ISTITUTO DI LINGUISTICA COMPUTAZIONALE "A. ZAMPOLLI"

2017

Rapporto Annuale

Consiglio Nazionale delle Ricerche

Sommario

Prefazione

A nome dell'Istituto è per me un grande piacere introdurre il terzo Rapporto Annuale dell'Istituto di Linguistica Computazionale "Antonio Zampolli" (ILC) del Consiglio Nazionale delle Ricerche (CNR) che - nell'ottica della trasparenza e del desiderio di condivisione con il personale interno e con tutte le parti interessate - raccoglie e sintetizza le principali attività svolte nel corso dell'anno 2017.

Tra i risultati più significativi del 2017 si segnalano in particolare:

- *l'ulteriore ampliamento dei rapporti con la comunità scientifica nazionale, anche grazie al ruolo dell'ILC come rappresentante del consorzio italiano di CLARIN-ERIC, l'Infrastruttura di ricerca per le Scienze Umane e Sociali (SSH). Tramite il Coordinatore Nazionale per l'Italia è stata condotta una capillare opera di disseminazione presso la comunità dei potenziali utenti nell'ambito della Linguistica Computazionale e delle SSH. Al consolidamento e all'estensione dei rapporti con la comunità scientifica nazionale hanno contribuito anche: la partecipazione di ricercatori dell'ILC alle attività dell'Associazione Italiana di Linguistica Computazionale (AILC); la partecipazione a bandi regionali, nazionali e internazionali; la sottoscrizione di numerosi protocolli di intesa, dichiarazioni di intenti, accordi quadro e convenzioni con università, enti di ricerca e istituzioni culturali, così come la collaborazione con altri istituti CNR per lo sviluppo di progetti interdisciplinari;*

- *il consolidamento e il potenziamento dell'ampia rete di relazioni internazionali sia nell'ambito della Linguistica Computazionale sia delle SSH, grazie alla posizione dell'ILC come referente tecnologico nazionale nell'ambito dell'azione ELRC - European Language Resource Coordination, nonché alle visite di studiosi stranieri presso la sede dell'Istituto; l'avvio o la prosecuzione di collaborazioni scientifiche internazionali extra-progettuali con università, enti di ricerca e istituzioni culturali internazionali;*

- *il proseguimento delle attività editoriali, di disseminazione e divulgazione scientifica condotte tramite: la direzione di tre riviste scientifiche del settore; l'organizzazione di importanti conferenze/workshop nazionali e internazionali; la partecipazione a eventi divulgativi rivolti al grande pubblico, tra cui è da segnalare la partecipazione alla prima edizione di Fiera Didacta Italia;*

- *le numerose iniziative di alta formazione, che ha visto impegnati i ricercatori ILC in attività didattiche in corsi universitari, master, scuole estive, corsi di formazione professionale e in progetti di collaborazione con scuole per avvicinare gli studenti ai temi della Linguistica Computazionale e delle Digital Humanities.*

La tabella che segue riassume in cifre le attività dell'anno:

I risultati 2017 della ricerca dell'ILC	
Articoli in rivista	13
Capitoli di libri / contributi in volumi	6
Contributi in atti di convegno	31
Curatele	3
Direzione Scientifica di riviste	3
Supervisione di tesi di LM e di dottorato	10
Insegnamento corsi universitari	6
Insegnamento di Moduli in Scuole estive	3
Progetti di ricerca	21
Infrastrutture di ricerca	2
Riconoscimenti e premi	2

Per quanto concerne la valutazione dei risultati della ricerca, è da evidenziare che ben il 66% dei prodotti dell'ILC valutati nel Rapporto "Valutazione della Qualità della Ricerca 2011-2014 (VQR 2011-2014)", pubblicato dall'ANVUR - Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca, ha ricevuto un giudizio "eccellente" o "elevato".

Tra i principali elementi di criticità si segnalano in particolare: le difficoltà nel reperimento di fondi da destinare allo sviluppo di attività di ricerca su obiettivi strategici; l'eccessiva macchinosità delle pratiche amministrative; la difficoltà di programmazione delle attività di ricerca a medio-lungo termine (principalmente a causa della mancanza di fondi per la ricerca di base); l'impossibilità di pianificare il reclutamento di nuovo personale a medio-lungo termine, con le conseguenti problematicità nella gestione di personale precario, sotto-inquadrato e demotivato.

Pisa, febbraio 2018

Simonetta Montemagni
Direttrice dell'ILC-CNR

1 Introduzione

L'Istituto di Linguistica Computazionale (ILC) è un centro di riferimento, a livello nazionale e internazionale, nel settore della Linguistica Computazionale. L'Istituto afferisce al Dipartimento Scienze Umane e Sociali, Patrimonio Culturale (DSU) del Consiglio Nazionale delle Ricerche (CNR) e svolge attività di ricerca nei settori scientifici strategici della disciplina, oltre ad attività editoriali, di formazione e di trasferimento tecnologico.

Fin dalle origini, la missione e le attività di ricerca dell'ILC si collocano programmaticamente all'interno dell'area umanistica, in un rapporto di costante interazione interdisciplinare con competenze di base eterogenee, che vanno dalle varie anime della linguistica (formale, tipologica, cognitiva e applicata) all'informatica e alle infrastrutture digitali, dalla psicologia della cognizione allo studio dei sistemi complessi e alle neuroscienze. Ne sono testimonianza l'afferenza dell'Istituto al Dipartimento di Scienze Umane e Sociali, Patrimonio Culturale (DSU) e la sua collocazione nella mappa delle competenze disciplinari del CNR all'interno di un'area disciplinare di tipo umanistico (N), corrispondente al settore ERC SH4_6 e all'Area 10 della classificazione ANVUR.

La varietà delle linee di attività e dei progetti di ricerca rendono l'ILC una realtà unica nel panorama italiano e una delle poche a livello internazionale dove si affiancano: ricerche innovative nel settore delle Digital Humanities; attività volte alla definizione di standard e infrastrutture di ricerca distribuite; definizione di metodi e di tecniche avanzate per la ricerca e la gestione "intelligente" dell'informazione all'interno di basi documentali in linguaggio naturale disponibili sul Web o su Intranet locali; creazione di modelli computazionali dell'apprendimento linguistico in contesti ecologici di interazione comunicativa.

