalizzazione. In breve, si è scelto di aderire all'interpretazione per cui il messaggio verbale ed la gestualità che lo accompagna avrebbero una medesima origine a livello cognitivo nel parlante: il contenuto informativo è quindi unico ma viene espresso attraverso due differenti modalità. Secondo McNeill, i due canali presentano degli stretti rapporti, indice di questa origine comune, e corrispondenti a tre tipi di "sincronia" riscontrabili nella comunicazione audio-visiva: sincronia fonologica, semantica e pragmatica. Molto semplicemente il primo tipo di sincronia attesta la sincronizzazione temporale fra la fase espressiva del gesto e la sillaba più prominente dell'unità linguistica co-occorrente. Gli altri due tipi di sincronia affermano rispettivamente la concordanza semantica dei contenuti espressi attraverso le due modalità e similmente una concordanza a livello di caratterizzazione pragmatica. Quest'ultimo tipo di sincronia non è stato per ora preso in considerazione. Questi semplici assunti teorici nascondono tuttavia una questione ancora aperta: è infatti difficile definire con precisione la natura degli oggetti "sincronizzati", soprattutto per quanto riguarda il canale verbale (parole, sintagmi, frasi, unità concettuali?). Questo sarà un aspetto fondamentale di cui verrà tenuto conto nel modello proposto. Di pari importanza è la caratterizzazione assegnata agli elementi gestuali. La denotazione di gesto non è infatti univoca nella letteratura. La scelta operata è stata quella di limitare l'interesse alla gestualità spontanea, solo parzialmente codificata socialmente e ampiamente documentata negli studi condotti sulla base delle teorie di McNeill. I gesti verranno perciò considerati come sintetici, ossia non in grado di combinarsi tra loro, e molto spesso privi di un livello formale-simbolico indipendente dal contenuto semantico espresso. Concretamente perciò il modello tenterà di soddisfare i seguenti requisiti: inclusione dei due tipi di sincronia presi in considerazione, rappresentazione dei gesti come entità prevalentemente

semantiche e costruzione di un singolo messaggio informativo a partire dai contenuti dei due canali. La proposta che qui viene presentata consiste nell'estensione dello schema generale delle Grammatiche Categoriali, sfruttando l'impianto composizionale su cui sono basate. Il modello necessita infatti di potere comporre diversi tipi di oggetti (stringhe, gesti e loro combinazioni) rispettando diversi schemi di composizione: sul singolo canale verbale gli elementi saranno combinati secondo una grammatica mentre la combinazione fra canali diversi avverrà rispettando le tanto citate sincronie. Inoltre tali grammatiche sono basate sull'assunto dell'esistenza di un isomorfismo fra composizione grammaticale e composizione del significato, risultando perciò particolarmente adatte per riprodurre i caretteri individuati per il canale gestuale. Il formalismo utilizzato è quello delle Type Logic Grammars derivate dal ben noto Lambek Calculus. Questo tipo di grammatiche descrive tramite una logica modale la dinamica della composizione grammaticale. Il primo passo perciò è stato quello di estendere la rappresentazione logico-semantica (in una struttura di Kripke) degli oggetti presi in esame dalle Type Logic Grammars, includendo gesti e combinazioni di stringhe con gesti. A questo arricchimento semantico è poi corrisposto un potenziamento del linguaggio logico con l'aggiunta di una terna di operatori, atti a riprodurre la dinamica della composizione multimodale. La teoria è stata poi completata con gruppo di assiomi in grado di governare il corrispettivo logico della composizione grammaticale e di quella multimodale, fornendo una preliminare prova di correttezza e completezza per questo sistema logico. Per fornire una effettiva metodologia di "parsing" (inteso come analisi composizionale), e seguendo in toto l'esempio classico del Lambek calculus, il modello è stato trasformato in un equivalente calcolo di sequenti alla Gentzen. Questo ha permesso anche di sfruttare la nota corrispondenza di Curry-Howard per ottenere una analisi semantica in parallelo con quella formale, semplicemente codificando la semantica multimodale tramite dei lambda termini (offrendo tra l'altro un grande potere espressivo). Il modello è stato poi implementato nel prototipo di un interfaccia di controllo per un ambiente domestico (applicazione particolarmente adatta per sistemi di interazione multimodale). Ai fini implementativi il calcolo logico è stato approssimato attraverso un approccio basato su regole, molto simile all'ormai storico calcolo di Ajdukiewicz - Bar-Hillel. Sfruttando l'ormai sperimentata tecnica del parsing deduttivo, l'analisi composizionale e la parallela ricostruzione dell'informazione semantica avviene con una modalità bottom-up, partendo dai dati provenienti da una coppia di riconoscitori e cercando di derivare una categoria bersaglio che rappresenta un tipo di interazione. Al momento il prototipo permette di comunicare tramite un riconoscitore vocale per l'inglese (Sphinx-4), accompagnando il parlato con selezioni e click del mouse su una semplice schematizzazione grafica di un possibile ambiente domestico (riproducendo perciò una sorta di gesto deittico). Test preliminari hanno confermato la possibilità di sfruttare questa metodologia per l'applicazione concreta. Si sta ora tentando di estendere tanto il modello teorico quanto quello applicato in modo da rendere dei molti aspetti della comunicazione multimodale non ancora considerati (livello prosodico e pragmatica soprattutto).

**Mirko Grimaldi**
Università di Lecce, Dipartimento di Filologia, Linguistica e Letteratura

## Cosa possono offrire le neuroscienze alla teoria linguistica?

«I segni linguistici, pur essendo essenzialmente psichici, non sono delle astrazioni; le associazioni ratificate dal consenso collettivo che nel loro insieme costituiscono la lingua, sono realtà che hanno la loro sede nel cervello» (Saussure 1916: 25). A distanza di oltre un secolo dalle riflessioni di Saussure, pare sempre più urgente ed empiricamente motivato chiedersi: cosa può offrire lo studio del cervello alla teoria linguistica? Le diverse e articolate posizioni, che stanno alimentano la discussione sull'argomento, oscillano, in estrema sintesi, fra due opposte idee di fondo, che possiamo così Anche se un neuroscienziato scoprisse che quando noi "pensiamo laÿriassumere: parola cane" – o che quando un altro parlante x pensa la stessa parola nella sua lingua y – nel cervello si attiva una 'configurazione' neurale C non ci direbbe nulla di rilevante né sul processo di significazione legato a quella parola né sul processo di significazione in generale (Putnam 1988). Per cui «[...] it is, for the moment, only a hope that "neurological terms" are relevant for the unification [of the mind-body] problem. [...] Naturalistic inquiry into the mind yields theories about the brain, its states and properties: UG [Universal Grammar], for example. No one knows how to begin to relate these theories to properties of atoms, cells, neurons, or other known structures of the brain [...] Non c'è nessun motivo – date le attuali conoscenze sulÿ(Chomsky 2000: 103). cervello – perché le teorie linguistiche debbano necessariamente essere 'astratte' e non formulate in termini concreti, per esempio in termini neurali. Questo dovrebbe essere invece l'obiettivo futuro della linguistica. Se vogliamo comprendere il linguaggio in una prospettiva biologica non possiamo poi esimerci dall'essere 'concreti' nell'ipotizzare processi e rappresentazioni cerebrali. Per fare ciò ci deve essere lo sforzo di esplorare gli spazi di confine dove le attuali conoscenze neuroscientifiche e le teorie linguistiche entrano in conflitto. «Using neuroscientific knowledge and data for guiding linguistic theorizing appears to be fruitful as well. Thus, neuroscientific data could then constrain linguistic theory. The reverse may also be true. [...] Availability of a brain-language interface of this type, a neuronal language theory, may be a necessary condition for deciding between alternative approaches to grammar as it could be a tool for exploring neuron circuits specific to the human brain» (Pulvermüller 2002: 272). Raccogliere la sfida lanciata dal campo delle neuroscienze vuol dire sviluppare una vocazione (e quindi iniziare a pensare anche a una formazione) fortemente interdisciplinare. L'intervento mirerà a capire come il ricorso ai nuovi strumenti per lo studio dell'attività cerebrale possa facilitare questo percorso. In particolare ci concentreremo su un tipo di strumentazione altamente sofisticata – presente presso il Centro di Ricerca Interdisciplinare sul Linguaggio (CRIL), appena costituito presso l'Università di Lecce – che consente da un lato il monitoraggio dell'attività della corteccia cerebrale in soggetti normali e patologici durante l'espletamento di compiti

linguistici, dall'altro lo studio dei movimenti oculari e dei processi cognitivi che sottostanno alla comprensione del linguaggio scritto. Attualmente di notevole interesse è l'impiego della repetitive Transcranial Magnetic Stimulation (rTMS), che, a quanto pare, meglio di altre tecniche di "neuroimaging funzionale" come la PET o la fMRI, può aiutare a collegare l'attivazione di una o più popolazioni di neuroni a processi inerenti la 'competenza' linguistica (cfr. Pascual-Leone et al. 1997). Infatti, la possibilità di fotografare un'area cerebrale specifica che "si accende" durante una qualsiasi attività cognitiva non risolve una questione fondamentale, cioè sapere se quella "accensione" è neuralmente connessa con la performance svolta dal soggetto, cosa invece possibile con la rTMS (quella presente al CRIL è dotata di 64 canali). Un'ulteriore possibilità di indagine molto efficace sfrutta invece i classici metodi elettrofisiologici come l'Elettroecefalogramma (EEG). Si tratta dei cosiddetti potenziali evocati (ERPs, Event-Related Potentials), cioè piccole modificazioni dell'attività elettrica cerebrale spontanea che riflettono l'attività coordinata di una rete neurale. Tale attività può essere provocata da uno stimolo definibile sperimentalmente, come un nome, un verbo o una frase, per indagare, ad esempio, se verbi e nomi producono differenti potenziali evento correlati, magari a seconda dei contesti in cui si trovano (cfr. Shapiro e Caramazza 2004). Ad un altro livello di analisi si collocano, invece, le registrazioni elettro-oculografiche (movimenti saccadici, ecc.), mediante un sistema a raggi infrarossi. Capire se effettivamente la rilevazione dei movimenti oculari fornisca misure sensibili del tempo necessario per elaborare porzioni di linguaggio vuol dire scoprire se esiste un rapporto fra il processo visivo e la successiva elaborazione cognitiva. Ipotizzando poi di poter utilizzare in sincronia la rTMS e le registrazioni elettro-oculografiche potremmo pensare di individuare quali aree della corteccia producono questo tipo di processi e soprattutto come li producono, e così via. Costruire una teoria dei meccanismi cerebrali del linguaggio è un compito certamente arduo ma che sicuramente dovrà poggiare su fondamenta interdisciplinari. A sua volta la ricerca interdisciplinare presuppone una forte tensione fra studiosi che hanno programmi differenti, strumenti di ricerca diversi, e soprattutto diversi modelli teorici per rispondere alle domande che si pongono. Il successo di ricerche interdisciplinari, e quindi degli obiettivi prefissi, dipende solo dallo sforzo comune di rendere compatibili queste differenze.

Riferimenti bibliografici:

Chomsky, N., 2002, New Horizons in the Study of Language and Mind, Cambridge, Cambridge University Press.
Pascual-Leone, A., Grafman, J., Cohen, L.G, Roth, B.J., Hallett, M., 1997, "Transcranial Magnetic Stimulation: A New Tool for the Study of Higher Cognitive Functions in Humans", in F. Boller, J. Grafman (eds.), Handbook of Neuropsychology, Amsterdam, Elsevier.
Pulvermüller, F., 2002, The Neuroscience of Language, Cambridge, Cambridge University Press. Putnam, H, 1988, Representation and reality, Cambridge, MA,

MIT Press.
Saussure, F. De, 1967 [1916], Corso di linguistica generale, trad. it. a cura di T. De Mauro, Roma-Bari, Laterza.
Shapiro, K., Caramazza A., 2004, "The Organization of Lexical Knowledge in the Brain: The Grammatical Dimension", in M. Gazzaniga (ed.), 2004, The Cognitive Neuroscience, Cambridge, MA, MIT Press, III Edition.

**Alessandro Panunzi e Marco Fabbri**
Università di Firenze, Dipartimento di Italianistica

**Estrazione automatica di parole chiave da documenti in ambiente multilingue. Una infrastruttura linguistica nel progetto IST AXMEDIS.**

Il poster presenta una tecnologia sviluppata all'interno del progetto integrato IST AXMEDIS per l'indicizzazione automatica dell'informazione testuale presente nei contenuti multimediali. Tale ambiente si caratterizza per due qualità rilevanti ai fini dell'indicizzazione: a) il multilinguismo dei contenuti; b) la mancanza di restrizione di dominio nei contenuti stessi. (Panunzi, Fabbri, Moneglia, 2005). La tecnologia sviluppata, per essere adeguata alle richieste dell'utenza, non poteva quindi far ricorso a lessici o ontologie specifiche e doveva essere in grado di trattare l'informazione linguistica nelle principali lingue europee (Italiano, Francese, Tedesco, Inglese, Spagnolo). Si doveva inoltre assicurare una indicizzazione del contenuto sufficiente a consentirne l'identificazione, senza sfruttare le coordinate offerte dall'appartenenza dei contenuti ad uno specifico dominio (per definizione assenti in questo tipo di risorse). Il poster, accompagnato da una demo che estrae keyword da documenti in italiano e in inglese, descrive la struttura e gli algoritmi su cui è costruita lo strumento e i principali risultati relativi alla sua validazione. L'estrazione di keyword da testi è stata affrontata in letteratura attraverso due strategie tra loro interrelate: a) l'uso di tecnologie statistiche; b) l'uso di tecnologie e basi di dati linguistiche. Le tecnologie statistiche comparano i testi da cui le keyword debbono essere estratte con corpora di riferimento (Drouin, 2003). Le tecniche basate sullo sfruttamento dell'informazione linguistica si inseriscono in genere su tali risultati statistici. Per esempio, Van der Plas et al. (2004) utilizza basi di dati lessicali come WordNet or EDR. L'approccio qui proposto lavora su tre livelli. Al primo livello sfrutta dati di frequenza lessicale derivati da corpora di riferimento generali delle varie lingue implementate e l'annotazione linguistica fornita da un PoS tagger multilingue (TreeTagger, http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/). Ad un livello più sottile è utilizzata l'informazione semantica derivabile da basi di dati semantiche (WordNet, WordNetDomains), e infine sono sfruttati in modo nuovo concetti ben noti nella letteratura linguistica, come il concetto di collocazione, per aumentare la predittività semantica dell'indicizzazione. Il documento analizzato è etichettato sulle PoS attraverso TreeTagger. Tale procedura fornisce un'informazione linguistica di livello più accurato rispetto a quella prodotta dagli stemmer, e

permette l'estrazioni dei nomi presenti nel testo (considerati gli elementi lessicali che identificano primariamente un testo). La frequenza dei termini nel documento (term frequency, TF) è rapportata alla sua distribuzione del corpus di riferimento (BNC per l'inglese, un corpus di laboratorio per l'italiano), attraverso la misura della frequenza inversa del termine nella collezione di documenti (inverse document frequency, IDF). Con queste due misure viene calcolato il valore TF.IDF (Salton, 1989), alla base del punteggio di key-ness. L'algoritmo genera quindi una serie ordinata di nomi: si considerano parole identificative del contenuto di uno specifico documento quelle che sono più frequenti nel documento stesso e parallelamente meno diffuse nel corpus di riferimento. Al secondo stadio di analisi tali keyword, potenzialmente ambigue a livello semantico, sono disambiguate facendo ricorso al database WordNet (http://wordnet.princeton.edu/w3wn.html). La procedura di disambiguazione associa ad ogni parola dell'insieme di keyword estratte la serie dei suoi Synset in Wordnet e stima la similarità semantica tra tali Synset, producendo una disambiguazione attraverso la misura di distanza di Lesk (1986). Tale procedura, che sottostà ai limiti delle basi di dati utilizzate, ha due finalità: a) assicurare la traducibilità delle keyword; b) determinare il dominio semantico a cui fanno capo le keyword attraverso la loro associazione con i domini selezionati in WordNet Domains (http://wndomains.itc.it/wordnetdomains.html). La determinazione del dominio semantico generico del documento potrebbe portare, in linea di principio, ad un aumento significativo della predittività delle parole chiave dal punto di vista dell'utente ai fini dell'identificazione del contenuto. Le parole chiave che sono attestate in un documento, e che sono estratte con TF.IDF, non identificano infatti in se stesse il dominio a cui appartengono, che deve essere ricostruito dall'utente (se un documento parla di tigri e gazzelle, TF.IDF estrarrà come keyword la parola tigre e la parola gazzella, ma non darà l'informazione generale che si parla di "animali della savana"). Il terzo livello della procedura elaborata permette una più precisa caratterizzazione del contenuto attraverso l'ulteriore specificazione delle keyword identificate (concetti sotto-ordinati), piuttosto che per dominio di appartenenza (concetto sopra-ordinato). A questo livello, vengono prese in considerazione le associazioni lessicali individuate dalle collocazioni ad alta frequenza (Sinclair, 1991) presenti nel testo. L'idea sottostante è che se una keyword meglio specificata, aumenta la sua descrittività rispetto al contenuto del documento (tigre della Malesia, estinzione della tigre). La procedura estrae gli n-grammi dei termini presenti nel documento e scarta quelli linguisticamente non adeguati ad essere considerati descrittori (per es. le sequenze "pro + prep"). Il punteggio viene attribuito sulla base di una valutazione che tiene conto di: - frequenza dell'n-gramma all'interno del documento - frequenza relativa dell'n-gramma rispetto alle singole parole che lo compongono; - TF.IDF dei nomi all'interno dell'n-gramma; - un valore di normalizzazione, al fine di generare una lista unica che contenga sia le keyword singole che quelle complesse. Questa procedura, differentemente da altre presenti in letteratura (vedi Witten et al. 1999), non confronta gli n-grammi di un documento con quelli di un corpus di riferimento,

ma valuta solo quelli presenti all'interno del documento in esame. Infatti la finalità non è quella di studiare ed analizzare le associazioni lessicali che caratterizzano la lingua in cui è scritto il testo che viene elaborato, ma è quella di scoprire le associazioni lessicali interne al testo. L'algoritmo ha subito una valutazione che considera i risultati delle tre procedure in termini di precisione e li compara tra loro dal punto di vista dell'utente. Il valore aggiunto introdotto dalla disambiguazione e dall'individuazione dei domini dei documenti non produce risultati apprezzabili probabilmente a causa delle caratteristiche di Wordnet e Word-net Domain. Al contrario, i risultati dell'implementazione delle parole chiave attraverso le collocazioni sono molto soddisfacenti. Si è notato che l'adeguatezza complessiva delle serie di 5 parole chiave per l'identificazione del contenuto cresce di un fattore del 100% nel caso di keyword di più parole. Parallelamente le keyword di più parole, se comparate singolarmente con quelle singole, sono considerate meno vaghe e più adeguate all'identificazione del contenuto per un fattore del 30%.