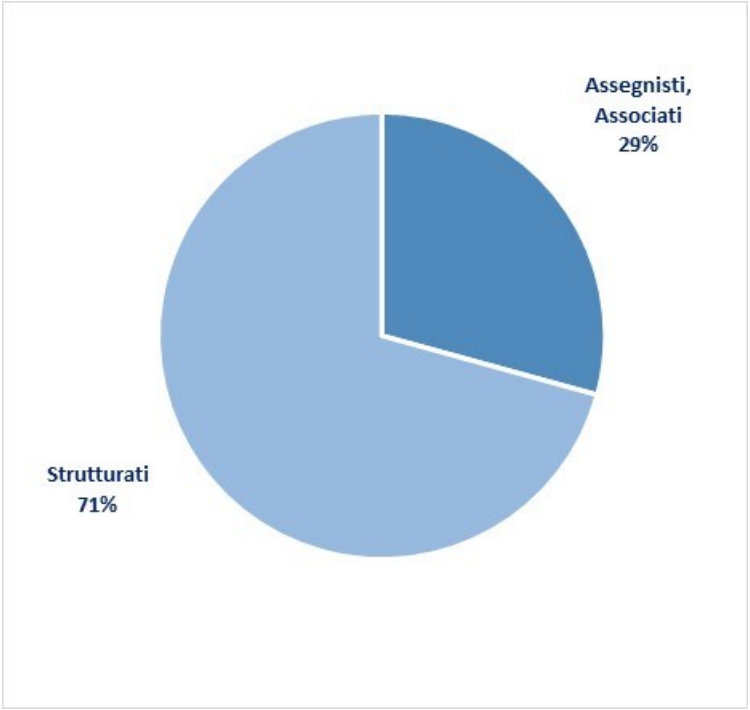
La ricerca all'ILC combina ricerca di base, con un investimento su temi di frontiera, e ricerca applicata, all'interno di un circolo virtuoso con ricadute significative sulla società e, in particolare, sul contesto socio-economico e culturale. Le attività sono condotte all'interno di una rete consolidata di collaborazioni a livello nazionale e internazionale con Istituti di ricerca, Università ed Enti Pubblici, così come con industrie e piccole e medie imprese, nell'ambito di numerosi progetti di ricerca.

2 L'ILC nel 2017: fatti e cifre

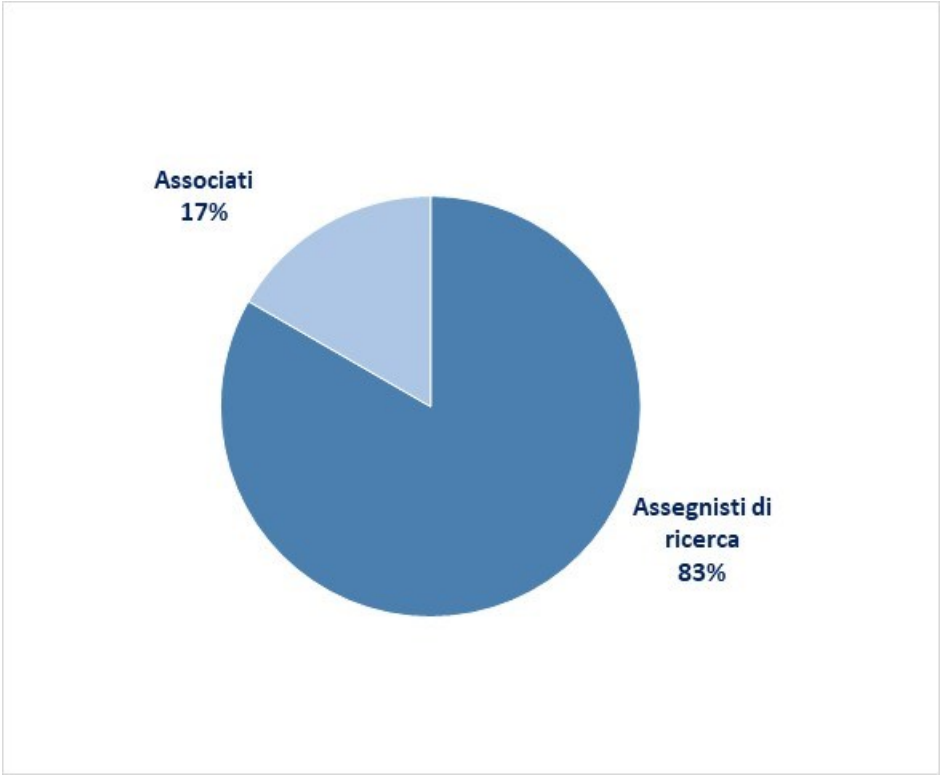
2.1 Personale

Il personale dell'ILC è composto da ricercatori, tecnologi e personale tecnico-amministrativo sia di supporto alla ricerca sia di area gestionale-amministrativa. Nel corso del 2017, il numero delle unità di personale ha subito delle variazioni: 5 unità con profilo di assegnista di ricerca e 1 unità con profilo di associato hanno terminato il loro contratto nel corso dell'anno.

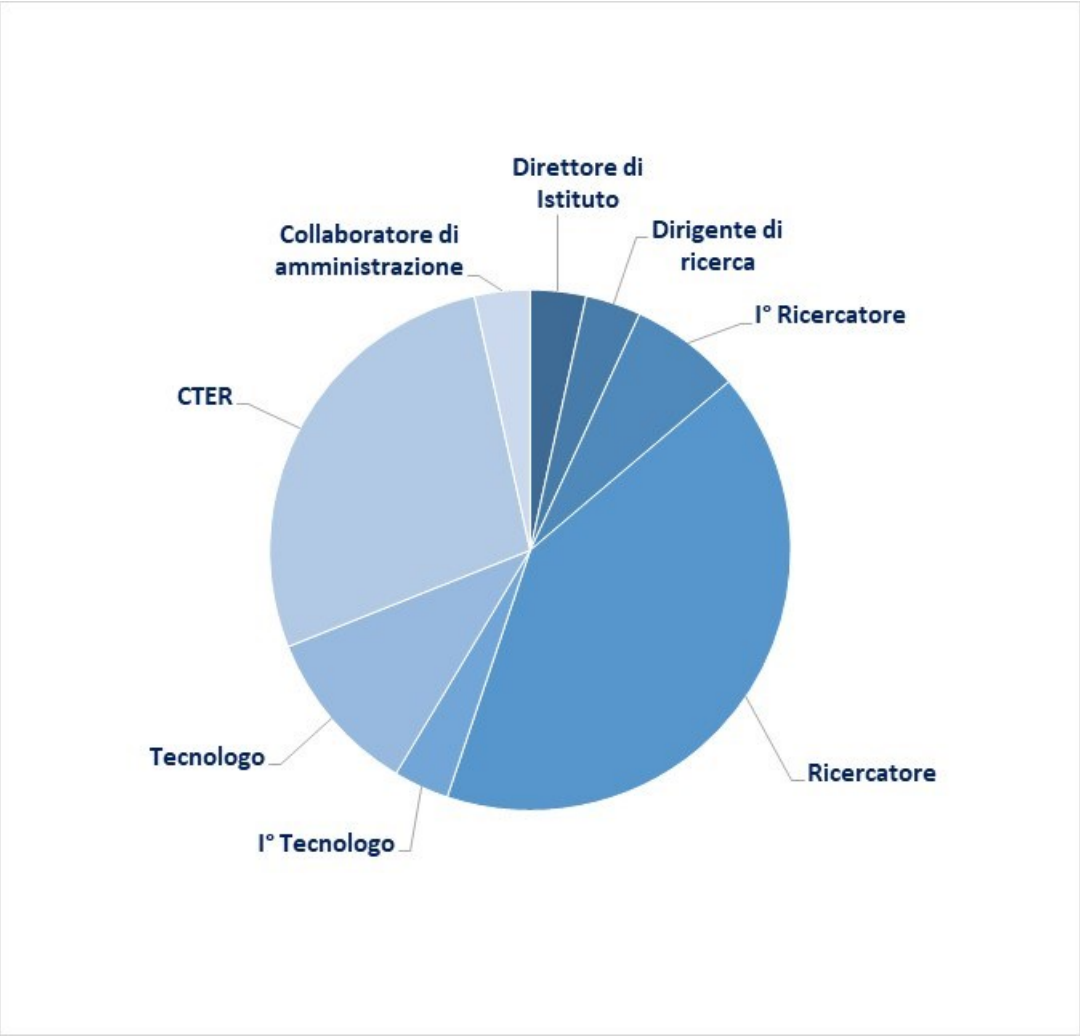
Oltre al Direttore dell'Istituto, lo staff è composto da 28 unità di personale strutturato (a tempo indeterminato e determinato) distribuite tra la sede di Pisa (26) e la sezione staccata di Genova (2), a cui si aggiungono 12 unità (11 a Pisa e 1 a Genova) tra assegnisti e personale associato: i dati fanno riferimento alla data del 31 dicembre 2017. Per maggiori dettagli v. appendice: *Personale ILC*.



RAPPORTO PERSONALE STRUTTURATO VS ALTRO PERSONALE



RIPARTIZIONE ALTRO PERSONALE



SUDDIVISIONE PERSONALE STRUTTURATO PER PROFILI E LIVELLI

2.2 Organizzazione interna

2.2.1 Responsabili e Organi di Governo

Direttore	Simonetta Montemagni
Consiglio di Istituto	Felice Dell'Orletta, Antonella Gadducci, Claudia Marzi, Monica Monachini, Simonetta Montemagni, Gabriella Pardelli, Vito Pirrelli
Responsabile amministrativo	Antonella Gadducci
Responsabile dei sistemi informativi	Alessandro Enea
Responsabile della sicurezza	Roberto Bartolini (referente per il collegamento con il Servizio di Prevenzione e Protezione – SPP e responsabile del Documento di Valutazione Rischi - DVR)
Responsabile della biblioteca	Gabriella Pardelli
Referente per la formazione	Claudia Marzi

2.2.2 Organizzazione interna

Ufficio Supporto Progetti	Paola Baroni, Antonella Gadducci, Eva Sassolini, Noemi Terreni
Comitato di autovalutazione e premiale	Monica Monachini, Simonetta Montemagni, Vito Pirrelli
Commissione comunicazione	Paola Baroni, Michela Carlino, Alessandro Enea, Simonetta Montemagni, Valeria Quochi, Claudia Soria
Comitato scientifico per i seminari	Federico Boschetti, Felice Dell'Orletta, Monica Monachini

Rappresentanti ILC all'interno dei Comitati dell'Area della Ricerca – CNR di Pisa:

Commissione Comunicazione	Paola Baroni, Claudia Soria
Commissione Relazioni Internazionali	Valeria Quochi
BRIGHT - La Notte dei Ricercatori in Toscana	Marcello Ferro
AREAPERTA - Comitato di Redazione	Emiliano Giovannetti

2.2.3 Organizzazione della ricerca

Nel 2017 è proseguito il processo di razionalizzazione e riorganizzazione delle attività di ricerca e sviluppo, in modo da organizzare le attività secondo quanto previsto dai nuovi regolamenti del CNR. La Direzione insieme agli organi di governo dell'istituto ha promosso una riflessione interna che ha coinvolto tutti i ricercatori e tecnologi per minimizzare le sovrapposizioni e massimizzare le sinergie tra aree di ricerca. È stato adottato un modello organizzativo "a matrice", in cui aree di competenza e progetti rappresentano rispettivamente gli assi verticali e orizzontali della matrice: le competenze maturate nell'ambito delle singole aree si combinano in modo vario, innovativo e sinergico all'interno dei progetti di ricerca. All'interno di questo assetto organizzativo, i laboratori e i gruppi di ricerca rivestono un ruolo centrale e strategico come luoghi di sviluppo di competenze e come spazio per la ricerca "curiosity-driven", mentre l'asse orizzontale della matrice coincide con la riorganizzazione delle attività in progetti, come richiesto dai regolamenti CNR.

Quattro le principali aree di attività attive:

- DH - Digital Humanities;
- LRI - Risorse, standard e infrastrutture di ricerca;
- MIND - Modelli (bio-)computazionali dell'uso linguistico;
- TAL - Trattamento automatico del linguaggio naturale ed estrazione di conoscenza;

AREE DI ATTIVITÀ e DESCRIZIONE	
DH **Digital Humanities**	La linea di attività DH intende coniugare le acquisizioni e conoscenze delle scienze informatiche con gli approcci metodologici e i modelli teorici dell'analisi testuale e della filologia del testo, contribuendo così alla trasformazione delle modalità di conservazione, fruizione, studio e pubblicazione dei documenti letterari, archivistici e bibliotecari, e offrendo nuove prospettive di indagine e condivisione. Le soluzioni tecnologiche messe a punto in questo ambito si integrano in un sistema "multi-modulare" a componenti indipendenti ma interconnessi, da cui le diverse metodologie di accesso, gestione, studio e revisione del testo possono trarre vantaggio e opportunità di interazione/integrazione.
LRI **Risorse, standard e infrastrutture di ricerca**	La linea di attività LRI si propone di facilitare la ricerca nel settore dell'ingegneria delle lingue e di ottimizzare il ciclo di produzione delle risorse linguistiche. Questo chiama in causa l'adozione di standard, lo scambio di buone pratiche per l'interoperabilità, il riciclo e il riutilizzo dei risultati disponibili in termini di dati e strumenti. Le attività principali di quest'area riguardano la definizione di modelli per la creazione, rappresentazione, estensione e mantenimento di lessici computazionali, repertori terminologici e ontologici, corpora e tecnologie linguistiche, ma anche lo sviluppo di soluzioni tecnologiche per la creazione di un'infrastruttura di ricerca distribuita e cooperativa, volta a stabilire nuove funzionalità di accesso, interoperabilità e condivisione di risorse e strumenti linguistici.
MIND **Modelli (bio-)computazionali dell'uso linguistico**	La linea di attività MIND è dedicata allo studio dei fattori che governano i processi di comprensione, produzione, apprendimento e variazione di una lingua e le interazioni dinamiche tra di essi. In particolare, promuove lo sviluppo di modelli teorici dell'uso linguistico e la loro verifica empirica attraverso: l'uso di metodi probabilistici per lo studio di corpora, lessici e basi di dati; simulazioni computazionali; lo studio di evidenza linguistica di natura sperimentale, clinica e acquisizionale. Gli obiettivi di questa linea di attività sono perseguiti coniugando metodologie di rappresentazione formale e modellazione simbolica con i metodi, i dati e gli strumenti di indagine di settori disciplinari più orientati all'analisi dell'uso linguistico in contesti finalizzati e controllati, quali la psico- e neuro-linguistica, la sociolinguistica e la glottodidattica.
TAL **Trattamento automatico del linguaggio naturale ed estrazione di conoscenza**	La linea di attività TAL si occupa di sviluppare metodi e tecniche per consentire l'accesso automatico al contenuto di un testo, rendendo espliciti quei nuclei di conoscenza linguistica che rispondono a una vasta gamma di bisogni informativi dei parlanti, dall'accesso su base semantica al contenuto testuale, alla valutazione della struttura del testo come indicatore della sua accessibilità ed efficacia comunicativa. Le soluzioni tecnologiche sviluppate in questo ambito rispondono alle necessità di ricerca e gestione "intelligente" dell'informazione contenuta all'interno di grandi basi documentali in continua evoluzione, e sono propedeutiche allo sviluppo di un ampio ventaglio di applicativi commerciali. TAL si focalizza sullo studio di modelli probabilistici del linguaggio e sullo sviluppo di algoritmi di apprendimento automatico per l'annotazione del testo e l'estrazione di conoscenza.