Riferimenti bibliografici:

P. Drouin (2003) "Term Extraction Using Non-technical Corpora as a Point of Leverage", Terminology, 9(1), Benjamins, pp. 99-115.
Lesk, M. (1986). "Automatic Sense Disambiguation: How to Tell a Pine Cone from an Ice Cream". In Proceedings of the SIGDOC Conference. New York, NY: ACM, pp. 24-26.
Panunzi, A., Fabbri, M., Moneglia, M. (2005) "Keyword Extraction in Open-Domain Multilingual Textual Resources". In Proceedeing of 1st International Conference on Automated Production of Cross media Content for Multi-channel Distribution, IEEE Press.
Salton, G. (1989) "Automatic text processing: the transformation, analysis and retrieval of information by computer".
Sinclair, J. (1991) "Corpus Concordance Collocation". OUP.
Van der Plas, L., Pallotta, V., Rajman M., Ghorbel, H. (2004), "Automatic Keyword Extraction from Spoken Text" Proceeding of LREC 2004.
Witten, I., Paynter, G., Frank, E., Gutwin, C., Nevill-Manning, C. (1999) "KEA: Practical Automatic Keyphrase Extraction". In Proceedings of the Fourth ACM Conference on Digital Libraries. Berkeley, CA pp. 254-255.

**Andrea Sansò, Federica Da Milano, Caterina Mauri**
Università di Pavia, Dipartimento di Linguistica Teorica e Applicata

**La variazione linguistica in Europa attraverso un database tipologico. L'esperienza del Pavia Typological Database.**

I database tipologici sono strumenti di ricerca e di classificazione dei dati linguistici sempre più popolari nella comunità dei tipologi. Le ragioni di questa

popolarità risultano evidenti a chiunque si occupi di tipologia, se è vero, come osserva DAHL (2001) e come dimostrano esperienze recenti come il World Atlas of Language Structure (cfr. HASPELMATH ET AL. 2005), che lo studio della distribuzione areale dei tratti linguistici è fondamentale nella ricerca tipologica anche in assenza di aree linguistiche vere e proprie. Tuttavia, all'interesse crescente per queste risorse linguistiche non corrisponde ancora un interesse altrettanto significativo per la definizione di procedure condivise di immagazzinamento e annotazione dei dati tipologici. Diversamente da quanto avviene nel settore, assai più avanzato, dei corpora testuali, per i quali esistono pratiche di annotazione universali (su tutte, le linee guida proposte dal consorzio della Text Encoding Initiative, cfr. SPERBERG-MCQUEEN / BURNARD 2002), il creatore di database tipologici si trova a operare in relativa "solitudine" e deve affrontare e risolvere problemi di codifica dell'informazione linguistica adattando, laddove possibile, standard e procedure nati per la codifica di materiale testuale. Solo di recente ci si è posti il problema della rintracciabilità delle informazioni linguistiche raccolte nei database tipologici: presso l'Utrecht Institute of Linguistics (v., p.es., MONACHESI ET AL. 2002, DIMITRIADIS / MONACHESI 2002) è tuttora in corso un progetto volto alla creazione di un meta-database che raccoglie informazioni sui database tipologici esistenti, progetto che, evidentemente, opera "a valle", senza suggerire alcunché riguardo alla creazione di nuovi database. D'altro canto, la linguistica tipologica ha elaborato, sin dagli anni novanta, procedure sempre più sofisticate per la raccolta e la glossatura dei dati (cfr. LEHMANN ET AL. 1994, LEHMANN 2001, BI-CKEL ET AL. 2004, LEHMANN 2004, 2004b), adottate come norma nei maggiori progetti internazionali. Chi si accinge alla costruzione di un database tipologico non può non tener conto degli importanti suggerimenti che vengono dalla moderna prassi tipologica, e che contribuiscono a rendere più sopportabile la "solitudine" di cui si è detto. Alla luce di queste premesse, la nostra presentazione si propone un duplice scopo: (i) si cercherà, innanzitutto, attraverso l'illustrazione di alcuni case studies, di enucleare le caratteristiche di portabilità (nel senso di BIRD / SIMMONS 2003) che rendono pienamente utilizzabile dal linguista un database tipologico; (ii) in secondo luogo si porranno al centro della discussione alcune scelte teoriche e metodologiche adottate nel Pavia Typological Database (PTD; http://www.unipv.it/paviatyp), un'iniziativa nata nel 2003 nell'ambito del programma di ricerca FIRB "Europa e Mediterraneo dal punto di vista linguistico: Storia e prospettive" (coord. centrale Paolo Ramat). Il PTD (cfr. SANSÒ 2006 per una descrizione più tecnica) è un database costruito utilizzando le immense potenzialità del linguaggio di marcatura XML, e della famiglia di linguaggi ad esso correlati (XSL, Xpath, Xquery), che contiene, allo stato attuale, 5 moduli che trattano fenomeni morfosintattici diversi (le strategie di relativizzazione; i sintagmi nominali possessivi; i nomi d'azione; le costruzioni coordinate; gli elementi deittici – avverbi e dimostrativi), ma che sarà ulteriormente accresciuto con nuovi dati. Si cercherà di mostrare, in sintonia con la filosofia del congresso SLI, come il tipologo creatore di database debba assumere una mentalità che è propria di chi

opera nel paradigma della computazione, e si sottolineerà come questo paradigma rappresenti una sfida concettuale e non solo empirica per chi opera quotidianamente con le idiosincrasie e le specificità della variazione linguistica, oggetto di indagine d'elezione della linguistica tipologica.

Riferimenti bibliografici:

BICKEL ET AL. 2004 = B. BICKEL / B. COMRIE / M. HASPELMATH, The Leipzig glossing rules. Con-ventions for interlinear morpheme-by-morpheme glosses, Leipzig, Max-Planck-Institut für Evolutionäre Anthropologie, 2004 (www.eva.mpg.de/lingua/files/morpheme.html).
BIRD / SIMMONS 2003 = S. BIRD / G. SIMMONS, Seven Dimensions of Portability for Language Documentation and Description, «Language» 79:3, 2003, pp. 557-582.
DAHL 2001 = Ö. DAHL, Principles of Areal Typology, in Language typology and language universals: an international handbook, ed. by M. HASPELMATH / E. KÖNIG / W. ÖSTERREICHER / W. RAIBLE, Berlin-New York, 2001, pp. 1456-70.
DIMITRIADIS / MONACHESI 2002 = A. DIMITRIADIS / P. MONACHESI, Integrating different data types in a Typological Database System, in Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, 27 May – 2 June 2002.
HASPELMATH ET AL. 2005 = M. HASPELMATH / M.S. DRYER / D. GIL / B. COMRIE (EDS.), The World Atlas of Language Structure, Oxford, 2005.
LEHMANN ET AL. 1994 = CH. LEHMANN / D. BAKKER / Ö. DAHL / A. SIEWIERSKA, EUROTYP Guidelines, Strasbourg, European Science Foundation (EUROTYP Working Papers), 2nd ed., 1994.
LEHMANN 2001 = CH. LEHMANN, Language documentation: a program, in Aspects of typology and universals, ed. by W. BISANG, Berlin, 2001, pp. 83-97.
LEHMANN 2004a = CH. LEHMANN, Data in linguistics, «Linguistic Review» 21, 3-4, 2004.
LEHMANN 2004b = CH. LEHMANN, Interlinear morphemic glossing, in Morphologie. Ein internationales Handbuch zur Flexion und Wortbildung, ed. by CH. LEHMANN ET AL. 2. Halbband, Berlin-New York, 2004.
MONACHESI ET AL. 2002 = P. MONACHESI / A. DIMITRIADIS / R. GOEDEMANS / A.-M. MINEUR, A unified system for accessing typological databases, in Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas de Gran Canaria, Spain, 27 May – 2 June 2002, pp. 1029-1035.
SANSÒ 2006 = A. SANSÒ, Documenting variation across Europe: The Pavia Typological Database, in Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006), Genova, 21-27 May 2006.
SPERBERG-MCQUEEN / BURNARD 2002 = C.M. SPERBERG-MCQUEEN / L. BURNARD (eds.), TEI P4: Guidelines for Electronic Text Encoding and

Interchange, XML-compatible edition, The TEI Consortium, 2002 (http://www.tei-c.org/P4X/).

**Renata Savy e Marina Castagneto**
Università di Salerno, Università di Cagliari

**Funzioni Comunicative e Categorie d'Analisi Pragmatica: dal Testo Dialogico allo Schema XML e Viceversa.**

Premessa: dall'analisi all'annotazione dei dati linguistici In tempi più o meno recenti, sempre maggior fortuna hanno avuto gli studi linguistici basati sull'analisi di corpora di lingua (scritta o orale), grazie ai quali è possibile sostanziare osservazioni e teorie su dati quantitativi e considerazioni statistiche. Perché un insieme di dati costituisca un corpus, è necessario che esso sia in primo luogo rappresentato, organizzato e dotato di una struttura (Llisterri, 1997:1-2), attraverso un'operazione di 'codifica' che comprende vari stadi, tra cui quello dell'annotazione (mark-up) o etichettatura (tagging). La codifica richiede un'attenta analisi linguistica del materiale: tutti i tipi di rappresentazione di un corpus sono risultato di un'attività di interpretazione e classificazione dei dati (Gibbon et al., 1997:146). Le operazioni di tagging comportano, quindi, necessariamente alcuni problemi di definizione e identificazione di categorie che l'analisi tradizionale può trascurare o considerare in termini di gradatum. L'attribuzione di un'etichetta, infatti, presuppone una scelta precisa della categoria (o, come si dice in gergo, del metadato) che andrebbe definita in modo netto ed univoco, senza consentire ambiguità o sovrapposizioni tra categorie. Nella costituzione di un corpus, pertanto, ogni livello di annotazione porta con sé qualche problema di categorizzazione, e ciò vale soprattutto per quei settori della grammatica o della lingua meno strutturati (e perciò parzialmente trascurati dalla linguistica teorica), come ad esempio il livello della pragmatica del testo, in particolar modo del discorso parlato. Lo standard XML per l'annotazione dei dati linguistici Tra i formati di rappresentazione linguistica, si è imposto da tempo come standard in numerosi progetti nazionali e internazionali l'XML (Extensible Mark-up language), un metalinguaggio che descrive linguaggi di mark-up, definendo i tag (le etichette) e la struttura dei dati e metadati. Promosso dalla TEI (Text Encoding Initiative), l'uso di XML ha numerosi vantaggi computazionali (di portabilità, semplificazione e potenza descrittiva) facilita lo scambio delle risorse e soprattutto si presta molto bene ad annotazioni multilivello e alla generazione di database linguistici. Xml è adatto in particolar modo al tagging di livelli dotati di una buona 'struttura' e rappresentabili per componenti e categorie 'gerarchizzate'. E', ad esempio, utilizzato per le annotazioni sintattiche e per il PoS-tagging (annotazione morfosintattica, Part-of-Speech tagging). Gli schemi di annotazione pragmatica In numerosi progetti internazionali, sono stati sviluppati diversi schemi di annotazione pragmatica, il cui scopo è quello di

identificare la funzione pragmatica che ogni atto dialogico ha all'interno di uno specifico contesto comunicativo. Tra questi, si segnalano lo schema multidimensionale DAMSL (e la sua estensione SWBD-Damsl – http://www.colorado.eduy/ling/jurafsky/manual.august1.html), che riflette lo schema degli atti linguistici di Searle, Molti altri schemi, nati per scopi specifici (es. Verbmobil, Alparon, Flammia, Linlin, Coconut), sono invece monodimensionali; tra questi il più famoso è lo schema messo a punto per l'annotazione dei MapTask dialogues del HCRC-corpus di Edimburgo (http://www.hcrc.ed.ac.uk). Quest'ultimo, anche nella sua versione estesa (Ferrari&Castagneto,2003), è un sistema pensato per uno specifico dominio di applicazione, con una struttura piuttosto rigida e limitata, che ha ricevuto una codifica Xml sia nella versione inglese, che in alcune applicazioni all'italiano (Crocco et al 2003). Il nostro progetto Presenteremo e discuteremo una nostra recente proposta (Castagneto, Savy, De Leo, 2006) di un nuovo schema annotativo, messo a punto integrando il sistema MapTask con categorie provenienti dal DAMSL e con alcune innovazioni. Lo schema è stato costruito su alcuni dialoghi task-oriented meno strutturati del Map-Task, provenienti dal corpus CLIPS, e prevede etichette basate sullo schema di Transaction, Games e Moves del MapTask, rivisitate in una struttura multidimensionale. La novità della proposta consiste, infatti, nel tentativo di una rappresentazione gerarchica tra le categorie d'analisi che distingue tra: a) livelli di Mosse Autonome (come il Comment, Self-Talk, Ready, Interruption Open, TransactionBegin, End, TransactionClosure) che non sono condizionate da e non condizionano lo sviluppo semantico del dialogo; b) livelli di macrocategorie corrispondenti a sottoclassi di Apertura e Chiusura (Influencing, Question, Understanding, Answer) che codificano il contributo comunicativo primario dell'atto linguistico; c) livelli di categorie 'fini', corrispondenti a Mosse Terminali (come Explain, Check, Acknowledgement, Reply, ecc.), che realizzano una specifica funzione comunicativa legata al tipo di compito e allo sviluppo micro-testuale del dialogo. Tale struttura gerarchizzata è stata successivamente formalizzata in una DTD (Document Type Definition) per la codifica Xml, in cui i vari livelli sono stati trattati come "Elementi" del testo o "Attributi" degli elementi stessi. Se a monte dello schema sussiste, com'è ovvio, una riflessione linguistico-pragmatica sulla struttura del testo, la cui difficoltà consiste essenzialmente nella identificazione di categorie discrete per l'analisi di un continuum sfuggente come quello delle funzioni comunicative dell'atto dialogico; a valle dell'annotazione, ulteriori riflessioni scaturiscono dalla trasformazione di uno schema d'analisi libero in uno schema rigido di rappresentazione. La costruzione della DTD implica infatti scelte precise sullo statuto stesso delle categorie d'analisi (elemento o attributo), sul trattamento e sul tipo di attributi, nonché sul tipo di relazione esistente tra attributi di livello diverso (ad esempio, i rapporti di implicazione tra mosse terminali e macrocategorie). Obiettivi L'obiettivo generale del nostro sistema è quello di arrivare, attraverso l'annotazione di un corpus piuttosto ampio di testi, alla costruzione di un database strutturato sul quale validare le nostre ipotesi interpretative, sia in termini di categorie, che in termini di struttura.

Presenteremo, dunque, un primo screening di testi dialogici annotati, tutti della stessa tipologia (Test Differenze), ma appartenenti ad aree geografiche diverse e realizzati da parlanti diversi, allo scopo di misurare le percentuali di occorrenza di ciascuna categoria e degli attributi definiti nella DTD e l'omogeneità o variabilità delle condizioni di realizzazione. Il primo test consentirà di perfezionare, in stadi successivi, lo schema e renderlo quanto più generico e duttile possibile, per poter giungere, in futuro, ad un sistema di misurazione dei testi (di tipologie variate) in termini di categorie e/o attributi associati.

**Francesca Tini Brunozzi, Silvia Guazza, Enrico Zovato**
Loquendo S.p.A. Torino

**Atti illocutivi e segnali discorsivi. Un contributo linguistico a un sistema TTS verso la sintesi vocale espressiva.**

Questo lavoro si inserisce nell'ambito di un obbiettivo più ampio rivolto all'analisi, alla modellizzazione e alla riproduzione di stili di lettura marcati [1] (frasi interrogative, parlato emozionale, foci contrastivi, ecc.) per un sistema di sintesi vocale Text to Speech (TTS). In particolare, si fa qui riferimento al contributo linguistico che ha orientato i risvolti applicativi del cosiddetto "GildedTTS" per il TTS di Loquendo verso la riproduzione acustico-prosodica di determinate funzioni pragmatiche. Questo lavoro riprende il percorso tracciato per un un analogo progetto linguistico-testuale di frasi interrogative considerate, non più per il mero aspetto della selezione di unità acustico-prosodiche da una base dati acustica specializzata, ma soprattutto, per la funzione pragmatica di essere 'atti di domanda', essere cioè azioni linguistiche (ad es. domande) che, nel contesto del dialogo, elicitano altre azioni linguistiche (ad es. risposte). Entrambi gli orientamenti di ricerca fanno quindi riferimento allo sviluppo di applicazioni che utilizzino la sintesi vocale nell'interazione uomo-macchina. Per lo specifico della cosiddetta sintesi espressiva (sia stilistica, sia emozionale) [15] [16], il presupposto che ha orientato la proposta applicativa, è stata la scelta di ovviare alla questione della rappresentazione delle emozioni tout court per risolvere, allo stato attuale della ricerca sulle tecniche di manipolazione del segnale [18] [19], il problema della coloritura espressiva di sintesi neutra, senza tuttavia compromettere la naturalezza timbrica raggiunta con la tecnica Unit Selection [2]. Prendendo spunto da un lavoro [12] in cui è suggerito l'utilizzo di elementi espressivi paralinguistici che performino la variazione stilistica sul testo mediante opportuna selezione prosodica, si è quindi tentato un approccio ancora più drastico. Si è infatti temporaneamente sospeso l'obbiettivo della variazione stilistica sui testi sintetizzati, arricchendo però il reportorio acustico di elementi espressivi (expressive cues) [10] [17] da integrare ai messaggi di sintesi. Da qui la scelta di orientare l'espressività delle voci sintetiche all'ambito pragmatico, focalizzando l'attezione sugli atti illocutivi e sui segnali discorsivi a

partire dagli studi sui correlati acustico prosodici ad essi relativi [13] [14]. Per ogni lingua (dall'italiano il progetto testi è stato esteso ad altre lingue come Francese, Castigliano, Catalano, Spagnolo Sudamericano, Portoghese, Inglese, Tedesco, Olandese, Americano) è stato quindi progettato un repertorio di formule classificabili in atti linguistici (confermare, rifiutare, annunciare, proibire, augurare, ringraziare, scusarsi, ecc.) [11], segnali discorsivi [3] [4], eventi paralinguistici (risate, schiarite di voce, sospiri, ecc.), che diano quindi non solo una coloritura emotiva al testo sintetizzato, ma che contribuiscano, per mezzo della correlazione tra formula testuale e tratti soprasegmentali, a connotare in funzione pragmatica il significato stesso del messaggio [5] [6] [7] [8] [9].

Riferimenti bibliografici:

[1] Avesani C., Vayra M., 2000, "Costruzioni marcate e non-marcate in italiano: il ruolo dell'intonazione", X Giornate di Studio del GFS (Gruppo di Fonetica Sperimentale), Napoli, 2000, pp. 1-15.
[2] Balestri M., Pacchiotti A., Quazza S., Salza P., Sandri S., 1999, "Choose the Best to Modify the Least: a New Generation Concatenative Synthesis System", in Proceedings of EUROSPEECH 1999, Budapest, pp. 2291-2294.
[3] Bazzanella C., 1994, Le Facce del Parlare. Un Approccio Pragmatico all'Italiano parlato, La Nuova Italia, Firenze.
[4] Bazzanella C., 1999, "I segnali discorsivi tra parlato e scritto", in Dardano M., Pelo A., Stefinlongo A. (a cura di), Scritto e parlato. Metodi, testi e contesti, Casa editrice Aracne, Roma, pp. 79-97.
[5] Bazzanella C., 2002a, Sul dialogo. Contesti e forme di interazione verbale, Milano, Guerini e associati.
[6] Bazzanella C., 2002b, "Emozioni e contesto: cenni d'analisi", in Bazzanella C., Kobau P. (a cura di), Passioni, emozioni, affetti, McGraw-Hill, Milano, 2002, pp. 291-302.
[7] Bazzanella C., 2005, Linguistica e pragmatica del linguaggio. Roma-Bari, Laterza.
[8] Bazzanella C., i.c.s., "Segnali discorsivi e sviluppi conversazionali", in Albano Leoni F., Giordano R. (a cura di), Italiano parlato. Analisi di un dialogo, Napoli, Liguori.
[9] Bazzanella C., Bosco C., 2001, "Multimodalità e contesto", in Multimodalità e Multimedialità nella Comunicazione, Atti delle XI Giornate di Studio del Gruppo di Fonetica Sperimentale (A.I.A.), Padova, 29-30 novembre 1 dicembre 2000, Unipress, Padova, 2001, pp.69-74.
[10] Campbell N., 2002, "Towards a grammar of spoken language: incorporating paralinguistic information", in 7th ISCA International Conference on Spoken Language Processing, Denver, Colorado, USA, September 16-20, 2002.
[11] Cresti E., Moneglia, M. et alii., 2002, "The C-ORAL-ROM Project. New methods for spoken language archives in a multilingual romance corpus", in M. C. Rodriguez e C. Suarez Araujo (a cura di), Proceedings of III° International Conference on Language Resources and Evaluation Vol 1, pp. 2-10, ELRA, Paris.

[12] Eide E., Aaron A., Bakis R., Hamza W., Picheny M., Pitrelli J., 2004, A corpus-based approach to expressive speech synthesis, 5 th ISCA Speech Synthesis Workshop Carnegie Mellon University, Workshop Proceedings, Pittsburgh.
[13] Firenzuoli V., 2001, "Verso un nuovo approccio allo studio dell'intonazione a partire da corpora di parlato: esempi di profili intonativi di valore illocutivo dell'italiano", in Maraschio N., Poggi Salani T., (a cura di), 2001, Atti del XXXIV Congresso internazionale degli studi della SLI, 19/21 ottobre 2000 - Firenze, Bulzoni, Roma.
[14] Gili Fivela B., Bazzanella C., i.c.s., 2004, "Sviluppo emozionale ed allocutività. Analisi di un caso", in Emanuela Magno Caldognetto e Piero Cosi (a cura di) Atti del Convegno Nazionale del Gruppo di Studio sulla Comunicazione Parlata della SLI (GSCP), Padova, 2004.
[15] Magno Caldognetto E., 2002, "I correlati fonemici delle emozioni", in Bazzanella C., Kobau, P. (a cura di), Passioni, Emozioni, Affetti, McGraw-Hill, Milano, pp. 197-214.
[16] Magno Caldognetto E., Zmarich C., Ferrero F.E., "Indici acustici macroprosodici dello stato emotivo del parlante", in Atti del XXVI Congresso Nazionale di Acustica, 1998, pp. 263-268.
[17] Schulze M., "Substitution of paraverbal and nonverbal cues in the written medium of IRC", in Neumann B. (ed.), Dialogue Analysis and the Mass Media, Niemeyer, Tübingen, 1999.
[18] Zovato E., Pacchiotti A., Quazza S., Sandri S., "Analisi di una base dati di parlato emozionale", Atti delle XIV Giornate del GFS, Viterbo, 4-6 dicembre 2003.
[19] Zovato E., Pacchiotti A., Quazza S., Sandri S., "Towards emotional speech synthesis: a rule based approach", 5th ISCA Speech Synthesis Workshop Carnegie Mellon University, Workshop Proceedings, Pittsburgh USA, June 14-16 2004.

**Marco Tomatis**
Università di Torino

**SMORFIA: un analizzatore della morfologia verbale dell'italiano moderno per gli apprendenti di lingua italiana**

Introduzione Nel corso degli ultimi anni si è potuto assistere allo sviluppo, in numerosi istituti di ricerca europei, di un numero sempre crescente di sistemi per l'analisi morfologica. Tuttavia, nonostante la richiesta pressante di strumenti per l'apprendimento elettronico, la maggior parte di detti sistemi di analisi risulta progettato con lo scopo primario di coprire funzionalità interne in programmi di complessità maggiore. Infatti, al fine di ottimizzare il processo di analisi e fornire al sistema le informazioni flessionali strettamente necessarie, tali progetti hanno trovato la loro chiave di sviluppo seguendo un modello

descrittivo fondamentalmente di tipo "Item-and-Process". Sebbene una tale scelta si riveli vincente all'interno di applicazioni rivolte all'NLP (Natural Language Processing), purtroppo proprio a causa delle sue caratteristiche intrinseche essa mostra tutti i suoi limiti quando inserita in contesti di e-learning. Il problema di fondo degli analizzatori morfologici basati su un approccio "Item-and-Process", infatti, consiste nella perdita pressoché totale delle informazioni legate alla struttura del verbo oggetto di analisi. Il presente articolo ha come scopo quello di descrivere le principali caratteristiche di "SMORFIA", un analizzatore morfologico capace di mostrare all'utente l'intera struttura fonomorfologica dei verbi italiani. In merito al nome, SMORFIA rappresenta l'acronimo di "SMOR Finite-state Italian Analyser", in quanto il sistema sfrutta una tecnologia basata su trasduttori a stati finiti (FST). Più precisamente, l'analizzatore è stato implementato mediante l'utilizzo di SMOR, un formalismo sviluppato all'IMS (Institut für Maschinelle Sprachverarbeitung) di Stoccarda da Helmut Schmid. Per quanto riguarda il suo funzionamento, il programma agisce come un normale strumento per l'analisi morfologica: può accettare in ingresso sia interi documenti, sia singole parole introdotte tramite tastiera, permettendo inoltre il ridirezionamento dell'uscita sia su schermo, sia su un file specifico. Approccio metodologico L'approccio metodologico adottato durante lo sviluppo del sistema oggetto del presente articolo può essere considerato sotto certi aspetti innovativo in quanto si allontana da quelli generalmente implementati in progetti analoghi. E' un dato di fatto inconfutabile che volendo trattare mediante strumenti di apprendimento elettronico fenomeni legati al linguaggio, sia necessario formulare programmi capaci di fornire all'utente una descrizione il più fedele possibile della lingua che si intende descrivere. Pertanto, nel caso di una lingua morfologicamente complessa e articolata quale l'italiano, un sistema di analisi morfologica deve essere in grado di comunicare allo studente non solo i diversi valori che l'elemento flessionale può di volta in volta assumere, ma anche la struttura completa del verbo, correttamente suddivisa nei suoi costituenti principali. E' altresì pacifico che tali informazioni debbano essere mostrate sequenzialmente all'interno della stessa stringa di testo come coppie attributo-valore. Pertanto, al fine di riuscire ad ottenere una struttura così elaborata, una soluzione adeguata consiste nell'organizzare le regole di analisi del trasduttore a stati finiti in una maniera tale da garantire un approccio di tipo "Item-and-Arrangement" in luogo del classico "Item-and-Process". E' necessario ancora precisare che l'analisi morfologica condotta da SMORFIA si basa su un modello teorico che presuppone la divisione del verbo in due parti: la base e la flessione. A sua volta, nella maggior parte dei casi, la base può essere ulteriormente divisa negli elementi "radice" e "vocale tematica". Nel sistema descritto in questa sede, per vocale tematica non si intende unicamente quella particella presente all'interno delle tre principali coniugazioni dell'infinito, ossia "A re", "E re" e "I re", bensì anche quella appartenente a contesti differenti (es. participio passato "perd U to"). Poiché il progetto SMORFIA punta a costituire uno strumento funzionale per aiutare i discenti a comprendere meglio la complessità della morfologia verbale

italiana, il sistema è stato programmato per mostrare, laddove possibile, anche la struttura interna alla flessione. I seguenti esempi di analisi flessionale consentono di chiarire con maggiore efficacia la differenza che sussiste tra gli approcci "Item-and-Process" e "Item-and-Arrangement". Uscita di uno strumento di analisi progettato su un modello di tipo "Item-and-Process": parola = cacciassi risultato = cacciare(Cat V)(Mod Subj)(Tense Imperf)(Pers 1 2 sing) Uscita di SMORFIA (Item-and-Arrangement): parola = cacciassi risultato = cacciarecaccassi<1,2> Al fine di migliorare l'accuratezza dell'analisi morfologica, il programma in questione è in grado di gestire con modalità diverse quei caratteri appartenenti esclusivamente al livello ortografico, in quanto del tutto privi di valore fonematico. E' proprio per tale motivo che nell'esempio di cui sopra il grafema "i" di "cacciare", unicamente utilizzato per esprimere la palatalizzazione della "c" ad esso precedente, è stato inserito tra parentesi uncinate "< >". Organizzazione dei paradigmi L'intero sistema paradigmatico italiano è stato implementato all'interno del trasduttore dividendo i verbi in piccoli gruppi caratterizzati dalla condivisione dello stesso tipo di radice. Il numero complessivo di tali "parti di paradigma" raggiunge un totale di 190 unità, distribuite secondo il seguente schema: - 3 parti di paradigma sufficienti a coprire l'intera gamma dei verbi regolari presenti nelle tre coniugazioni. - 1a coniugazione: 18 parti per i verbi irregolari. 9 parti per i verbi difettivi. - 2a coniugazione: 82 parti per i verbi irregolari. 41 parti per i verbi difettivi. - 3a coniugazione: 29 parti per i verbi irregolari. 3 parti per i verbi difettivi. Sebbene tali valori possano suscitare non poche perplessità, è opportuno chiarire che l'origine di una simile ridondanza è principalmente dovuta a due fattori. Il primo di essi è legato alla scelta di mantenere il sistema il più vicino possibile alla reale struttura fonologica del verbo. Da ciò si deduce che l'analizzatore sarà costretto a utilizzare due parti di paradigma differenti al fine di mostrare l'esistenza di due radici foneticamente diverse, sebbene queste possano mostrarsi del tutto identiche da un punto di vista ortografico (es. "produc + e" c = /tS/ vs. "produc + o" c = /k/). Il secondo elemento di ridondanza, invece, è strettamente legato al comportamento assolutamente imprevedibile dei verbi difettivi, nonché dell'esistenza, sviluppata in particolare nella 2a coniugazione, di forme verbali caratterizzate da allotropia a livello del participio passato (es. "perduto" vs. "perso"). Sviluppi futuri Il sistema di analisi descritto nel presente articolo risulta tutt'ora in corso d'opera. Attualmente il programma è in grado di gestire l'analisi morfologica di oltre 11900 verbi differenti, tuttavia il trasduttore può essere ulteriormente migliorato al fine di garantire la completa analisi linguistica di tutti i restanti aspetti di natura flessionale e derivazionale caratteristici del lessico italiano.

Riferimenti bibliografici:

Aronoff, Mark. 1994. Morphology by itself. Stems and Inflectional Classes. Cambridge: MIT Press.
Koskenniemi, Kimmo. 1983. Two-level morphology: a general computational

model for word-form recognition and production. Publication No. 11. University of Helsinki: Department of General Linguistics.
Pirrelli, Vito. 2000. Paradigmi in morfologia: un approccio interdisciplinare alla flessione verbale dell'italiano. Istituti editoriali e poligrafici internazionali.
Schmid, Helmut; Fitschen, Arne & Heid, Ulrich. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. Proceedings of the IVth International Conference on Language Resources and Evaluation (LREC 2004), 1263-1266. Lisbon, Portugal.
Stump, Gregory T. 2001. Inflectional Morphology: A Theory of Paradigm Structure. Cambridge: Cambridge University Press.