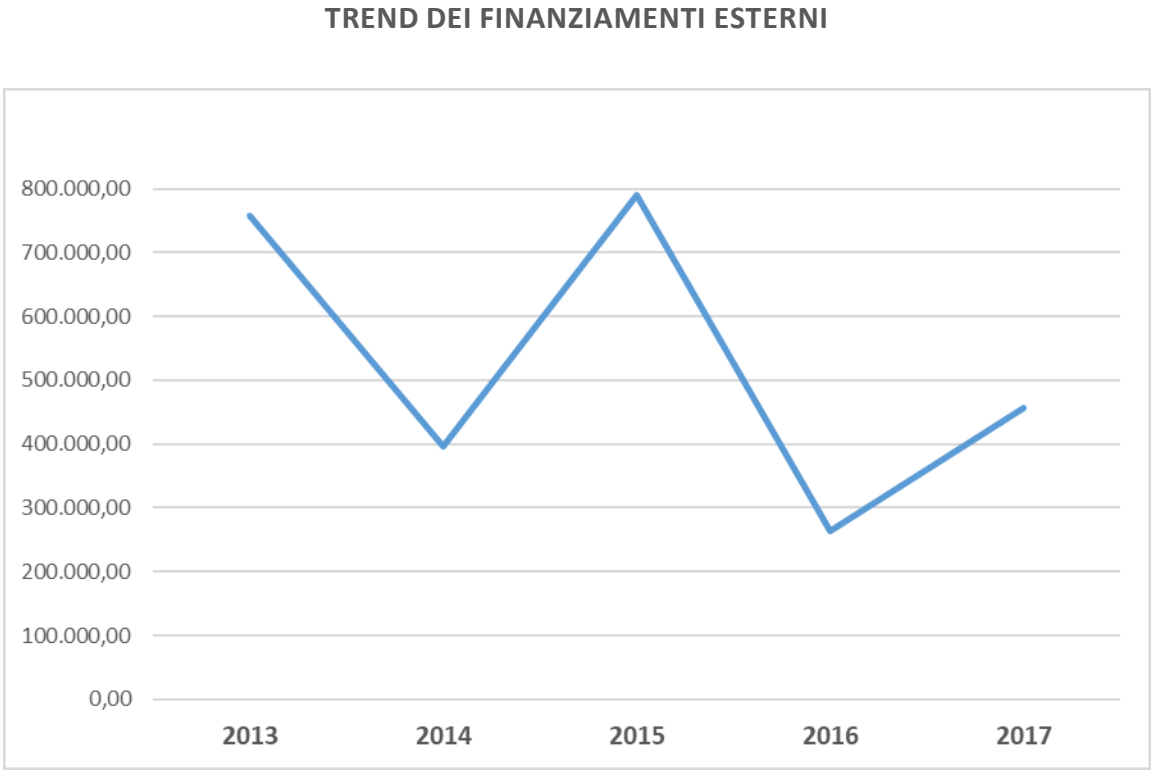
A queste aree si affianca una quinta linea "ART – Attività di Ricerca Trasversali" riguardante il perfezionamento linguistico, statistico e informatico nonché, nel campo della gestione della ricerca e della conoscenza dei sistemi di ricerca europei e internazionali, la valorizzazione dei risultati della ricerca e la protezione della proprietà intellettuale.

LABORATORI e GRUPPI di RICERCA	
ComPhys Lab	*Obiettivo*: Progettazione e sviluppo di modelli (bio-)computazionali del comportamento linguistico con l'obiettivo di comprendere e spiegare le relazioni tra competenza grammaticale, uso linguistico e correlati neuro- e psico-linguistici della comunicazione verbale e dei suoi disturbi. *Responsabile:* Vito Pirrelli *Sito:* www.comphyslab.it
CoPhiLab	*Obiettivo*: Formalizzazione delle entità e delle relazioni nel dominio della filologia collaborativa; creazione di risorse digitali; progettazione e sviluppo di componenti software, in particolare per le lingue classiche. *Responsabile:* Federico Boschetti *Sito:* http://cophilab.ilc.cnr.it:8080/CoPhiLabPortal

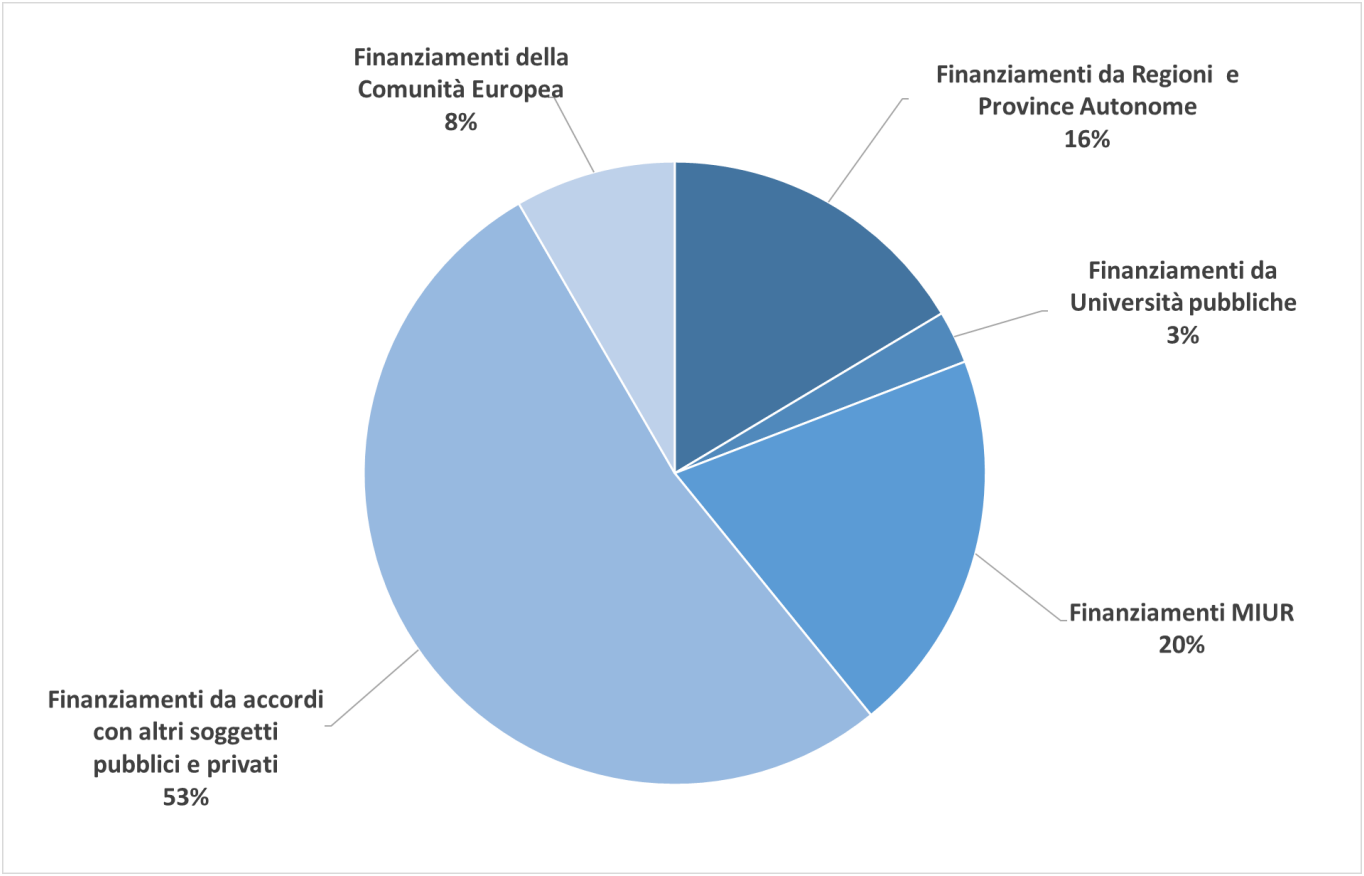
ItaliaNLP Lab	*Obiettivo*: Progettazione e sviluppo di modelli, metodi, algoritmi e tecnologie per il Trattamento Automatico del Linguaggio e per l'estrazione di conoscenza, con particolare attenzione alla lingua italiana. Principali linee di attività: annotazione linguistica multi-livello di testi; estrazione di conoscenza da collezioni documentali; sviluppo di prototipi applicativi. *Responsabile:* Felice Dell'Orletta *Sito:* www.italianlp.it
LaRI Group	*Obiettivo*: Il gruppo "Risorse e Infrastrutture Linguistiche" si propone di facilitare la ricerca nel settore dell'ingegneria delle lingue e di ottimizzare il ciclo di produzione delle risorse linguistiche. *Responsabile:* Monica Monachini *Sito:* http://lari.ilc.cnr.it
Literary Computing Group	*Obiettivo*: Il gruppo "Literary Computing" si occupa di modelli testuali, ontologie dei testi, traduzione assistita da computer, progettazione e sviluppo di applicazioni per il "Literary Computing". *Responsabile:* Emiliano Giovannetti *Sito:* http://licolab.ilc.cnr.it