**Sara Tonelli, Rodolfo Delmonte**
Università Ca' Foscari Venezia - Dipartimento di Scienze del Linguaggio

**Knowledge-poor and knowledge-rich approach in anaphora resolution algorithms: a comparison**

Anaphora resolution (AR) is a prominent topic in NLP, since in most application domains such as Question Answering, Text Summarization or Information Extraction there is an increasing need to collect and to implement semantically consistent information. It's been observed that anaphora extraction process should rely on a rule-based approach rather than on a statistical one, as anaphora (in our case pronoun binding) is a complex task involving functional and semantic aspects beside the syntactic ones. In our study we carried out an evaluation of three AR algorithms – GuiTAR, JavaRAP and MARS - based on a subpart of the Susanne Corpus. The results were then compared to the GETARUNS system developed at Dipartimento di Scienze del Linguaggio, Università Ca' Foscari, Venezia, by prof. Delmonte. The texts used in similar evaluations were usually portions of scientific manuals, thus they were poor on pronouns and rich on nominal description. Out testbed, on the contrary, could be better compared because it was a collection of texts taken from newspaper articles and stories, counting 35,000 tokens and about 1,000 third person pronominal expressions. Our aim was to highlight differences between the systems, but also to see if knowledge-poor approaches are suitable for AR, as they are much less labour-intensive and time-consuming than knowledge-rich systems. Two of the algorithms we evaluated – GuiTAR and JavaRap – use Charniak's statistical parser output, whereas MARS relies on a more sophisticated input provided by Connexor FDG parser (Tapanainen, Järvinen 1997). JavaRAP is based on the RAP algorithm (Lappin, Leass 1994) and identifies inter- and intrasentential antecedents of third person pronouns. Basically it uses the output of Charniak's parser to recover head-argument/head-adjunct relations and grammatical roles. Besides, it identifies pleonastic pronouns through lists of modal adjectives and cognitive verbs. JavaRAP assigns salience weights to antecedent candidates relying on several criteria such as

sentence recency, subject emphasis, distance from the current sentence. This allows to rank all the possible antecedents and to choose the best candidate. GuiTAR was developed by its authors (Poesio, M. and Mijail A. Kabadjov, 2004) to be as modular as possible so as to embed it into different NLP systems. The AR algorithm implemented in our evaluation was the one proposed by MARS, which will be discussed below. The application does not only identify personal pronouns but also (partially) definite descriptions. MARS is the system based on Ruslan Mitkov's anaphora resolution system using Connexor's output to extract candidate NPs and to check syntactic constraints. This algorithm seems to have the most original approach with a filtering module that computes preferential and impeding factors (a total of 14) such as Lexical Reiteration or Collocation Match to define the set of competing candidates. Several indicators are used to boost or reduce the candidates' scores. The NP with the highest score is then proposed as the antecedent. The first step in our evaluation was to manually compute all third person pronominal expressions including possessives, personals and reflexives. Apart from JavaRAP, none of the other systems including GETARUNS scored 100% coverage. As for accuracy, GuiTAR scored the best result with 54%, followed by JavaRap (50%) and MARS (43%): even if JavaRap and GuiTAR take the same parser input, they perform quite differently. The main problem with GuiTAR is that it fails systematically to recognize possessive adjectives, causing a limited coverage. On the other hand, MARS relies too heavily upon the concept of paragraph, as it doesn't perform anaphora resolution between elements in different paragraphs. Finally, JavaRAP seems to lack a consistent agreement check, especially for gender, which leads to very "basic" errors. These accuracy results suggest that knowledge-poor approaches are presently not suited for real-world tasks. For this reason, we decided to carry out a further comparison with a knowledge-rich AR module incorporated into the GETARUNS system. Differently from Connexor, the system elaborates both grammatical relations and semantic roles information for arguments and adjuncts. Besides, it proceeds in a "clause by clause" fashion, i.e. it tries first to resolve the anaphora locally and only if the resolution fails, it moves to a higher clause level. The system stores all entities in a push-down stack according to persistence principles; then it carries out the weighting procedure by taking into account several linguistic properties associated to each referring expression such as grammatical functions, semantic roles (framenet), animacy (according to 75 semantic features derived from Wordnet) and functional clause type. When trying to bound a personal pronoun, GETARUNS chooses the best antecedent by recursively trying to match features of the pronoun with the first available antecedent previously ranked by weighting. The procedure is repeated until all clauses are examined and all pronouns scrutinized and bound or left free. Pronouns left free (externals) will be matched tentatively with the best candidate provided by a centering-like algorithm. Accuracy overall results show that GETARUNS obtained the best score (62,7%), followed by JavaRap. For the GETARUNS system we also produced an evaluation of pronominal expressions in relation to their contribution at three different levels of AR: clause, utterance

and discourse level. We noticed that the highest percentage of pronouns found is at clause level, although the system performs better at discourse level. Compared to the other systems, there are at least three reasons why our system has the best performance: presence of both functional and semantic information; then the decision to treat utterance level pronominal expressions separately from discourse level ones, and finally the third reason is the way in which discourse level anaphora resolution is organized, with a centering algorithm hinging on a record of previously weighted best antecedents.

**Andrea Velardi e Alessio Plebe**
Università di Messina, Dipartimento di Scienze Cognitive

**Problemi teorici dei modelli: l'interazione tra approccio matematico e approccio cognitivo allo studio delle categorie semantiche.**

Uno dei campi in cui la linguistica computazionale si è sviluppata è senza dubbio quello dei modelli della memoria semantica cioè del sistema in cui è organizzata la conoscenza nella mente umana. Quasi tutti i modelli di questa specie operano su associazioni di parole e sui meccanismi della loro attivazione. I vari modelli sono stati costruiti ai fini della loro implementazione ma hanno cercato di fornire una spiegazione del funzionamento della mente (Velardi in corso di stampa). L'intervento vuole focalizzare due tipi di problemi contro cui deve misurarsi l'approccio computazionale: 1) I problemi di natura logico-sistemica connessi al trattamento delle categorie di parole per il quale si prenderà come esempio il noto problema della ereditarietà dei tratti o hyperonum problem (Levelt, Roelofs, Bock) e il trattamento dei concetti congiuntivi o complessi, per i quali non vale il principio logico della transitività (Hampton) 2) i problemi di natura più marcatamente psicologico-cognitiva connessi al trattamento delle molteplici variabili di matrice soggettiva e contestuale. Si discuteranno alcuni tentativi di produrre parametri computazionali adeguati al trattamento della variabile soggettiva come ad esempio le teorie dell'esemplare e specialmente uno dei suoi sviluppi: la teoria del contesto in cui viene formalizzata la attenzione selettiva dei soggetti (Medin,Smith,Nosofky). Questi temi sono approfonditi in Velardi (2005) dove si mostra come non ci sia tappa dello sviluppo dei modelli computazionali che non sia stata caratterizzata da una straordinaria interazione tra approccio teorico-qualitativo e approccio matematico-quantitativo. 1) Il problema dell' iperonimo (hyperonum problem) meglio detto problema della ereditarietà dei tratti è stata sollevato per la prima volta da Levelt (1989) ed è stato approfondito in seguito da Levelt, Roelofs, Meyer (1999), Bock (1995). Il problema dell'iperonimo emerge dall'ovvia constatazione che se una categoria iponima è contenuta in una categoria iperonima e questa categoria è a sua volta iponima di una categoria iperonima più grande allora le proprietà che definiscono la categoria maggiore saranno possedute certamente anche dalla categoria minore. Dato che la

categoria leone è iponima della categoria felino e quest'ultima categoria è iponima di ANIMALE, allora le proprietà della categoria superiore ANIMALE devono essere possedute necessariamente anche dalla categoria inferiore leone. A livello di elaborazione mentale questo vuol dire che i tratti della categoria superiore vengono trasportati in quella inferiore attraverso un processo simile a quello di vasi comunicanti. Le proprietà, o tratti, definitori vengono ereditate dalle categorie figlie o minori per il tramite delle categorie madri o superiori. Ma questo processo avviene anche quando parliamo? L'evidenza logica derivata dalla transitività categoriale non è così scontata a livello dei processi di elaborazione linguistica. Il problema dell'iperonimia ci dice che se l'attivazione dei tratti semantici di una parola provoca l'attivazione di tutte le altre che li contengono a vari livelli di astrazione, esso renderebbe troppo complicata e incerta la selezione di una voce lessicale specifica per un referente specifico. Bock (1995) mostra che se il passaggio dei tratti fosse automatico e meccanico l'attenzione della mente non potrebbe concentrarsi su una sola categoria. Essa dovrebbe per forza comunicare con la categoria superiore quando invece la selezione lessicale delle parole e la stessa idea di pertinenza semantica dimostrano che il soggetto non è obbligato a trasportare i tratti da una categoria superiore ad una inferiore come deve fare un calcolatore in preda ad un modello logico astratto. 2) Hampton (1993) ha mostrato che l'approccio prototipico non sia, al pari di quello classico, troppo astratto e troppo vuoto dal punto di vista teorico. Per quanto riguarda il primo dei problemi Hampton (1982) ha mostrato che i soggetti classificano classi di oggetti tipo car-seats o ski-lifts come sedie ma non sono disposti a classificarli come mobili. La categorizzazione espressa dalla asserzione: "Una sedia è un tipo di mobile" non è valida nel caso dei concetti complessi. Questo risultato è molto importante perché dimostra che la classificazione dei soggetti può essere di tipo intransitivo cioè contravvenire ad un noto principio logico quello appunto detto della transitività. Questa legge dice che: se A è uguale a B e B è uguale a C anche A è uguale a C. La regola non si applica solo al caso dell'equivalenza ma anche a quello dell'appartenenza o dell'inclusione : Se A contiene l'insieme B e 'insieme B appartiene a C allora C appartiene anche ad A. Nel caso delle car-seats e degli ski-lifts questa regola è infranta: Se la categoria mobile (A) contiene la categoria sedia (B) e la categoria sedia (B) contiene l'oggetto car seats (C) la car-seats ( C) non è inclusa nella categoria mobile (A). Hampton (1987,1988a, 1988b, 1991, 1993) ha mostrato che non solo gli approcci monotetici, ma anche la stessa logica fuzzy, sono inappropriate a spiegare la natura delle interazioni semantiche che intercorrono nei concetti congiuntivi. 3) Il problema della focalizzazione dei soggetti è presente nella teoria dell'esemplare e soprattutto nella versione che ne ha fornito Nosofsky nel suo modello generalizzato del contesto (Nosofsky 1984a; Nosofsky 1984b; Nosofsky 1985b; Nosofsky 1986; Nosofsky 1987; Nosofsky 1989; Nosofsky 1992 a; Nosofksy 1992 b; Nosofsky 1994; Nosofsky,Palmeri,McKinley 1994.Una sintesi in Murphy 2002 a). In questo approccio i soggetti immagazzinano nella memoria esemplari individuali delle categorie, con decisioni sulla classificazione basate sulla similarità dello stimolo con gli

esemplari immagazzinati (Medin e Schaffer 1978). Essi percepiscono gli esemplari in un continuum multidimensionale e con un' attenzione diseguale, soggettiva e contestualizzata, degli esemplari e dei loro attributi. Quest'ultima variabile viene inserita nel modello e calcolata come parametro di attenzione selettiva. Il processo di attenzione selettiva opera sulla rappresentazione percettiva che può condurre a cambiamenti sistematici nella struttura dello spazio psicologico e relativi cambiamenti nelle relazioni della similarità fra stimoli (Shepard 1964). L'attenzione selettiva è modellizzata da pesi differenziati delle dimensioni delle componenti nello spazio psicologico.

Riferimenti bibliografici:

Velardi A., Il nuovo paradigma. Categorie, prototipi e semantica cognitiva, EDAS,Messina, 2005.
Velardi A., Linguaggio e memoria, Bruno Mondadori, in corso di stampa.

**Alessandro Vietti**
Libera Università di Bolzano, Centro Ricerca Lingue

**Variabilità e grammatiche probabilistiche: il contributo di VARBRUL.**

La proposta avanzata da Labov negli anni 60-70 di interpretare in senso probabilistico le regole context sensitive della fonologia generativa di Chomsky, Halle (1968) fu generalmente considerata incompatibile con la nozione di competenza grammaticale ritenuta allora categorica e non tendenziale. Da allora il quadro metodologico (direi tecnologico) e teorico è molto mutato. Da un lato, la linguistica computazionale e dei corpora ha sviluppato modelli statistici della grammatica (proprio a partire dalle grammatiche formali della gerarchia chomskyana), mentre dall'altro, nell'ultimo decennio, all'interno della linguistica teorica si è andato consolidando un vasto ed eterogeneo paradigma funzionalista, orientato ai dati e all'uso (usage-based) che pone al centro la natura probabilistica dei fatti di lingua (Bybee, Hopper, 2001; Bod, Hay, Jannedy, 2003). L'idea di una competenza grammaticale all'interno della quale ci sia spazio anche per la frequenza è oggi molto più che un'ipotesi. In questo senso dunque i modelli probabilistici elaborati dal paradigma variazionista (americano), spogliati dal formalismo delle regole variabili, risultano oggi in sintonia con il contesto teorico e le attuali possibilità di trattare grandi quantità di dati. Per questo motivo VARBRUL, software nato negli anni 70 per l'analisi della variazione, sta conoscendo in questi ultimi anni una sorta di rinascita in ambito variazionista, anche grazie alle recenti versioni per Windows, GOLDVARB 2001 (Robinson, Lawrence, Tagliamonte, 2001) e GOLDVARB X (Sankoff, Tagliamonte, Smith, 2005; v. Paolillo, 2001). Il programma è per così dire purpose specific, cioè elaborato dal matematico canadese David Sankoff (Cedergren, Sankoff, 1974) con lo scopo di analizzare il comportamento di una

variabile linguistica categorica (dicotomica ma anche politomica) in relazione a più variabili indipendenti. La regressione logistica è la tecnica statistica utilizzata che consente di computare i diversi pesi esercitati dai fattori linguistici e/o socio-situazionali. Gli obiettivi della presentazione sono pertanto due: a) illustrare il funzionamento del programma attraverso esempi di indagini presenti in letteratura (p.e. Labov, 1969; Anttila, 1997), ovvero condotte da chi scrive (su situazioni di contatto linguistico); b) concentrarsi sulle possibili implicazioni che tali modelli probabilistici possono avere nello studio della variazione. Per il punto (a) si presenterà il processo di analisi dalla codifica dei dati a partire da un corpus fino all'esecuzione della regressione logistica semplice (1-level analysis). Particolare attenzione sarà dedicata alla possibilità di migliorare il modello attraverso la cosiddetta step-wise regression che consente di valutare la significatività di diversi modelli. La possibilità di recoding consente invece di modificare le condizioni di analisi assemblando delle macro-variabili, suddividendo variabili in sotto-variabili, eseguendo analisi su sottocampioni e così via. Per quel che riguarda il punto (b) si forniranno in primo luogo alcune ragioni per l'utilizzo di VARBRUL al posto dei più consueti pacchetti generali di statistica come SPSS: è infatti un software gratuito e piuttosto semplice da imparare ad utilizzare, permettendo anche di valutare con facilità la bontà dei vari modelli. In secondo luogo, si proporranno alcune questioni che sembrano sorgere dai risultati di simili analisi e riguardanti per così dire il rapporto con il cuore di un teoria della variazione, ovvero il locus e il grado della variazione. In particolare si cercherà di capire (facendo ricorso alle analisi presentate) quali siano i rapporti tra i cosiddetti fattori interni (linguistici) ed esterni (genericamente sociali) della variazione. Concludendo si indicheranno alcune possibili prospettive di integrazione del variazionismo in una linguistica probabilistica non necessariamente formalista (v. per la fonologia l'utile rassegna di Pierrehumbert, 2001).

Riferimenti bibliografici:

Anttila A., 1992, "Deriving variation from grammar: A study of Finnish genitives". In: Hinskens F. / van Hout R. / Wetzels L. (eds.), Variation, Change, and Phonological Theory, Amsterdam / Philadelphia, Benjamins: 35-68.
Bod R. / Hay J. / Jannedy S. (eds.), 2003, Probabilistic Linguistics, Cambridge Mass., MIT Press.
Bybee J. / Hopper P. (eds.), 2001, Frequency and the Emergence of Linguistic Structure, Amsterdam / Philadelphia, Benjamins.
Cedergren H. / Sankoff D. 1974, "Variable rules: Performance as a statistical reflection of competence". Language 50/2: 333-355.
Chomsky N. / Halle M., 1968, The Sound Pattern of English, New York, Harper and Row.
Labov W., 1972, "The social stratification of (r) in New York City department stores". In: Linn, Michael D. (ed.), 1998, Handbook of Dialects and Language Variation, San Diego, Academic Press: 221-244.

Paolillo J.C., 2002, Analyzing Linguistic Variation. Statistical Models and Methods, Stanford, CSLI Publications.
Pierrehumbert J., 2001, "Stochastic Phonology". Glot International 5/6: 195-207.
Robinson J. / Lawrence H. / Tagliamonte S., 2001, GOLDVARB 2001. A Multivariate Analysis Application for Windows,
http://www.york.ac.uk/depts/lang/webstuff/goldvarb/manual/manualOct2001.html
Sankoff D. / Tagliamonte S. / Smith E., 2005, GOLDVARB X. A Multivariate Analysis Application,
http://individual.utoronto.ca/tagliamonte/Goldvarb/GV_index.htm

# CALENDARIO DELLE MANIFESTAZIONI LINGUISTICHE

a cura di Emanuele Banfi

## 2006

### Giugno 2006

1-3 / Beograd

INTEX/NooJ Workshop, 9th. Belgrade, Serbia & Montenegro. Abstract deadline: 15 March 2006. Informazioni: max.silberztein@univ-fcomte.fr; http://nooj.matf.bg.ac.yu

5-6 / Tokyo

Logic & Engineering of Natural Language Semantics 2006 LENS 2006. Tokyo, Japan. Abstract deadline: 15 February 2006. Informazioni: mccready@lang.osaka-u.ac.jp; http://www.lang.osaka-u.ac.jp/~ogata/LENLS2006.html

5-7 / Honolulu

Pragmatics in the Chinese,Japnese, Korean Classroom: The State of the Art. Honolulu, HI. Informazioni: nrcea@hawaii.edu; http://www.hawaii.edu/nrcea/CJKCallforPapers.htm.