2.3 Finanziamenti

Rispetto all'anno precedente, nel 2017 si registra un'inversione di tendenza per quanto riguarda la capacità di attrarre risorse esterne per la ricerca, sia su bandi competitivi (regionali, nazionali e internazionali) sia attraverso contratti con committenza esterna (pubblica e privata) e interna al CNR. Nel 2017, infatti, l'importo dei finanziamenti esterni risulta quasi raddoppiato rispetto all'anno precedente.



RIPARTIZIONE DELLE ENTRATE - ANNO 2017



2.4 Progetti

2.4.1 Progetti europei

Progetti europei coordinati dall'ILC

DLDP - The Digital Language Diversity Project
Progetto triennale finanziato dalla Comunità Europea nell'ambito del programma Erasmus+.
Responsabile Scientifico del Progetto e Responsabile Scientifico Unità di Ricerca ILC: Claudia Soria
Il progetto Erasmus + (Attività KA2, Azione chiave 2 Partenariati Strategici) si propone di far avanzare la sostenibilità delle lingue regionali e minoritarie europee nel mondo digitale, incrementando nei loro parlanti la conoscenza e le abilità per creare e condividere contenuti sulle apparecchiature digitali usando le loro lingue minoritarie.
Nel corso del 2017, oltre alle attività di coordinamento del progetto, sono state svolte attività volte a: l'elaborazione di una scala per la valutazione della vitalità linguistica digitale; la progettazione e la realizzazione di questionari per la raccolta di informazioni relative all'uso e usabilità delle lingue regionali e minoritarie su Internet e per mezzo di strumenti digitali; l'analisi dei dati; la progettazione della struttura e del contenuto di un corso on-line rivolto a parlanti di lingue regionali e minoritarie per promuovere la presenza e l'uso di queste lingue in ambito digitale; la preparazione dei report di progetto; le attività di networking e di disseminazione.
http://www.dldp.eu

Progetti europei ai quali l'ILC partecipa come partner

PARTHENOS - Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies
Progetto Horizon 2020 di durata quadriennale finanziato dalla Comunità Europea.
Responsabile Scientifico Unità di Ricerca ILC: Monica Monachini
Il progetto mira a rafforzare la coesione della ricerca nell'ampio settore relativo a studi linguistici, scienze umane, patrimonio culturale, storia, archeologia e settori collegati, attraverso un cluster tematico di infrastrutture di ricerca europee, integrando iniziative, infrastrutture elettroniche e altre infrastrutture di punta, e costruendo ponti tra campi diversi, sebbene strettamente interconnessi. Il progetto raggiungerà questo obiettivo attraverso la definizione e il supporto di standard comuni, il coordinamento di attività congiunte, l'armonizzazione della definizione e dell'implementazione di policy e lo sviluppo di servizi comuni e di soluzioni condivise agli stessi problemi. ILC, in particolare, contribuisce alla creazione di strumenti per facilitare le attività di ricerca specifiche nel riutilizzo dei dati.
http://www.parthenos-project.eu

Iniziative europee in cui l'ILC è sottocontraente

ELRC - European Language Resource Coordination
Azione finanziata nell'ambito del CEF SMART 2014/1074 Programme, contratto Ref. Ares(2014)2275366, dalla Commissione Europea.
Responsabile Scientifico Unità di Ricerca ILC: Simonetta Montemagni
L'azione "European Language Resources Coordination" (ELRC) si colloca all'interno del programma 'Connecting Europe Facility' (CEF): finanziata dalla Commissione Europea, si propone di migliorare e di estendere la copertura e la qualità del sistema di traduzione automatica sviluppato dalla DG Translation in vista della sua integrazione nei servizi pubblici online di tutti i Paesi europei. Tale obiettivo è perseguito attraverso l'individuazione e la raccolta di risorse linguistiche multilingui delle amministrazioni e istituzioni governative in tutti i 30 Paesi europei che partecipano al programma CEF. Nell'ambito di questa azione, l'ILC ha il compito di supportare il processo di individuazione e raccolta delle risorse linguistiche (lessici e corpora testuali, multi- e mono-lingui) in Italia.
Nell'ambito del progetto, Simonetta Montemagni è il referente tecnologico nazionale per l'Italia. I referenti tecnologici nazionali dell'azione ELRC costituiscono il Language Resources Board, l'organo di governo di ELRC.
http://www.lr-coordination.eu

Altre iniziative internazionali in cui l'ILC è partner

Progetto DiTMAO - Dictionary of Old Occitan medico-botanical terminology
Progetto internazionale finanziato dalla DFG (Deutsche Forschungsgemeinschaft)
Responsabile Scientifico Unità di Ricerca ILC: Emiliano Giovannetti
Il progetto DiTMAO mira alla creazione di un sistema di lessicografica digitale contenente termini medico-botanici in occitano antico derivanti da testi medievali in latino e in caratteri ebraici. In particolare, l'attività dell'ILC è dedicata alla definizione di modelli lessico-ontologici e al supporto alla progettazione dell'editor LexΩ.
https://www.uni-goettingen.de/en/ditmao/487498.html

2.4.2 Progetti nazionali e regionali

Progetti finanziati a livello nazionale

CHROME - Cultural Heritage Resources Orienting Multimodal Experiences
Progetto triennale (febbraio 2017 - febbraio 2020) finanziato dal MIUR (PRIN 2015) e sviluppato dall'ILC insieme all'Università degli Studi di Napoli "Federico II", all'Università degli Studi "Roma Tre", all'Università degli Studi di Salerno e all'Istituto di Scienze Applicate e Sistemi Intelligenti "Eduardo Caianiello" del Consiglio Nazionale delle Ricerche (ISASI-CNR).
L'obiettivo principale del progetto è l'elaborazione di una metodologia per raccogliere, rappresentare e analizzare dati multimodali relativi ai beni culturali e per presentarli attraverso agenti artificiali il cui comportamento è ispirato da un'analisi accurata di guide esperte, curatori di musei e tour operator. Il progetto, sviluppato congiuntamente da umanisti e informatici, permetterà di modellare il comportamento che i custodi adottano quando presentano il

patrimonio culturale. Tale modello sarà utilizzato per controllare un robot umanoide progettato per seguire simili strategie di presentazione.
http://www.chrome.unina.it

CITTÀ EDUCANTE
Progetto PON Ricerca e Competitività 2007-2013 – MIUR, Linea di intervento CTN-TSC - Tecnologie per le Smart Communities.
Responsabile Scientifico Unità di Ricerca ILC: Lucia Marconi
Il progetto intende contribuire a generare una "città educante", ossia un modello in cui il momento educativo è basato sulla reciprocità dei soggetti coinvolti: chi educa è anch'egli educato e il suo sapere prende forma nell'atto dell'educazione. Educare non è solo formare, è anche costruire identità e futuro. Ciò in un ambiente dove ognuno possa contribuire a progettare e realizzare un maggiore benessere comune. Il progetto individua tre aree tematiche e mira a:
- scuola/educazione: offrire alla scuola di ogni grado di istruzione un insieme di saperi, strategie e applicazioni tecnologiche che innovino l'approccio educativo. Un metodo che aiuti ad essere cittadini attivi, accoglienti e consapevoli. Rivolto a bambini e adulti, come soggetti di life-long learning, nei processi ageing e nell'accoglienza della diversità.
- società: sviluppare nuove connessioni tra scuola, aziende e territorio.
- tecnologia: creare nuove piattaforme, servizi e applicazioni ICT grazie a specifiche attività di ricerca nei temi del cloud computing, del collaborative sourcing, dei social network, dell'analisi automatica di testi, video e dati 3D, della big data analysis, delle interfacce naturali e interattive, degli ausili robotici, sensoriali e pervasivi, dell'apprendimento automatico e dei sistemi di search. Nel corso del 2017 sono state sviluppate metodologie di estrazione e comparazione di informazioni linguistiche dalle produzioni scritte degli studenti prima e dopo un evento specifico.
http://www.cittaeducante.it

TALMUD - Traduzione Talmud Babilonese
Progetto quinquennale finanziato dal MIUR.
Responsabile Scientifico Unità di Ricerca ILC: Emiliano Giovannetti
Il progetto ha come obiettivo la traduzione in lingua italiana del Talmud Babilonese. La traduzione commentata, con testo originale a fronte in lingua ebraica e aramaica, è realizzata da un team di traduzione formato da circa 70 studiosi da tutte le parti del mondo tramite l'utilizzo di Traduco, una piattaforma Web collaborativa sviluppata dall'ILC che include strumenti editoriali avanzati e componenti per il trattamento del testo e della conoscenza basati su metodi e tecniche della linguistica computazionale.
http://www.talmud.it

TOTUS MUNDUS - The World is our Home
Progetto nazionale finanziato dall'Università La Sapienza di Roma.

Responsabile Scientifico Unità di Ricerca ILC: Emiliano Giovannetti
Il progetto è finalizzato alla creazione di una piattaforma digitale mediante la quale sarà possibile accedere all'edizione annotata da Pasquale D'Elia dell'Atlante mondiale prodotto nel 1602 a Pechino da Matteo Ricci e Li Zhizao. A tale scopo è stata attivata una sinergia tra l'Istituto di Informatica e Telematica del Consiglio Nazionale delle Ricerche - IIT-CNR (capofila del progetto) e l'Università La Sapienza. L'ILC, nell'ambito di un sottoprogetto, collabora in particolare per la definizione di modelli lessico-ontologici e per il supporto alla creazione di una risorsa lessico-ontologica bilingue cinese-italiano.
http://www.totusmundus.it

VOCI DELLA GRANDE GUERRA
Progetto di 18 mesi (2016-2018) finanziato dalla Presidenza Consiglio dei Ministri nell'ambito dell'avviso pubblico per la Commemorazione del Centenario della Grande Guerra.
Responsabile Scientifico Unità di Ricerca ILC: Simonetta Montemagni

Il progetto mira a preservare e diffondere le memorie della Prima Guerra Mondiale attraverso la costruzione e la pubblicazione online di un corpus digitale di testi (lettere, bollettini di guerra, giornali, diari, ecc.) rappresentativi delle diverse modalità di sentire e raccontare l'Italia in guerra da parte dei suoi protagonisti. Questo patrimonio culturale collettivo è valorizzato grazie all'utilizzo di tecniche avanzate di linguistica computazionale, Web semantico e visualizzazione dell'informazione. L'archivio digitale, che comprende documenti storici appartenenti a un ampio spettro di registri e varietà linguistiche, permette di ricostruire la polifonia delle lingue dell'Italia in guerra: la voce ufficiale della propaganda e la voce dei soldati, la voce dei giornali e la voce delle lettere, la voce delle élite degli intellettuali e la voce popolare, la voce del consenso e la voce del dissenso. Un'interfaccia online dotata di funzionalità di ricerca innovative consente di consultare i testi tramite chiavi di ricerca semantiche e, ove possibile, di visualizzare le immagini delle edizioni cartacee originarie. L'ILC, in particolare, si occupa dell'acquisizione digitale del testo mediante tecniche innovative di OCR, della supervisione al lavoro di correzione manuale dei testi digitalizzati, nonché della specializzazione di componenti per l'annotazione linguistica del testo e per l'estrazione di informazione nella forma di eventi e georeferenziazione dei luoghi della guerra.
http://www.vocidellagrandeguerra.it

Progetti regionali

PERFORMA ARCO CNR - Personalizzazione di pERcorsi FORMativi Avanzati
Progetto POR FSE 2014-2020 Asse A – Occupazione, Avviso pubblico per progetti congiunti di alta formazione attraverso l'attivazione di assegni di ricerca (anno 2017), Linea A
Il progetto si propone di definire nuove metodologie basate sull'uso di tecnologie del linguaggio per la creazione di percorsi di formazione a distanza personalizzabili a diversi livelli. Il risultato sarà lo sviluppo di un sistema di profilazione dei materiali didattici fruibili all'interno di piattaforme di e-learning in grado di valutarne l'adeguatezza rispetto alle competenze linguistiche dei corsisti e rispetto al dispositivo di lettura (tablet, pc, smartphone) dal quale il corso può essere fruito. Per ciascuna delle due dimensioni, il grado di adeguatezza dei materiali sarà valutato sia dal punto di vista della complessità linguistica dei testi di studio sia dal punto di vista dei contenuti proposti. Obiettivi del progetto saranno la definizione e l'adattamento di nuovi algoritmi basati su strumenti di trattamento automatico della lingua per la valutazione della complessità del testo e dei contenuti proposti nei corsi rispetto alle competenze degli utenti che usufruiranno dei corsi e alle modalità di fruizione offerte dai vari dispositivi.