7-10 / Nijmegen

Speech Motor Control, 5th SMC. Nijmegen, Netherlands. Informazioni: infor@slp-nijmegen.nl; infor@slp-nijmegen.nl/smc2006/

8-10 / Trieste

Diachronic Generative Syntax, 9th DiGS 9. Trieste, Italy. Informazioni: digs9@units.it; http://www.units.it/~digs9

8-10 / Oslo

Explicit & Implicit Information in Text: Information Structure across Languages. Oslo, Norway. Informazioni: m.f.krave@ilos.uio.no; http://www.hf.uio.no/forskningsprosjekter/sprik/english/activities/index.html

8-10 / Amsterdam

Universality & Particularity in Parts-of-Speech Systems PoS2006. Amsterdam, Netherlands. Informazioni: PoS Cte, c/o Gerdien Kerssies, Dept Ling, U

Amsterdam, Spuistraat 210, 1012 VT Amsterdam, Netherlands; pos-fgw@uva.nl; http://home.hum.uva.nl/pos

9-10 / Birmingham

Revisiting Advanced Varieties in L2 Learning. Birmingham, UK. Abstract deadline: 31 January 2006. Informazioni: e.labeau@aston.ac.uk

10-11 / Tokyo

Japanese Society for Language Sciences, 8th JSLS2006. Tokyo, Japan. Abstract deadline: 20 January 2006. Informazioni: kei@aya.yale.edu; http://www.cyber.sccs.chukyo-u.ac.jp/JSLS/JSLS2006

14-17 / Firenze

IX Convegno internazionale della SILFI: "Prospettive nello studio del lessico italiano".
Informazioni: prof.ssa Emanuela Cresti, Dipartimento di Italianistica, Università degli Studi, piazza Savonarola 1, 50132 Firenze.
Informazioni: E-mail: elicresti@unifi.it; silfi2006@gmail.com

15-17 / Budapest

Temporal Representation & Reasoning TIME 2006. Budapest, Hungary. Abstract deadline: 30 January 2006. Informazioni: jamesp@cs.brandeis.edu; http://time2006.org

17-20 / Montreal

American Association for Applied Linguistics/Association Canadienne de Linguistique Appliquée/Canadian Association of Applied Linguistics AAAL/ACLA/CAAL. Montreal, PQ, Canada. Informazioni: AAAL, 3416 Primm Ln, Birmingham, AL 35216; 205-824-7700; toll free: 866-821-7700; fax: 205-823-2760; aaal@primemanagement.net

23-24 / Roma

Università di Roma Tre: Giornate su "La struttura dell'informazione". Informazioni: Prof.ssa Lunella Mereu mereu@uniroma3.it; Prof. Giorgio Banti gbanti@unior.it.

23-25 / Bellingham

North American Conference on Chinese Linguistics, 18th NACCL18.

Bellingham, WA. Informazioni: naccl18@wwu.edu; http://www.wwu.edu/naccl18

27-30 / Tomsk

Languages of Europe & North & Central Asia, 3rd LENCA-3. Tomsk, Russia. Theme: Grammar & Pragmatics of Complex Sentences. Informazioni: tomsk@eva.mpg.de

28-30 / Coimbra

Associacao de Crioulos de Base Lexical Portuguesa e Espanhola. Coimbra, Portugal. Informazioni: tjerk@netcabo.pt; http://www.umac.mo/fsh/dp/acblpe/.

**Luglio 2006**

1-4 / Paris

Teaching & Language Corpora, 7th TaLC2006. Paris, France. Abstract deadline: 20 February 2006. Informazioni: talc7@eila.jussieu.fr; http://talc7@eila.jussieu.fr

2-3 / Jerusalem

Israel Association for Theoretical Linguistics, 22nd IATL 22. Jerusalem, Israel. Abstract deadline: 5 March 2006. Informazioni: Ariel Cohen, Dept For Lits & Ling, Ben-Gurion U Negev, Beer Sheva 84105, Israel; arikc@bgu.ac.il

4-6 / Brisbane

Pacific Second Language Research Forum, 5th PacSLRF06. Brisbane, QLD, Australia. Abstract deadline: 15 January 2006. Informazioni: PacSLRF06, EMSAH, Michie Level 4, U Queensland, Brisbane, QLD 4072, Australia; pacslrf06@uq.edu.au; http://www.emsah.uq.edu.au/pacslrf06/

5-9 / Paris

Digital Humanities 2006. Paris, France. Informazioni: lisa.lena.opas-hanninen@oulu.fi; http://www.allc-ach2006.colloques.paris-sorbonne.fr/

10-14 / Albi

Digital Documents & Interpretation: Corpora in Humanities & Social Sciences. Albi, France. Informazioni: LPE2@ext.jussieu.fr

10-14 / Jalisco

International Circle of Korean Linguistics, 15th ICKL 15. Jalisco, Mexico. Informazioni: http://www.ickl.or.kr

13-15 / Minneapolis

Society for Text & Discourse, 16th ST&D. Minneapolis, MN. Abstract deadline: 1 February 2006. Informazioni: ST&DO6, Dept Psych, U MN, 75 E River Rd, Minneapolis, MN 55455; info@std06.org; http://std06.org/index.html

17-20 / Paris

Language, Culture, & Mind 2 LCM 2006. Paris, France. Informazioni: lassegue@lcm2006.net; http://ww.lcm2006.net

17-21 / Sevilla

Advances in Native South American Historical Linguistics LING 19. Sevilla, Andalucia, Spain. Informazioni: peviegas@hotmail.com; http://www.52ica.com

17-21 / Sydney

Computational Linguistics, 21st/Association for Computational Linguistics, 44th COLING/ACL 2006. Sydney, Australia. Abstract deadline: 28 February 2006. Informazioni: tim@csse.unimelb.edu.au; http://www.acl2006.org

17-21 / Sevilla

The Languages of Central America Caribbean Coast: Articulating Society, Culture in the Present, Past, & Future. Sevilla, Spain. Informazioni: icacaribe@purdue.edu; http://www.personal.us.es/tutatis/521CA

## Agosto 2006

5-6 / New York

Japanese Language Education 2006. New York, NY. Informazioni: icjle@yahoo.co.jp; http://www.japaneseteaching.org/icjle

9-12 / Seoul

Seoul International Conference on Generative Grammar, 8th SICOGG 8. Seoul, S Korea. Abstract deadline: 20 March 2006. Informazioni: parkmk@dgu.edu; http://www.kggc.org

9-13 / Turku

Organization in Discourse 3: The Interactional Perspective OID3. Turku, Finland. Informazioni: OID3 Conf, c/o Dept Engl, U Turku, FIN-20014 Turku, Finland; oid3@utu.fi; http://www.hum.utu.fi/engfil/oid3.html

17-19 / Osaka

Formal Approaches to Japanese Linguistics, 4th FAJL4. Osaka, Japan. Abstract deadline: 1 March 2006. Informazioni: fajl4@lang.osaka-u.ac.jp; http://www2005.lang.osaka-u.ac.jp/~fajl4/

17-27 / Beijing and Hong Kong

Applied Linguistics & Language Teaching, 7th. Beijing & Hong Kong, PRC. Abstract deadline: 31 May 2006. Informazioni: kevin@buaa.edu.cn; http://www.isallt2006.org

21-25 / Bergamo

English Historical Linguistics, 14th 14 ICEHL. Bergamo, Italy. Informazioni: 14icehl@unibg.it; http://www.unibg.it/anglistica/slin/14icehl-home.html

24-26 / Bolzano-Bozen

Multilingualism across Europe: Findings, Needs, Best Practices. Bolzano, Italy. Abstract deadline: 28 February 2006. Informazioni: mputz@eurac.edu; http://www.eurac.edu/Org/LanguageLaw/Multilingualism/Projects/tag06_1.profil.htm

39th Annual Meeting of the SLE. University of Bremen: "Relativism and Universalism in Linguistics".
Informazioni: www.fb10.uni-bremen.de/sle2006

**Settembre 2006**

1-3 / Tokyo

Construction Grammar, 4th ICCG4. Tokyo, Japan. Informazioni: iccg4komaba@ecs.c.u-tokyo.ac.jp; http://gamp.c.u-tokyo.ac.jp/~iccg2006/iccg2006.html

5-6 / Sendai

Strength Relations in Phonology. Sendai, Japan. Abstract deadline: 16 April

2006. Informazioni: rprg@tscc.tohoku-gakuin.ac.jp; http://www.tscc.tohoku-gakuin.ac.jp/~rprg

5-7 / Bristol

Association for French Language Studies AFLS. Bristol, UK. Theme: Variations. Abstract deadline: 31 January 2006. Informazioni: Kate.Beeching@uwe.ac.uk; http://www.afls.net/Conference2006.html

6-8 / Mannheim

Linguistic Colloquium, 41st. Mannheim, Germany. Abstract deadline: 28 February 2006. Informazioni: 41.lingkoll@uni-mannheim.de; http://www.phil.uni-mannheim.de/lingkoll06/

6-9 / Torino

EURALEX International Congress, 12th EURALEX 2006. Torino, Italy. Informazioni: http://www.euralex2006.unito.it

6-9 / Oxford

8th International Conference on Late and Vulgar Latin. Informazioni: roger.wright@liv.ac.uk

7-9 / Berlin

Typology of Tone & Intonation TTI. Berlin, Germany. Abstract deadline: 1 May 2006. Informazioni: downing@zas.gwz-berlin.de

7-9 / Utrecht

Romance Turn II. Utrecht University. Information: www.let.uu.nl/romanceturn/

8-10 / Toronto

Laboratory Approaches to Spanish Phonology, 3rd LASP3. Toronto, ON, Canada. Abstract deadline: 17 February 2006. Informazioni: lasp@utoronto.ca; http://www.chass.utoronto.ca/spanish_portuguese/phonology/

11-13 / Potsdam

Brandial 06: Semantics & Pragmatics of Dialogue, 10th Brandial06: Semdial 10. Potsdam, Germany. Abstract deadline: 12 May 2006. Informazioni:

das@ling.uni-potsdam.de; http://www.ling.uni-potsdam.de/brandial
13-16 / Antalya

European Second Language Association, 16th EUROSLA 2006. Antalya, Turkey. Abstract deadline: 31 January 2006. Informazioni: eurosla2006@boun.edu.tr; http://www.eurosla2006.boun.edu.tr

14-16 / Zaragoza

European Association of Languages for Specific Purposes, 5th AELFE. Zaragoza, Spain. Abstract deadline: 15 March 2006. Informazioni: llantada@unizar.es; http://www.unizar.es/aelfe2006/

14-16 / Barcelona

Forensic Linguistics/Language & the Law, 2nd IAFL European Conference. Barcelona, Spain. Abstract deadline: 31 January 2006. Informazioni: forensiclab@upf.edu; http://www.iula.upf.edu/agenda/iafl_bcn_06/iafl01uk.htm

14-16 / Praha

Romani Languages, 7th 7ICRL. Prague, Czech Rep. Abstract deadline: 20 May 2006. Informazioni: 7icrl@email.cz; http://ulug.ff.cuni.cz/7icrl/index.php

21-22 / Nijmegen

Workshop on Writing Systems, 5th. Nijmegen, Netherlands. Theme: Constraints on Spelling Changes. Abstract deadline: 1 May 2006. Informazioni: a.neijt@let.ru.nl; http://www.ru.nl/WrittenLanguage

21-23 / Vercelli

Linguistica e modelli tecnologici di ricerca, XL Congresso della Società di Linguistica Italiana. Informazioni: www.lett.unipmn.it/sli2006/

21-23 / Warsaw

Communicating across Age Groups: Age, Language, & Society GlobE 2006. Warsaw, Poland. Abstract deadline: 31 March 2006. Informazioni: globe@ils.uw.edu.pl; http://globe.ils.uw.edu.pl

26-28 / Regensburg

Anglicisms in Europe AiE. Regensburg, Germany. Informazioni: AiE@sprachlit.uni-regensburg.de; http://www.AiE2006.uni-regensburg.de

**Ottobre 2006**

8-10 / Valencia

Gender & Language Association, 4th IGALA-4. Valencia, Spain. Abstract deadline: 15 July 2006. Informazioni: IGALA 4, Dept Filol Angl & Alemanya, U Valencia, Av Blasco Ibanez 32-6, Valencia 46010, Spain; 34-96-386-42-62; fax: 34-96-386-41-61; jose.santaemilia@uv.es; http://www.uv.es/~santaemj/

12-14 / Bologna

Congresso della ASLI: „Storia della lingua e storia del teatro. L'italiano in scena".
Informazioni: Prof. Fabrizio Frasnedi: fabrizio.frasnedi@unibo.it

13-15 / Berlin

Kolloquium am Institut für Romanistik der Humboldt-Universität: "Rumanian und Europa. Transversale – für einen neuen Diskurs des Anschluss".
Informazioni: michele.mattusch@romanistik.hu-berlin.de

26-28 / Pisa

XXXI Convegno annuale della Società Italiana di Glottologia.
Categorie del verbo. Diacronia, Teoria, Tipologia
Informazioni: www.unimc.it/sig/convegni.htm

**Novembre 2006**

16-19 / Washington, DC
American Association for the Advancement of Slavic Studies, 39° AAASS.
Informazioni: walker@fas.harvard.edu

**2007**

**Agosto 2007**

5-11 / Montreal

XVIIIth International Conference on Historical Linguistics / XVIIIe Colloqui international de Linguistique historique.
Informazioni: listproc@uqam.ca

## Settembre 2007

3-8 / Innsbruck

Università di Innsbruck: XXV Congrès International de Linguistique et Philologie Romanes CILPR 2007
Info: http://www.cilpr2007.at
Alcune delle informazioni sono dovute alla cortesia di Hermann W. Haller, Iørn Korzen, Klaus Müllner e Elisabetta Jezek. A loro il grazie di tutta la SLI.

I soci sono invitati ad inviare informazioni per questa rubrica a Emanuele Banfi, Facoltà di Scienze della Formazione, Università degli Studi di Milano-Bicocca, P.zza dell'Ateneo Nuovo 1, 20126 Milano. Telefono: 02-64484817 / Fax: 02-64486995.
E-mail: emanuele.banfi@unimib.it

# PUBBLICAZIONI DEI SOCI

a cura di Emanuele Banfi

Csaba Földes, Szergej Tóth, Ágota Fóris (eds.), *Lexikológiai, lexikográfiai látkép: problémák, paradigmák, perspektívák.* (Fasciculi Linguistici Series Lexicographica 3. Szeged:Generalia (http://www.lib.jgytf.uszeged.hu/alknyelv, e-mail: alknyelv@lib.jgytf.u-szeged.hu), 2004, pp. 276, HU ISSN 1416-8081, ISBN 963 9167 85 1.

Il volume raccoglie i trentatré contributi della sezione di Lessicologia e Lessicografia dei due Congressi di Linguistica Applicata Ungherese tenuti a Pécs nel 2001, e a Szeged nel 2002. La rilevanza del volume è che la Lessicografia in Ungheria negli ultimi anni ha avuto uno sviluppo enorme. Il volume permette di conoscere i risultati ed i problemi emersi nel recente dibattito tra lessicografi ungheresi.

Ágota Fóris, Miklós Pálfy (eds.), *A lexikográfia Magyarországon.* Budapest: Tinta Kiadó (http://www.tintakiado.hu, e-mail: info@tintakiado.hu), 2004, pp. 197, 2100 Ft, HU ISSN 1419-6603, ISBN 963 7094 14 8.

La prima parte del volume contiene sette articoli che affrontano la situazione panoramica della lessicografia in Ungheria: problemi teorici e pratici dei vocabolari specializzati, a proposito del *Vocabolario tecnico-scientifico ungherese-italiano* (Á. Fóris), sulla grandezza dei vocabolari (Cs. Földes), sulla lemmatizzazione del grande vocabolario ungherese – *Nagyszótár* – (G. Kiss), la relazione tra lessicologia e lessicografia (T. Magay), sulla tipologia dei vocabolari (M. Pálfy), cambiamenti nella struttura dei vocabolari elettronici (G. Prószéky), e la denominazione delle piante prima del secolo XVII, a proposito del *Lexicon nominum herbarum, arborum fructicumque linguae Latinae I-IV* (J. Stirling).

Altri otto articoli del volume presentano i lavori lessicografici dei partecipanti al programma del Dottorato di Linguistica Applicata dell'Università degli Studi di Pécs (M. Balaskó, A. Kordásné Kenesei, Z. Máté, A. Németh, Gy. Pomázi, Cs. Szabó, T. Wallendums, A. Zrínyi).

Lunella Mereu, *La sintassi delle lingue del mondo,* Roma-Bari, Editori Laterza, 2004, 200 pp., 20 Euro.

Nata da un lavoro di sintesi dei risultati ottenuti da due tra le principali correnti della linguistica del 900, la grammatica generativa e la tipologia, una nuova proposta di studio della sintassi sulla base della comparazione di lingue diverse, che contempla insieme comportamenti linguistici e aspetti teorici e consente di ripensare la natura multidimensionale del linguaggio.

Lunella Mereu e Edoardo Lombardi Vallauri (a cura di), *What do we speak about when we speak of linguistics?* in "The Linguistic Review" 21.3-4, Berlin - New York, Mouton de Gruyter, 2004, pp.171-433, 42 USA $.

Il volume contiene gli Atti di un colloquio internazionale organizzato a Roma Tre nel luglio 2002 sullo statuto attuale della linguistica come scienza. Il volume include 10 saggi intesi come risposte a gruppi di quesiti organizzati intorno alle seguenti tematiche: I. il problema dei dati (Christian Lehmann – Lunella Mereu); II. metodi e formalizzazioni (Raffaele Simone); III. categorie e nozioni operative (Claude Hagège – Annarita Puglielli); IV. il problema della variazione (Gaetano Berruto); V. il linguaggio, la mente e le altre scienze (Luigi Rizzi – Edoardo Lombardi Vallauri); VI. I fondamentali della linguistica (Gilbert Lazare – Paolo Ramat).

Angela Ferrari (a c. di) 2004, *La lingua nel testo, il testo nella lingua,* Torino, Istituto dell'Atlante Linguistico Italiano, pp. 219.

I contributi raccolti affrontano lo studio dell'interazione tra la dimensione linguistica e la dimensione testuale nel discorso scritto, osservandola dal punto di vista della microsintassi (Angela Ferrari, Magda Mandelli, Luciano Zampese), della punteggiatura (Luca Cignetti, Letizia Lala) e del lessico (Anna-Maria De Cesare). L'obiettivo a cui tendono i lavori - definendone taglio, organizzazione e tono - è duplice. Data una forma linguistica, si tratta di vedere, paragonandola a possibili alternative espressive, quale sia il contributo che essa offre alla costituzione delle architetture testuali; e, attraverso questa analisi, di valutare in che misura e in che modo essa codifichi questo stesso contributo. Per molta parte del loro contenuto, gli studi proposti in questa miscellanea si pongono dunque idealmente nell'ambito della cosiddetta "pragmatica integrata", delimitando più precisamente uno spazio a cui si potrebbe dare il nome di "testualità integrata". Si osserva in tutti i suoi aspetti "la lingua nel testo", e si definisce in che modo e in che misura "il testo" sia già prefigurato nella "lingua".

Carla Bazzanella, *Linguistica e pragmatica del linguaggio*, Roma-Bari: Laterza, 2005, pp. 265, Euro 28, ISBN 88-420-7519-1.

L'obiettivo specifico di questo manuale (utilizzabile sia per un primo livello introduttivo, sia per approfondimenti) è presentare una prospettiva *pragmatica* della linguistica. Più specificamente, nella prima parte del volume si introducono i problemi relativi alle proprietà e funzioni della lingua, alle lingue storiche, alle dimensioni di variazione, alla problematica del discreto e continuo, ai livelli della lingua (fino alla nozione di 'grammatica emergente'), agli strumenti d'analisi. Nella seconda parte si tratta la nascita e lo sviluppo della

pragmatica, le complesse tematiche relative a contesto e deissi, le diverse teorie degli atti linguistici, la proposta teorica di Grice e gli sviluppi successivi, per concludere con una presentazione delle prospettive di analisi e strategie specifiche dell'interazione verbale, secondo un approccio complesso che tiene conto delle diverse componenti in gioco.

Michele Loporcaro, *Sintassi comparata dell'accordo participiale romanzo*, Torino: Rosenberg & Sellier, 1998, pp. VIII, 272, Lire 65.000.

Il volume si occupa di un problema di sintassi romanza (l'accordo del participio passato nei tempi composti) mettendo a frutto il ricchissimo materiale dialettale italoromanzo e proponendo una nuova sistematizzazione teorica che dà ragione della precisa corrispondenza tra acquisizione, variazione geografica e sviluppo diacronico del fenomeno.

Michele Loporcaro, *Cattive notizie. La retorica senza lumi dei mass-media italiani,* Milano, Feltrinelli, 2005, pp. 224, Euro 14.

Il volume analizza con gli strumenti del linguista come parla l'informazione dei mass-media italiani e mostra come questa, agli antipodi rispetto all'ideale illuministico dell'informazione come quarto potere, sia fatta, strutturalmente, in modo da anestetizzare, anzi, da prevenire il formarsi di una pubblica opinione.

Marina Castagneto, *Chiacchierare, bisbigliare, litigare... in turco: il complesso intreccio tra attività linguistiche, iconismo,* Cagliari, Arxiu de Tradicions de l'Alguer (via Carbonazzi, 17, 09123 Cagliari, arxiudetradicions@libero.it), 2004, pp. 202 [s.i.p.].

Nel libro vengono esaminati i complessi rapporti tra il lessico che designa il parlare quotidiano e la particolare struttura morfologica, a base reduplicativa, di questa componente del lessico, nel turco e non solo. Questo tipo di attività linguistiche, infatti, pertinentizza linguisticamente il rumore prodotto dalla fonazione durante il suo svolgimento. La base iconica delle parole interessate si rivela tanto nella presenza di un sistema fonosimbolico, tanto nella loro struttura morfologica reduplicativa (si tratta, in questo caso, di una iconicità diagrammatica). Nel libro viene proposta un'analisi morfo-semantica dei tipi di reduplicazione riscontrati nella lingua turca: ad ogni tipo di reduplicazione riscontrato (red. di una base monosillabica, di una base bisillabica, con apofonia vocalica, con apofonia consonantica) si associa senza eccezioni un tipo di significazione. Anche il turco, lingua agglutinante e derivativa per eccellenza, dimostra dunque di avere un sottosistema che si organizza su basi morfo-sintattiche completamente diverse, dove l'iconicità vince sulla arbitrarietà della grammatica.

Marco Biffi, Omar Calabrese, Luciana Salibra (a cura di), *Italia linguistica: discorsi di scritto e di parlato. Nuovi studi di linguistica italiana per Giovanni Nencioni,* Siena, Protagon, 2005, pp.334, euro 23,00.

Il volume raccoglie contributi di allievi più giovani e meno giovani, che nella loro eterogeneità - storia della lingua, analisi stilistica, fonetica, sintassi, linguistica teorica, analisi grammaticali -, testimoniano, oltre che gli interessi specifici di ciascuno studioso in questo momento della sua ricerca, anche l'impronta di un insegnamento genialmente versatile, spesso pionieristico, sempre di largo respiro.

Marina Cecchini (a cura di), Fa*re, conoscere, parlare. Abilità linguistiche, capacità operative e processi d'apprendimento,* "Collana Giscel" n° 2, Milano, Franco Angeli, 2004, pp.434, Euro 29,00.

Il volume indaga l'intreccio tra abilità linguistico-cognitive, capacità operative e processi d'apprendimento, riconsiderando il peso della dimensione biologica e mentale nella costituzione ed elaborazione di conoscenze, saperi, competenze. L'interazione sinergica di corpo e mente, motricità e pensiero, percezioni sensoriali, linguaggio e ambiente costituisce l'oggetto di discussione teorica dei contributi di autorevoli studiosi (T. De Mauro, P. Guidoni, A. Oliverio, A. Oliviero Ferraris, C. Pontecorvo, R.C. Schank), a cui si affiancano studi e ricerche più centrati sulla operatività didattica.

Barbara Hans-Bianchi, *La competenza scrittoria mediale. Studi sulla scrittura popolare,* Beihefte zur Zeitschrift für romanische Philologie 330, Tübingen, Max Niemeyer, 2005, pp. viii + 351, Euro 68,00 / sFr 114,00.

Quali fattori sono da ritenere costitutivi della competenza ortografica, interpuntoria e formale? Questo il quesito posto nell'analisi di un corpus di 'scrittura popolare' toscana del '900: una trentina di testi di scriventi toscani con istruzione scolastica elementare. Sullo sfondo teorico della scrittura come sistema e come processo, nonché della sua acquisizione, l'analisi empirica mira ad individuare i fattori socioculturali e situazionali rilevanti per la competenza scrittoria mediale e per la selezione funzionale alla comunicazione.

Elena Tamborrino, *Indicare il tempo. Osservazioni nell'area salentina. Prefazione di Maria Teresa Romanello,* Bari, Palomar, 2006, pp. 233, Euro 22,00.

Questo testo descrive gli usi dei parlanti relativamente a avverbi e locuzioni di tempo.L'area di pertinenza è quella salentina. Le modalità di raccolta si riferiscono a materiali linguistici attuali, prodotti in situazioni comunicative

diversificate.L'analisi si fonda anche sulle acquisizioni della dialettologia storica. In tal modo, l'interpretazione dei materiali -sottoposti anche ad analisi quantitativa- può fornire riflessioni su un più generale discorso di rapporti tra le diverse varietà del repertorio nell'area.

I soci sono invitati ad inviare informazioni per questa rubrica a Emanuele Banfi, Facoltà di Scienze della Formazione, Università degli Studi di Milano-Bicocca, P.zza dell'Ateneo Nuovo 1, 20126 Milano
Telefono: 02-64486817 / Fax: 02-64486995
E-mail: emanuele.banfi@unimib.it oppure banfi@planet.it

Nell'inviare dati relativi a libri di esclusivo interesse scientifico, i soci sono pregati di attenersi al seguente schema:

- nome e cognome dell'autore o del curatore
- titolo ed eventuale sottotitolo
- luogo di stampa
- editore (se si tratta di editore locale privo di rete distributiva o di editore non italiano, indicare tra parentesi l'indirizzo)
- data di edizione
- numero di pagine
- prezzo di copertina.

Tutti i dati devono essere redatti in carattere tondo, senza sottolineature.
A ciò si aggiunga una breve nota (non più di 5 righe) sul contenuto del libro.
La SLI si riserva di modificare, per esigenze di uniformità redazionale, i testi inviati. Non si potrà tener conto di libri inviati senza la scheda redatta secondo le norme sopra riportate.

L'ordine di pubblicazione delle schede bibliografiche rispetta l'ordine di arrivo delle singole segnalazioni.

# BOZZA TEMARIO XLI CONGRESSO SLI

## Alloglossie e comunità alloglotte nell'Italia contemporanea

*Teorie, applicazioni e descrizioni, prospettive.*

Il congresso si propone di offrire spunti di riflessione e confronto di opinioni a livello teorico e descrittivo in merito ad una delle componenti essenziali della nuova situazione plurilinguistica del paese. Pertanto l'attenzione sarà rivolta ai tre assi tematici sotto indicati e alle relative sottoarticolazioni.

1) Alloglossie
   Descrizione delle varietà alloglotte ai vari livelli di analisi;
   analisi contrastiva delle alloglossie in riferimento alle rispettive 'norme' standard;
   dinamiche linguistiche interne alle alloglossie dovute al contatto con l'italiano o con altre lingue e/o alla condizione di isolamento dell'alloglossia.

2) Comunità alloglotte
   Dinamiche sociali e sociolinguistiche delle alloglossie;
   relativismo linguistico, lingue e culture, interculturalità;
   gruppi/comunità di parlanti, interazioni, rapporti con le istituzioni.

3) Italiano L2
   Italiano come lingua di contatto con le alloglossie di antico e di nuovo insediamento;
   italiano L2 e interlingue in ambienti di apprendimento;
   italiano L2 a scuola.

*Sede e date proposte*
Università "G. D'Annunzio" di Chieti-Pescara
Polo didattico di Pescara, Viale Pindaro n. 42

27-28-29 settembre 2007 oppure
4-5-6 ottobre 2007.

*Comitato scientifico*
Tullio De Mauro
Emanuele Banfi
Augusto Carli
Vincenzo Orioles
Massimo Vedovelli
Paola Desideri
Carlo Consani

*Comitato organizzatore*
Carlo Consani
Paola Desideri
Francesca Guazzelli
Carmela Perta
Domenico Russo

# NOTIZIARIO

## Notiziario GISCEL
## Gruppo di intervento e Studio nel Campo dell'Educazione Linguistica

**http://www.giscel.org**

a cura di Adriano Colombo

Sede legale: presso Università di Roma «La Sapienza», Dipartimento di Studi filologici, linguistici e letterari, Piazzale Aldo Moro 5, 00185 Roma.
Indirizzo operativo: Casella postale n. 4, 40050 Funo di Argelato (BO)

### ORGANISMI NAZIONALI

**Segreteria nazionale**

**Segretario:** Adriano Colombo, [redacted]
**Consigliere:** Emanuela Piemontese, Dipartimento di Studi filologici, linguistici e letterari, via A. Cesalpino 12/14, 00161 Roma;
e-mail: emanuela.piemontese@uniroma1.it
**Consigliere:** Francesco De Renzo, Dipartimento di Filologia, Università della Calabria, Via Pietro Bucci, cubo 27 B. Rende (CS);
e-mail: francoderenzo@inwind.it

**Comitato scientifico della collana "Quaderni del Giscel"**

Adriano Colombo, Cristina Lavinio, Maria Pia Lo Duca, Maria Antonietta Marchese, Simonetta Rossi, Immacolata Tempesta

**Segreterie regionali sono presenti nelle seguenti aree:**

Abruzzo, Calabria, Campania, Emilia-Romagna, Friuli-Venezia Giulia, Giappone, Lazio, Liguria, Lombardia, Marche, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Veneto.

L'iscrizione al Giscel è subordinata alla iscrizione SLI ed è soggetta alle norme adottate dai singoli gruppi regionali secondo quanto previsto nello Statuto.

**Lettera del Segretario Nazionale GISCEL**

Care amiche e cari amici,

vi scrivo nel momento di un passaggio politico delicato (15 maggio), quando è impossibile fare previsioni su chi sarà il nuovo ministro dell'Istruzione; tanto meno sulla futura politica scolastica, di cui nessuno parla. Abbiamo in proposito una sola certezza: il tema non interessa un fico ai nostri politici, né del resto alla cosiddetta società civile (basta guardare a come i giornali parlano di scuola, nei rari casi in cui se ne ricordano).

Questo non diminuisce il nostro impegno e le nostre responsabilità. Presto saremo chiamati a pronunciarci su scelte importanti, e questo esigerà di avere canali efficaci e rapidi di discussione tra noi. Per parte mia, vi anticipo alcune opinioni personali, in parte già confortate dall'assemblea:

- la "riforma Moratti" è da abrogare (questo punto lo abbiamo votato), qualsiasi tentativo di rimediare parzialmente o dall'interno ai suoi effetti peggiori creerebbe solo ulteriore confusione normativa e stress per le scuole;
- la scuola ha bisogno di un momento di raccoglimento, di ritrovare le sue ragioni di fondo; l'esperienza ha dimostrato che le grandi riforme organiche in questo paese sono impossibili;
- ma anche per questo c'è molto da fare; per esempio, ci sono da riscrivere gli "obiettivi specifici di apprendimento" (che in sé non sono un'invenzione recente, sono voluti da una legge del 1999); c'è da ripensare (o mantenere) le strutture per la formazione iniziale degli insegnanti; e tali questioni richiederanno il contributo delle nostre competenze specifiche.

La prossima assemblea che terremo a Vercelli in occasione del Congresso nazionale SLI potrà essere un momento di discussione su queste questioni.

Si è chiuso da poco il nostro XIV Convegno nazionale di Siena, 6-8 aprile, con un notevole successo di pubblico presente e di qualità scientifica dei contributi. Rinnovo qui il ringraziamento al Comitato Scientifico e al gruppo organizzatore condotto da Massimo Vedovelli e da Monica Barni. È già al lavoro il Comitato Scientifico per il XV Convegno nazionale, che si terrà in Lombardia (quasi sicuramente a Pavia) nel 2008. Il tema che è stato scelto a Siena, *Descrizione, misurazione, valutazione delle competenze linguistiche*, è relativamente nuovo per la nostra tradizione di studi, e potrà aiutarci a mettere sempre meglio a fuoco un'idea complessiva di curricolo. Non si tratta, ovviamente, di fare del tecnicismo docimologico, ma di assumere un punto di vista strategico da cui guardare a tutto il processo dell'educazione linguistica.

Nel campo delle pubblicazioni, stiamo recuperando alcuni ritardi accumulati nel periodo del difficile cambio di casa editrice. È uscito il volume *Questioni linguistiche e formazione degli insegnanti*, a cura di Domenico Russo, e stanno per uscire due volumi ricavati dal Convegno nazionale di Lecce, 2004. Con questo la distanza tra un evento e la pubblicazione dei risultati si è accorciata a due anni; il prossimo obiettivo sarà tornare a un solo anno, come ai vecchi tempi (purtroppo questo non accadrà ancora per il volume sul trentennale delle *Dieci*

*Tesi*). Ricordo ancora una volta che è necessario il massimo impegno di tutta l'associazione in tutte le sedi, molto più che in passato, per assicurare almeno il minimo indispensabile di diffusione alle nostre pubblicazioni. Sono sempre disponibili copie delle schede-catalogo delle nostre collane, da diffondere in ogni occasione utile.

Quanto al nostro radicamento territoriale, è inutile negare che in alcune sedi stiamo vivendo qualche momento di difficoltà. Intanto però stanno nascendo nuovi gruppi regionali. Quando leggerete queste righe, si sarà già tenuta a Trento una Giornata di studio che prelude alla costituzione del gruppo trentino; il 7 settembre qualcosa di analogo succederà a Campobasso per il Molise; contatti sono avviati in altre sedi.

Ci sono insomma le premesse per acquisire quella maggiore consistenza organizzativa che dia più autorevolezza alle nostre tesi. Al volgere di questo anno ripeteremo una verifica generale del numero e della regolarità dei nostri iscritti, e speriamo di poter constatare che il GISCEL è anche formalmente in buona salute, pronto a far fronte ai suoi compiti in questo complicato e difficile paese.