UBIMOL - UBIquitous Massive Open Learning
Progetto biennale (2017-2019) finanziato dalla Regione Toscana (POR FESR 2014-2020 - Bando 2. Progetti di ricerca e sviluppo delle PMI).
Si tratta di un progetto sviluppato dall'ILC insieme alle società M.E.T.A (Capofila del Progetto), 01Sistemi, Viditrust e Persafe e al CoLing Lab del Dipartimento di Filologia, Letteratura e Linguistica dell'Università di Pisa. Il progetto è finalizzato allo sviluppo di piattaforme di e-learning arricchite con tecnologie innovative capaci di offrire corsi personalizzati rispetto al livello delle competenze linguistiche specifiche per ciascun profilo di discente. Tali piattaforme, sfruttando applicazioni basate sul trattamento automatico del linguaggio, permetteranno percorsi formativi guidati da processi di auto-valutazione in grado di valutare le competenze acquisite durante i corsi.

SMART NEWS - Social sensing for breaking news
Progetto biennale (2016-2018) finanziato dalla Regione Toscana nell'ambito del Bando FAR FAS 2014 – Linea A.
Responsabile Scientifico Unità di Ricerca ILC: Felice Dell'Orletta
Il progetto mira a realizzare un tool per la gestione delle breaking news che aiuti i giornalisti nelle diverse fasi di questo processo: dall'individuazione della breaking news, alla raccolta di informazioni, alla scrittura. Il tool "ascolterà" i social media e sarà in grado di individuare automaticamente le breaking news. A tale scopo sono messe a punto avanzate metodologie per il trattamento automatico della lingua in grado di: individuare ed estrarre dai documenti di social media entità rilevanti per la descrizione delle notizie; classificare i documenti in base alle informazioni estratte; identificare i testimoni di eventi; individuare il sentimento prodotto dai documenti; riepilogare automaticamente diversi documenti al fine di scrivere articoli di giornale.
Nel corso del 2017 l'ILC ha sviluppato algoritmi per l'analisi dei testi estratti dai social media, in particolare Twitter, e sistemi per: l'annotazione linguistica dei testi; l'estrazione dell'informazione; la classificazione dei tweet;

l'identificazione dei testimoni; la classificazione del sentimento del testo dei tweet e la creazione di riassunti automatici da collezioni di tweet.
http://www.smart-news.it

Progetti CNR

CLAVIUS - Clavius on the Web
Progetto nazionale.
Responsabile Scientifico Unità di Ricerca ILC: Simone Marchi
L'obiettivo del progetto è la conservazione e la valorizzazione di una parte dei manoscritti conservati nell'Archivio Storico della Pontificia Università Gregoriana. Il progetto prende in esame alcuni manoscritti relativi a Christophorus Clavius (1538-1612), matematico e astronomo gesuita. Tali manoscritti sono stati digitalizzati, trascritti, tradotti e analizzati dai punti di vista linguistico, lessicale e semantico. La terminologia e le entità di dominio individuate nel testo sono state strutturate in un lessico e un'ontologia e sono state collegate a risorse già disponibili in rete secondo i principi dei Linked Data. Nel corso del 2017 è stato finalizzato lo sviluppo della piattaforma per lo studio della corrispondenza di Christophorus Clavius con astronomi come Galileo Galilei, Tycho Brahe, ecc. In particolare, sono terminate la progettazione e implementazione delle funzionalità di ricerca full-text e di ricerca concettuale basata sul lessico costruito nell'ambito del progetto.
http://claviusontheweb.it

NINFA - iNtelligent Integrated Network For Aged people
Progetto di Interesse strategico CNR "Invecchiamento": innovazioni tecnologiche e molecolari per un miglioramento della salute dell'anziano.
Referente Scientifico Unità di Ricerca ILC: Lucia Marconi
Un sotto-progetto del progetto CNR nazionale "Invecchiamento" che intende analizzare e proporre soluzioni ai problemi relativi all'applicazione delle Tecnologie dell'Informazione e della Comunicazione (TIC) per un invecchiamento attivo e al monitoraggio di danni cognitivi. Il progetto affronta l'accettabilità di nuove TIC, la valutazione della condizione di benessere degli utenti e la gestione di eventi critici con un impatto minimo sugli utenti durante la fornitura di servizi domiciliari, attraverso soluzioni differenti, calibrate principalmente su aspetti riguardanti gli utenti finali. ILC, come partner del progetto Ninfa e nell'ambito del WP3 "Analisi e test d'implementazione di deficit cognitivo attraverso l'analisi del linguaggio", aveva come obiettivo quello di realizzare una base di dati strutturata costituita da un corpus di registrazioni e di trascrizioni dei singoli soggetti anziani monitorati nel tempo. Nel 2017 sono proseguite le analisi su alcune categorie grammaticali, in particolare sostantivi e verbi sui soggetti monitorati che hanno effettuato test di eloquio spontaneo.

Progetto ItaliaNLP – WAFI
Responsabile Scientifico Unità di Ricerca ILC: Felice Dell'Orletta
Progetto di collaborazione con il laboratorio WAFI dell'Istituto di Informatica e Telematica del CNR di Pisa (IIT-CNR) che mira a potenziare il sistema di analisi dei testi derivanti da social media nell'ambito della Cyber Intelligence. Nello specifico, l'ILC svolge le seguenti attività: consulenza su approcci e metodologie di analisi del testo; integrazione dello strumento T2K nella piattaforma di Cyber-Intelligence WAFI; sviluppo di prototipi per finalità specifiche, ad esempio per la rilevazione di hate-speech nei contenuti postati dagli utenti; realizzazione congiunta di pubblicazioni scientifiche.

SM@RTINFRA-SSHCH - Infrastrutture integrate intelligenti per l'ecosistema dei dati delle scienze sociali, umane e del patrimonio culturale
Progetto nazionale di durata triennale finanziato con il "Fondo Ordinario per gli Enti di Ricerca - Quota finalizzata al Finanziamento Premiale di Specifici Programmi e Progetti" del MIUR.
Responsabile Scientifico Unità di Ricerca ILC: Monica Monachini
Il progetto premiale mira a creare una struttura di governance di coordinamento nazionale dei nodi italiani delle infrastrutture di ricerca (RI) europee di Social Sciences and Humanities, Cultural Heritage (SSHCH). Il risultato finale prevede il potenziamento delle RI nazionali e della partecipazione dell'Italia come membro agli ERIC già costituiti

(CLARIN, ESS, SHARE) o in fase di costituzione (DARIAH). Le principali linee di attività sono: formazione per lo sviluppo delle competenze; ricerca nel settore delle tecnologie abilitanti fondamentali (Key Enabling Technologies - KETs) per l'avanzamento delle infrastrutture; networking, trasferimento tecnologico; diffusione dei risultati.

Altri progetti

COMMERCE NUMÉRIQUE

Progetto coordinato e finanziato dell'Università di Pisa, in collaborazione con ILC-CNR, Università di Perugia e Fondazione Caetani.

Obiettivo del progetto è la digitalizzazione della rivista letteraria francese "Commerce". Sono state condotte attività di acquisizione e ottimizzazione delle immagini digitali, di acquisizione del testo tramite OCR a partire dalle immagini digitalizzate e di supervisione al lavoro di correzione manuale dei testi digitalizzati.