Buon lavoro a tutte e a tutti.

Adriano Colombo

**Convocazione dell'Assemblea nazionale GISCEL**

L'assemblea nazionale GISCEL sarà convocata in occasione del XL Congresso della SLI (Vercelli, 21-23 settembre 2006, luogo e ora da definirsi).

Ordine del giorno:

1. Comunicazioni della segreteria
2. documenti e progetto per seminari di formazione
3. stato delle pubblicazioni
4. questioni di politica scolastica
5. varie ed eventuali.

**Verbale dell'Assemblea nazionale GISCEL – Siena, 8 aprile 2006**
(estratto – il verbale integrale appare sul sito www.giscel.org)

Il giorno 8 aprile 2005, presso l'Auditorium dell'Università per Stranieri di Siena, in Via Pispini 1, si riunisce, alle ore 17.30 (in seconda convocazione), l'Assemblea nazionale dei soci GISCEL per discutere e deliberare sul seguente ordine del giorno:

relazione del Segretario nazionale
approvazione del bilancio
rinnovo cariche sociali
modifica dello Statuto (sede legale)
collana Giscel
XV Convegno nazionale Giscel: norme, sede, tema
seminario nazionali e documento sulla formazione
varie ed eventuali.

Presiede il Segretario nazionale, Adriano Colombo. Verbalizza Rosa Calò, membro uscente della Segreteria nazionale. Sono presenti i soci elencati nell'allegato 1.

**1. Relazione del Segretario nazionale**

**1. La cornice politica**

Il futuro in cui possiamo collocare il ruolo della nostra associazione dipende da un cambiamento politico che deve verificarsi proprio in questi giorni. Tutti ci contiamo e tutti abbiamo contribuito al suo verificarsi. Credo che se il cambiamento ci sarà, la scuola dovrà aspettarsi prima di tutto un restauro: ha bisogno di ritrovare serenità, chiarezza di norme, di dedicarsi ai suoi compiti primari. Nel nostro seminario sulla formazione in servizio, e di conseguenza nel documento che propongo alla vostra approvazione, si è parlato e si parla di questo.

Per questo bisogno di serenità, nel prossimo futuro eviterei di parlare di grandi riforme (che sono possibili solo nei paesi normali). Certo con una

bisognerà fare i conti, quella dell'autonomia, che non è stata ancora veramente applicata. Se riteniamo che si debba ancora tentare questa strada, allora si pone il problema non solo di riscrivere i risibili obiettivi specifici di apprendimento che sono stati varati, ma di avere un sistema di valutazione nazionale serio, che oggi non c'è. Dovremo oggi decidere se confermare l'idea di un nostro seminario sulla valutazione per il prossimo giugno.

Indipendentemente dal cambiamento politico, c'è un grande mutamento in corso: una intera generazione sta uscendo dalla scuola, nei prossimi anni si prevedono 250.000 pensionamenti (tra loro, il 54% dei colleghi che abbiamo raggiunto nella nostra ricerca dello scorso anno sulla conoscenza delle *Dieci tesi*). Ci aspetta quindi un enorme lavoro, se vogliamo che le nostre idee e le relative pratiche continuino a vivere nella scuola.

## 2. Quel che abbiamo fatto e che stiamo per fare

Che cosa abbiamo fatto, quali possibilità di intervento abbiamo per far fronte a un tale compito? Comincio da quel che già c'è e si fa:

- abbiamo ottenuto la qualifica di "Ente di formazione";
- abbiamo diffuso ad oggi circa 1500 copie gratuite delle *Dieci tesi*, più 2000 copie del documento *Oltre la "riforma" Moratti*, che contiene ampi stralci del nostro testo fondativo;
- abbiamo la nostra presenza in molte SSIS, sia come docenti, sia come supervisori, e forse in alcuni corsi di laurea in Formazione primaria;
- ci sono poi le iniziative pubbliche, dalla più grande che si è svolta un anno fa per il trentennale delle *Dieci tesi* (e bisogna rinnovare il grazie alle amiche del Giscel Lazio che si sono prodigate per il suo successo), a quelle che si svolgono nelle sedi regionali, otto finora in questo anno;
- la collana dei quaderni Giscel ha ripreso un buon ritmo: questo anno vedrà l'uscita di tre o quattro volumi.

Per una nostra mobilitazione e presenza diffusa ci vorrebbe un progetto nazionale sul genere di quel che è stato il *Laboratorio di scrittura*, che ci aveva consentito un contatto con oltre mille insegnanti, un'avanguardia significativa. Ma le nuove norme sulla formazione in servizio la affidano interamente alle Direzioni regionali (U.S.R.), tranne i progetti nazionali per via telematica affidati all'INDIRE. Quanto a questi, non siamo finora entrati nel progetto "Neoassunti", per il quale le nostre forze non bastano e le cui condizioni non sono chiare. Siamo invece impegnati a fondo nel progetto "Abilità di base – Area linguistica" accanto a LEND e ADI: nelle prossime settimane inizierà una formazione dei formatori, tra i quali una ventina designati dal Giscel in seguito all'appello che ho lanciato nello scorso ottobre; si può sperare di creare un gruppo di formatori Giscel in parte nuovi, affiatati, in contatto fra loro, capaci di intervenire sia in sede telematica nazionale sia in sedi locali per attività formative in presenza.

Altri impegni derivano dall'assemblea di Milano e dal Seminario di Roma: la gestione diretta di corsi residenziali brevi interregionali, la formulazione di proposte formative da rivolgere a reti di scuole o ad autorità scolastiche locali.

Su questo si è lavorato, anche col contributo del Comitato scientifico della Collana, e in parte con un dibattito per posta elettronica: il risultato sono il documento di indirizzo sulla formazione e le linee generali per l'elaborazione di "pacchetti formativi" (da approvare). Chiede anche di confermare la delibera dell'assemblea di Milano sui corsi residenziali brevi interregionali. Spetterà poi ai gruppi regionali avanzare le nostre proposte formative alle autorità locali, ovunque sia possibile.

A Milano e a Roma si è parlato anche della possibilità di gestire uno o più "master" in Educazione linguistica in collaborazione con uno o più atenei. Le informazioni raccolte grazie a Maria Pia Lo Duca e a Emanuela Piemontese non sembrano per ora rendere realistica questa prospettiva.

### 3. Le nostre risorse

Il Giscel costituisce uno straordinario patrimonio di competenze e di conoscenza collettiva, ma il tempo che ciascuno può dedicare all'associazione è limitato, e qui sta la nostra debolezza organizzativa.

Abbiamo una presenza attiva in undici regioni (più l'attivissimo gruppo extraterritoriale del Giappone). Sono lieto di annunciare che due nuove realtà si stanno per aggiungere: il 18 maggio a Trento si terrà una Giornata di studio che sarà il lancio del gruppo trentino, il 7 settembre lo stesso accadrà a Campobasso per il costituendo gruppo Molise. Un grazie a Vito Maistrello e a Giuliana Fiorentino, promotori delle due iniziative.

I tentativi di risvegliare i Giscel addormentati (Piemonte, Liguria, per altri versi Marche, Friuli-Venezia Giulia) non hanno avuto finora successo. Ci sono gruppi nuovi o rinnovati di recente che mostrano una notevole vitalità, a cominciare dal gruppo giapponese e da quello toscano che ha organizzato questo convegno. Ma altre realtà tradizionalmente forti danno segni di stanchezza, a cominciare dalla mia Emilia-Romagna. Da varie parti si segnala la difficoltà di agganciare colleghi delle nuove leve (che, sappiamo, non sempre vuol dire giovani); una difficoltà comune del resto ad altre associazioni disciplinari. Si aggiunge a volte una qualche difficoltà di comunicazione fra la segreteria nazionale e quelle locali: certi stimoli non sembrano raccolti da tutti, o forse non sono gli stimoli giusti.

D'altra parte, esistono nell'associazione energie non pienamente utilizzate. Bisogna trovare il modo di inserire nel lavoro comune tutti e tutte coloro che possono dare un contributo, evitando che si formi, anche inconsapevolmente, un certo spirito di gruppo che tende a chiudersi (sia questo a livello nazionale, sia a livello locale). Chiunque ha voglia e disponibilità per fare deve trovare incoraggiamento: gli schemi organizzativi, nazionali e locali, devono avere la flessibilità necessaria per cogliere ogni opportunità.

I compiti che ci attendono richiedono un'associazione più dinamica, più aperta all'esterno. Su questo intendo proseguire e intensificare il mio impegno.

Sui temi proposti nella relazione intervengono Cristina Lavinio, Anna Rosa Guerriero, Valter Deon; Pinella Depau chiede che l'Assemblea prenda posizione

per l'abrogazione della Legge 53. L'Assemblea all'unanimità risulta favorevole all'abrogazione della 53.

**4. Approvazione del bilancio**

Il Segretario illustra il bilancio di cui all'allegato 2, che è approvato all'unanimità.

Cristina Lavinio chiede che sia inserita nel prossimo bilancio la voce relativa al contributo del Giscel nazionale al Convegno di Siena. Adriano Colombo fa presente che ha già assicurato tale contributo. Silvana Ferreri ricorda che la quota-contributo ai convegni nazionali era già stata fissata e che il Giscel nazionale ha sempre erogato un contributo; Cristina Lavinio propone che si fissi una quota (2.500 euro) passibile di eventuali compensazioni (in più o in meno secondo il caso). L'Assemblea delibera quest'ultima proposta all'unanimità.

**5. Rinnovo delle cariche sociali**

Il Segretario presenta le proposte messe a punto dal Comitato nomine e le pone in votazione.

**Segreteria nazionale**

Segretario: Adriano Colombo, al termine del mandato relativo al primo biennio, rieleggibile. Membri della segreteria nazionale: M. Emanuela Piemontese, al termine del mandato relativo al primo biennio, rieleggibile, e Francesco De Renzo. L'Assemblea approva all'unanimità.

**Comitato scientifico della Collana Giscel.**

Tre membri sono giunti al termine del mandato quadriennale: Silvana Ferreri, Maria Maggio e Francesca Romana Sauro. Si propone che siano sostituiti da Maria Pia Lo Duca, Simonetta Rossi, Immacolata Tempesta. L'Assemblea approva all'unanimità.

Per il prossimo biennio, il Comitato scientifico della Collana è quindi costituito da: Cristina Lavinio, Maria Antonietta Marchese, Maria Pia Lo Duca, Simonetta Rossi, Immacolata Tempesta e Adriano Colombo.

Al termine delle votazioni, si ringraziano i membri uscenti dalla Segreteria (Rosa Calò) e dal CS (Silvana Ferreri, Maria Maggio e Francesca Romana Sauro).

**6. Modifica dello Statuto (sede legale)**

Si propone la modifica dell'ultimo comma dell'art. 1 dello Statuto Giscel relativamente alla sede legale del Giscel che è la seguente: Presso l'Università di Roma "La Sapienza", Dipartimento di Studi filologici, linguistici e letterari, Piazzale Aldo Moro 5, 00185 Roma. L'Assemblea approva la modifica all'unanimità.

Tenendo conto del fatto che alcuni soci del Giscel Lombardia devono anticipare la partenza, il Segretario propone di discutere subito il punto 6 all'o.d.g. L'Assemblea approva.

## 7. XV Convegno nazionale Giscel: norme, sede, tema

Per quanto riguarda la sede del prossimo Convegno nazionale, il Segretario ha ricevuto una parziale disponibilità da parte del Giscel Lombardia. La Segretaria del Giscel Lombardia precisa che al momento si tratta di un'ipotesi da verificare, La sede più adatta potrebbe essere Pavia. Pensa di essere in grado di dare una risposta in settembre per l'Assemblea Giscel che si terrà a Vercelli durante il Convegno SLI.

Tullio De Mauro raccomanda che siano avviati contatti con l'Università di Pavia, dove c'è una presenza di docenti che potrebbero essere disponibili a collaborare.

Considerata la situazione attualmente incerta del Giscel Lombardia, Silvana Ferreri chiede la disponibilità del Giscel Veneto. Risponde Vittoria Sofia, segretaria del Giscel Veneto, la quale assicura una certa disponibilità a livello locale, ma non immediata, perché il gruppo sta vivendo una crisi di ricambio generazionale, il Veneto potrebbe sobbarcarsi l'organizzazione del Convegno nazionale successivo (2010).

Quanto al tema del XV Convegno, emergono dueproposte: 1° la valutazione (già proposto a Lecce da Edoardo Lugarini, Massimo Vedovelli e da altri soci), 2° la riflessione linguistica (tema che viene ora proposto dal Giscel Veneto).

A favore del primo tema intervengono Elda Padalino, Immacolata Tempesta, Tullio De Mauro, Werter Romani, Massimo Vedovelli, Francesco De Renzo, Monica Barni, Rosa Calò, Massimo Vedovelli. In particolare, Tullio De Mauro propone delle parole-chiave sul tema della valutazione: descrittori, misurazione, valutazione della competenza linguistica. Werter Romani raccomanda di evitare sulla valutazione i tecnicismi docimologici.

Si esprimono a favore del secondo tema Cristina Lavinio, Silvana Ferreri, Vittoria Sofia e Vannina Pudda.

Il Segretario mette ai voti le due proposte:

Primo tema: Descrizione,misurazione, valutazione delle competenze linguistiche. Voti espressi: 38, di cui 37 favorevoli e 1 contrario. Astenuti: 6.

Secondo tema: Riflessione linguistica. Voti espressi: 33, di cui 15 favorevoli e 18 contrari. Astenuti: 10

In merito alla composizione del comitato scientifico del Convegno. Il Segretario propone di stabilire una norma che limiti a 6 il numero dei componenti. Si dichiarano contrarie Cristina Lavinio, Silvana Ferreri, Rosa Calò e M. Emanuela. Sentiti i pareri difformi, il Segretario ritira la proposta.

Come componenti del Comitato vengono designati: Adriano Colombo, Tullio De Mauro, Valter Deon, Cristina Lavinio, Pietro Lucisano, Alberto Sobrero, Massimo Vedovelli. Un altro componente sarà individuato nell'area della Lombardia, in relazione alla sede del convegno.

Data l'ora tarda, si decide di rinviare la discussione del punto 5 (Collana Giscel). Riguardo al punto 7, si rinvia la discussione e l'approvazione dei documenti sulla formazione all'Assemblea di Vercelli, ma il Segretario chiede comunque all'assemblea di pronunciarsi in merito alla opportunità di organizzare

nel giugno prossimo un seminario nazionale sulla valutazione, secondo quanto suggerito all'Assemblea di Milano. Loredana Corrà esprime delle riserve su tale iniziativa, visto il tema scelto per il XV convegno nazionale e la prassi del "seminario intermedio" che affronterà comunque il tema nel giugno 2007. Il Segretario accoglie le osservazioni della collega.

Infine il Segretario dichiara che dopo l'Assemblea di Vercelli sarà opportuno organizzare un seminario nazionale sulla formazione iniziale, destinato in primo luogo ai soci che già operano nelle SSIS (ed eventualmente nei corsi di laurea in Formazione primaria), coi diversi ruoli di docente di corsi o laboratori e di supervisore.

La seduta è sciolta alle ore 20.15.

Siena, 8 aprile 2006
Rosa Calò, Adriano Colombo

## Allegato 1

### Presenti all'Assemblea nazionale GISCEL di Siena, 7.4.2006

Adriano Colombo, Emanuela Piemontese, Rosa Calò (segreteria naz.), Domenico Russo, Giuseppina Pani (GISCEL Abruzzo), Francesco De Renzo, Anna Chiara Monardo, Raffaelina Pizzini, Maria Laura Calabrese (GISCEL Calabria), Velia Damiani, Giovanna Sessa, Marina Brandi, Anna Rosa Guerriero, Fabio M. Risolo (GISCEL Campania), Werther Romani, Silvana Loiero (GISCEL Emilia-Romagna), Edda Serra (GISCEL Friuli-Venezia Giulia), Paola Peruzzi, Ignazio Gioè (GISCEL Giappone), Iolanda Salacchi, Lidia Alesini, Sparta Tosti, Anna Maria Lettieri, Tullio De Mauro (GISCEL Lazio), Miria Carpaneto (GISCEL Liguria) Letizia Rovida, Fioretta Mandelli, Francesca Gaudenzio (GISCEL Lombardia), Carla Marello (GISCEL Piemonte), Immacolata Tempesta, Rosaria Solarino (GISCEL Puglia), Pinella Depau, Cristina Lavinio, Alessia Defraia, Luisa Milia, M. Teresa Lecca, Rosanna Figus, Vannina Pudda (GISCEL Sardegna), Francesca Cappdonna, Rosalia Misuraca, Ignazio Mirto, Maria Antonietta Marchese, Silvana Ferreri (GISCEL Sicilia), Antonella Benucci, Giosuè Piscopo, Stefania Semplici, P. Diadori, D. Troncarelli, Alessandra Felici, Andrea Villarini, Maurizio Sarcoli, Elda Padalino, Andreina Sottile, Monica Barni (GISCEL Toscana), Gioconda Rilievo, Cornelia Cazzorla, Nella Cazzadori, M.Giuseppina Lo Duca, Valter Deon, Vittoria Sofia, Loredana Corrà, Giuseppina Colmelet, Vito Maistrello (GISCEL Veneto), Simonetta Rossi, Catherine Camugli, Isabella Totaro, Carmela Camodeca

## Allegato 2
## Bilancio 25.5.2005 – 31.3.2006

| | | |
|---|---|---|
| **Stato patrimoniale 25.5.2005** | | |
| c.c. Unicredit | € | 18.336,85 |
| Deposito titoli | € | 25.324,49 |
| credito Colombo (piccole spese) | € | -130,84 |
| **Stato patrimoniale 31.3.2006** | € | 43.530,50 |
| c.c. Unicredit | € | 18.472,51 |
| Deposito titoli | € | 25.000,00 |
| *maggiori spese: € 58* | € | 43.472,51 |
| **Entrate** | | |
| cedole BTP | € | 656,24 |
| contributo MIUR | € | 5426,71 |
| diritti RCS 2004 | € | 1219,38 |
| diritti FrancoAngeli 2004 | € | 196,50 |
| diritti FrancoAngeli 2005 | € | 464,32 |
| 20% su contratto IPRASE (Deon-Maistrello) | € | 600,00 |
| vendita libri | € | 818,10 |
| | € | 9381.25 |
| **Uscite** | | |
| **Spese di segreteria** | € | 544,74 |
| telefoniche | € | 75,60 |
| postali | € | 148,04 |
| viaggi del segretario | € | 121,92 |
| diverse | € | 199,18 |
| **Assemblee e riunioni** | € | 5474,39 |
| residuo Giornate di Roma 18.4.05 e Bologna 9.5.05 | € | 991,31 |
| assemblea e CS Roma 11-12.6.2005 | € | 2382,05 |
| C. S. Roma 29.8.05 | € | 412,16 |
| assemblea di Milano 23.9.05 | € | 252,25 |
| seminario di Roma 15.1.06 | € | 1012,52 |
| C.S. Roma 18.3.06 (parziale) | € | 424,10 |
| **Acquisto libri** | € | 1715,62 |
| **Pubblicazioni gratuite** | € | 1150 |
| stampa di 1000 fascicoli *Dieci tesi* | € | 707,20 |
| stampa di 1000 schede catalogo L.N.I. | € | 390,00 |
| spedizioni | € | 52,80 |
| **Partecipazione alla Fiera Docet 2006** | € | 60 |
| **Banca** | € | 479,50 |
| (creazione e gestione del deposito titoli, spese di conto) | | |
| | € | 9424,25 |

*Maggiori spese* € *43*
Differenza contabile € 13

# INDIRIZZARIO GISCEL

(aggiornato al 15.5.2006)

**GISCEL** c/o Adriano Colombo, [redacted]

**Segreterie regionali**

| | |
|---|---|
| **Giscel Abruzzo**<br>[redacted] | Daniela Campitelli<br>[redacted] |
| **Giscel Calabria**<br>francoderenzo@inwind.it<br>tel. 0984 493118 | Francesco De Renzo<br>c/o Dipartimento di Filologia.<br>Università della Calabria<br>Via Pietro Bucci, cubo 27 B<br>87936 RENDE (CS) |
| **Giscel Campania**<br>[redacted] | Fabio Risolo<br>[redacted] |
| **Giscel Emilia-Romagna**<br>wromani@libero.it<br>tel. 051 2098556 | Werther Romani<br>Dip. di Italianistica, via Zamboni 32<br>40126 - BOLOGNA |
| **Giscel Friuli-Venezia Giulia**<br>[redacted] | Edda Serra<br>[redacted] |
| **Giscel Giappone** *(segreteria provvisoria)*<br>sferreri@unitus.it<br>tel. 0039 0761 357602 | Silvana Ferreri<br>Facoltà di Lingue e letterature straniere<br>Largo dell'Università<br>01100 VITERBO |
| **Giscel Lazio**<br>[redacted] | Sparta Tosti<br>[redacted] |
| **Giscel Liguria**<br>[redacted] | M. Cristina Castellani<br>[redacted] |
| **Giscel Lombardia**<br>[redacted] | Letizia Rovida<br>[redacted] |
| **Giscel Marche**<br>[redacted] | Paola Desideri<br>[redacted] |

| | |
|---|---|
| **Giscel Piemonte** | Agostino Roncallo |
| **Giscel Puglia** | Elena Tamborrino |
| **Giscel Sardegna** | Maria Teresa Lecca |
| **Giscel Sicilia** | M. Antonietta Marchese |
| **Giscel Toscana** | Elda Padalino |
| **Giscel Veneto** | Vittoria Sofia |

**Comitato scientifico della collana GISCEL**

| |
|---|
| **Adriano Colombo** |
| **Cristina Lavinio** |
| **Maria Pia Lo Duca** |
| **Maria Antonietta Marchese** |
| **Simonetta Rossi** |
| **Immacolata Tempesta** |

# NOTIZIARIO GSCP

**Relazione del coordinatore sulle attività svolte dal GSCP nel triennio 2003-2006.**

Nel corso del triennio sono stati organizzati dal GSCP (che attualmente conta circa 130 aderenti) un convegno a Padova sul tema *La manifestazione fonica delle emozioni* (novembre 2004) e il congresso internazionale *La comunicazione parlata – Spoken communication* a Napoli (febbraio 2006). Gli atti del convegno di Padova sono pronti per la stampa che sarà in forma di e-book presso l'editore Liguori di Napoli nella collana "Quaderni di comunicazione parlata". E' in corso la raccolta dei testi del congresso di Napoli, la cui pubblicazione è prevista entro il 2006.

L'assemblea dei soci, tenutasi nel corso del convegno di Napoli 2006, ha deliberato che il prossimo congresso internazionale si terrà nel 2009, presumibilmente in febbraio, in luogo da stabilire. Ha inoltre espresso l'auspicio che nel corso del triennio 2006-2009 gruppi di soci interessati organizzino convegni o seminari su argomenti specifici. Sono stati suggeriti i seguenti temi: *La tematizzazione della 'comunicazione parlata' nel pensiero linguistico occidentale* (proposta di Albano Leoni); *Parlato e nuovi media* (proposta di Orletti); *Bidirezionalità tra scritto e parlato* (proposta di Pettorino); *La comunicazione parlata uomo/macchina* (proposta di Ferrari).

Nel corso dell'assemblea sono state rinnovate le cariche venute a scadenza (Albano Leoni, coordinatore 2006-2009; Voghera e Danieli, comitato di coordinamento 2006-2009). L'assemblea ha altresì approvato la proposta di istituzione del comitato nomine e la conseguente richiesta al CE della SLI di approvare la relativa modifica del regolamento (approvata dal CE del 21 aprile 2006).

E' stato attivato il sito del gruppo: www.comunicazioneparlata.org

Nel corso del triennio si sono verificati alcuni eventi che, pur non essendo espressione diretta e formale del GSCP si collocano nello stesso ambito di interesse.

E' stata istituita la rivista on line "Comunicazione parlata" presso l'editore Liguori di Napoli (dettagli, informazioni, fogli di stile e altro in http://www.liguori.it/areaautori/comunicazioneparlata.asp). Alla rivista è associata una collana "Quaderni di comunicazione parlata". Il primo volume, appena uscito è F. Albano Leoni e R. Giordano (a c. di), *Italiano parlato. Analisi di un dialogo* (con un cd-rom contenente il materiale audio), Napoli, Liguori, 2006. Il secondo sarà costituito dagli Atti del convegno padovano già ricordato.

## REGOLAMENTO AGGIORNATO

### Art. 1

In seno alla SLI è costituito il Gruppo di Studio per la Comunicazione Parlata, ai sensi dell'art. 21 dello Statuto.

### Art. 2

Il Gruppo ha la finalità di promuovere e coordinare gli studi sulla comunicazione parlata, favorendo la collaborazione e lo scambio tra quanti, a qualsivoglia titolo, si occupano di questo tema. A tal fine il gruppo può organizzare incontri di studio, convegni, seminari e curare pubblicazioni.

### Art. 3

Fanno parte del gruppo tutti i soci SLI che ne facciano richiesta. L'adesione di studiosi esterni è subordinata alla loro iscrizione alla SLI..

### Art. 4

Sono organi del Gruppo l'Assemblea degli aderenti, il Coordinatore, il Comitato di coordinamento e il Comitato per le nomine.

### Art 5

L'Assemblea è costituita da tutti i soci SLI che abbiano aderito al gruppo. Elegge a maggioranza semplice il coordinatore e con voto limitato i membri del Comitato di coordinamento su proposta del Comitato per le nomine. Approva le proposte del coordinatore e propone iniziative. L'assemblea si riunisce in occasione delle giornate di studio organizzate dal gruppo. Per materie di particolare importanza, come il rinnovo delle cariche sociali o le modifiche del presente regolamento, l'assemblea può venire consultata anche telematicamente. Le operazioni di voto sono gestite da una Commissione Elettorale nominata dal Comitato di coordinamento, al quale dovranno pervenire le candidature.

### Art. 6

Il Coordinatore è eletto a maggioranza semplice dall'Assemblea degli aderenti, dura in carica un triennio ed è rieleggibile consecutivamente una sola volta. Convoca e presiede l'assemblea e il comitato di coordinamento. Formula, d'intesa con il coordinamento e raccogliendo anche suggerimenti di singoli soci o di gruppi di soci, la programmazione delle attività. Tiene i contatti con studiosi esterni o con altre associazioni quando ciò sia necessario per la realizzazione di specifiche iniziative di studio. Riferisce periodicamente al Presidente e al CE della SLI sulle attività del gruppo.

### Art 7

Il Comitato di coordinamento è costituito, oltre che dal Coordinatore, che lo presiede, da altri 4 membri, di cui uno con funzione di segretario, eletti dall'Assemblea con voto limitato a 3 preferenze. I membri durano in carica un triennio e sono rieleggibili consecutivamente una sola volta. Il comitato coadiuva

il coordinatore nella formulazione dei programmi delle attività e nella loro gestione e cura l'elenco e l'indirizzario degli aderenti.

**Art. 8**

Il Comitato per le nomine è composto da tre aderenti in numero di due ogni anno e per la durata di tre anni ciascuno. Il socio che ha raggiunto il terzo anno di carica fungerà da Presidente del comitato. Il Comitato propone le candidature per il rinnovo degli organismi. Se il Segretario riceverà, almeno tre settimane prima dell'elezione, sei o più designazioni dello stesso Socio per la medesima carica, egli conferirà a questi la candidatura a parità di condizioni con il candidato designato dal Comitato per le nomine. I nominativi di coloro che avranno effettuato la designazione scritta al Segretario dovranno rimanere segreti.

**Art. 9**

La pubblicizzazione delle attività del Gruppo avviene tramite il bollettino della SLI.

**Art. 10**

Le attività del gruppo saranno autofinanziate e graveranno sulla SLI solo per ciò che concerne la diffusione delle notizie sul bollettino

**Art. 11**

Le modifiche al presente regolamento debbono essere approvate dai due terzi degli aderenti.

**Art. 12**

In fase transitoria e fino alla prima convocazione degli aderenti, l'assemblea è costituita dai soci proponenti, Federico Albano Leoni, Emanuele Banfi, Carla Bazzanella, Gaetano Berruto, Pier Marco Bertinetto, Tullio De Mauro, Wolfgang Dressler, Anna Giacalone Ramat, Michele Loporcaro, Emanuela Magno Caldognetto, Marco Mancini, Giovanna Marotta, Alberto Mioni, Massimo Pettorino, Paolo Ramat, Raffaele Simone, Alberto Sobrero, Massimo Vedovelli, Miriam Voghera e la funzione di coordinatore è svolta dal decano dei proponenti.

**Norma transitoria**

Nella prima applicazione del presente regolamento, i due membri del Comitato di coordinamento che, nelle votazioni indette nel corso della prima assemblea del gruppo, avranno ottenuto più voti o che, a parità di voti, avranno maggiore anzianità restano in carica per 4 anni per consentire, a regime, il periodico ricambio parziale del Comitato stesso.

**Norma transitoria bis**

Nella prima costituzione del comitato nomine, a partire dal 2007, esso sarà costituito dai due componenti del Comitato di Coordinamento che ottennero il maggior numero di voti all'atto della costituzione del Comitato di Coordinamento nel 2003 e da un terzo componente da loro designato. Il componente del comitato in assoluto più votato svolgerà le funzioni di Presidente, durerà in carica un anno e sarà sostituito nel 2008; il componente del comitato, secondo quanto a numero di voti, durerà in carica due anni, svolgerà la funzione di Presidente nel corso del secondo anno e sarà sostituito nel 2009.

# NOTIZARIO GSPL

**Società di Linguistica Italiana – Gruppo di Studio sulle Politiche Linguistiche**
**Associazione Italiana di Linguistica Applicata**

Giornata di Studio del 31 marzo 2006
«Quali politiche linguistiche per l'Europa e l'Italia?»

Si è tenuta a Milano, presso l'«Aula Massa» dell'Università di Milano-Bicocca, alla presenza di circa quaranta di intervenuti, la tavola rotonda che ha costituito la giornata di studi «Quali politiche linguistiche per l'Europa e l'Italia?», organizzata dal Gruppo di studio sulle politiche linguistiche e dalla Associazione Italiana di Linguistica Applicata. Sono intervenuti Carla Marcato (Presidente del Centro Internazionale sul Plurilinguismo), Francesco Sabatini (Presidente dell'Accademia della Crusca), Leonardo Savoia (Presidente della Società di Linguistica Italiana), Luca Tomasi (Dirigente della Direzione Generale per l'educazione e la cultura dell'Unione Europea), moderati da Gabriele Iannàccaro (coordinatore del Gruppo di studio sulle politiche linguistiche e segretario dell' Associazione Italiana di Linguistica Applicata). Erano in sala, fra gli altri, anche Augusto Carli (presidente dell' Associazione Italiana di Linguistica Applicata e membro del coordinamento del Gruppo di studio sulle politiche linguistiche) e Emanuele Banfi (membro del coordinamento del Gruppo di studio sulle politiche linguistiche).

Il dibattito, intervallato da una colazione in piedi, è stato ampio, interessante e articolato; dopo una presentazione da parte degli intervenuti delle attività e concezioni di politica linguistica delle Istituzioni che rappresentano, sono emersi alla discussione in particolare i seguenti temi:
- il ruolo delle lingue nazionali e in particolare dell'inglese nell'amministrazione dell'Unione Europea;
- il plurilinguismo europeo e il problema di una *lingua franca*;
- la gestione del plurilinguismo nell'ambito della Repubblica Italiana: la legge 482/99, le leggi e istituzioni regionali e locali demandate all'implementazione delle lingue minoritarie, le proposte di creazione di un Consiglio superiore della Lingua Italiana;
- la divulgazione della linguistica e l'affermazione di un corretto «sguardo linguistico» sui problemi della società.

È stata avanzata la proposta di costituire una camera permanente di discussione fra le Istituzioni presenti e altre individuate da contattare (come la Società Italiana di Glottologia e la Associazione per la Storia della Lingua Italiana) che possa, esaminando problemi e istanze concrete relative alla politica linguistica, giungere a posizioni comuni e che costitituire un interlocutore per le istituzioni amministrative (la Direzione Generale per l'educazione e la cultura dell'Unione Europea riconoscerebbe volentieri un tale interlocutore).

In tal senso si è proposto di trovarsi per un'altra Giornata di studio a Udine

presso il Centro Internazionale sul Plurilinguismo nel dicembre del corrente anno, dedicata alla discussione di uno o due temi concreti, individuati nel frattempo. Si sono perciò invitati i presenti e gli associati al Gruppo di studio sulle politiche linguistiche e all'Associazione Italiana di Linguistica Applicata a contribuire all'individuazione dei temi tramite proposte, suggerimenti e osservazioni, che possono confluire in un apposito *forum* di discussione aperto sul sito del Gruppo di studio sulle politiche linguistiche all'indirizzo web http://www.sli-gspl.net/forum/formprogetti.html.

MODULO PER IL VERSAMENTO DELLA QUOTA DI ASSOCIAZIONE ALLA SLI TRAMITE CARTA DI CREDITO

Nome e cognome .....

.....

indirizzo .....

.....

indirizzo di posta elettronica .....

tipo e numero della carta di credito .....

data di scadenza della carta di credito .....

importo pagato per l'associazione alla SLI .....

autorizzo la pubblicazione dei miei dati personali (nome e indirizzo) sull'indirizzario del bollettino e dle sito SLI

data.....

firma.....

**Elenco dei paesi con prodotto interno lordo pro capite superiore ai 10.000 dollari*.**

Lussemburgo, USA, Kuwait, Svizzera, Qatar, Singapore, Giappone, Canada, Norvegia, Emirati arabi uniti, Danimarca, Belgio, Austria, Germania, Francia, Australia, Islanda, Italia, Brunei, Gran Bretagna, Paesi Bassi, Svezia, Nuova Zelanda, Finlandia, Israele, Bahama, Irlanda, Cipro, Spagna, Maurizio, Arabia saudita, Portogallo, Bahrain, Malta, Grecia, Barbados, Corea del Sud.

* Fonti: Banca mondiale, FMI, ONU