ENCORE - ENgaging Content Object for Reuse and Exploitation of cultural resources

Progetto biennale (2017-2019) in collaborazione con le società M.E.T.A. Srl, F2 Glocal Innovation e l'Università degli Studi di Salerno.

*Il progetto mira allo sviluppo di approcci innovativi per la produzione, l'accesso e il riuso di "*cultural resources" *offrendo all'utente una narrativa personalizzata per l'accesso alla cultura e al turismo culturale*.

GREEK STUDIES IN XVTH CENTURY EUROPE

Progetto Marie Curie, coordinato da Paola Tomè (Università Ca' Foscari Venezia e University of Oxford, finanziato con borsa Marie Curie). ILC ha dato supporto esterno sul piano scientifico.

Obiettivo del progetto è lo studio del ritorno del greco in Occidente, grazie all'attività degli umanisti del Quindicesimo secolo. In collaborazione con U. Springmann (Ludwig Maximilians Universität München) ed Edoardo Bighin (WikiMedia), si è proceduto alla digitalizzazione del De Orthographia di G. Tortelli, Venezia 1471 tramite OCR e alla messa online su piattaforma WikiSource. Tra le attività di disseminazione è stata prevista anche la partecipazione a convegni nazionali e giornate di studio in licei classici della provincia di Venezia.

IL TESORO DELL'ILC

Progetto interno dell'ILC, coordinato da Manuela Sassi, in collaborazione con Eva Sassolini e Sebastiana Cucurullo

Il progetto è finalizzato al recupero, alla conversione, alla conservazione a lungo termine e alla valorizzazione dell'Archivio Testuale dell'ILC, che raccoglie importanti risorse testuali patrimonio dell'Istituto salvate in formati obsoleti (ASCII, ODBC, TCL ASCII). L'avvio del progetto ha portato alla realizzazione di diverse collaborazioni scientifiche per il recupero, la conservazione e la divulgazione degli archivi recuperati.

MUSEO VIRTUALE DELLA MUSICA BELLININRETE

Il progetto è basato su una collaborazione interistituzionale tra Comune di Catania, Istituto di Scienze e Tecnologie della Cognizione (ISTC) del CNR, Università degli Studi di Catania per il tramite del Dipartimento di Scienze Umanistiche e Fondazione Bellini. Nell'ambito del progetto l'ILC ha stipulato un accordo di collaborazione con ISTC-CNR. Responsabile Scientifico Unità di Ricerca ILC: Emiliano Giovannetti.

Obiettivo del progetto è la digitalizzazione dei documenti belliniani custoditi nel Museo Civico belliniano di Catania. L'ILC ha provveduto alla definizione di schemi di codifica del testo e alla consultazione di edizioni digitali di lettere manoscritte.

2.5 Collaborazioni scientifiche

2.5.1 Accordi Bilaterali

- **Sidi Mohamed Ben Abdellah University - MAROCCO**
  L'Oceano Universale: un sistema semi-automatico per la classificazione morfologica, sintattica e semantica dei verbi trilitteri arabi (accordo CNR/CNRST anni 2016-2017)
  Acquisizione di informazione lessicale araba, morfologia, trattamento automatico della lingua araba. Nel corso del 2017 le attività si sono incentrate sullo studio dell'interazione tra frequenza e regolarità ed emergenza della struttura morfologica in compiti di acquisizione lessicale di forme verbali in lingua araba.
  Responsabili: Vito Pirrelli (ILC), Mohammed El Mohajir

2.5.2 Accordi e Convenzioni

- **Accademia della Crusca**
  Sviluppo di funzionalità di analisi e ricerca avanzate (DBT-like), per l'indicizzazione e l'interrogazione del testo. Nell'ambito del Progetto PRIN coordinato dall'Accademia della Crusca "Corpus di riferimento per un Nuovo Vocabolario dell'Italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazioni" è stato implementato un modello di codifica per la conversione a XML TEI dell'intero corpus e un sistema di consultazione dei testi.
  Responsabile ILC: Eva Sassolini
- **Associazione NeuroCare Onlus**
  Collaborazione per lo sviluppo di sistemi per il monitoraggio in ambito neuro-linguistico e neuro-cognitivo.
  Responsabile ILC: Vito Pirrelli
- **Austrian Academy of Sciences** (Wien, Austria)
  Collaborazione in materia di "Corpus linguistics and Theoretical and computational modelling of Morphology".
  Responsabile ILC: Vito Pirrelli
- **Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI)**
  Raccolta di risorse linguistiche italiane che sono rilevanti per lo European Language Resource Coordination (ELRC).
  Responsabile ILC: Simonetta Montemagni
- **Diocesi di Alba (CN)**
  Progettazione e sviluppo di un modulo per una piattaforma web di filologia digitale atto alla fruizione dell'edizione digitale del "Rotulo di San Teobaldo".
  Responsabili ILC: Simone Marchi
- **Fondazione Maria Bianca Corno**
  Collaborazione per l'analisi testuale di diari clinici.
  Responsabile ILC: Federico Boschetti
- **European Language Resources Association - ELRA**
  Gestione, personalizzazione, aggiornamento e documentazione di risorse linguistiche e attività di divulgazione scientifica.
  Nel 2017 la prosecuzione del progetto CNR basato su un accordo bilaterale tra ILC ed ELRA ha avuto lo scopo di formalizzare, sviluppare e consolidare la già esistente cooperazione tra le due parti estendendone le aree e le modalità e promuovendone le azioni comuni in vari settori, con attenzione rivolta in particolare al mantenimento, alla "customizzazione", all'aggiornamento e alla produzione della documentazione delle risorse linguistiche e alle attività di disseminazione scientifica, come ad esempio l'organizzazione di workshop/conferenze e la gestione di attività editoriali.
  Responsabili ILC: Nicoletta Calzolari Zamorani, Sara Goggi, Monica Monachini, Claudia Soria

- **Istituto di Fisiologia Clinica (IFC-CNR)**
  Collaborazione nei settori della adeguatezza ed efficacia pragmatica della comunicazione verbale e dei disturbi evolutivi della comunicazione verbale e non verbale.
  Responsabile ILC*:* Vito Pirrelli

- **Istituto di Informatica e Telematica - Consiglio Nazionale delle Ricerche (IIT-CNR)**
  Attività di Social Media Intelligence, Cyber-Intelligence.
  Responsabile ILC: Felice Dell'Orletta

- **Istituto di Informatica e Telematica - Consiglio Nazionale delle Ricerche (IIT-CNR) e Università degli Studi di Roma "La Sapienza"**
  Costruzione di una risorsa lessicale per l'accesso al testo della Mappa di Ricci e alla relativa traduzione di Pasquale D'Elia secondo criteri lessico-semantici.
  Responsabile ILC: Emiliano Giovannetti

- **Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" - Consiglio Nazionale delle Ricerche (ISTI-CNR)**
  Meta Opac Pisano (MOP): aggiornamento, mantenimento e miglioramento dell'accesso ai cataloghi delle biblioteche partecipanti.
  Responsabile ILC: Monica Monachini

- **Istituto di Scienze e Tecnologie della Cognizione (ISTC-CNR)**
  Nell'ambito del Progetto BellinInRete, digitalizzazione dei documenti belliniani custoditi nel Museo Civico belliniano di Catania.
  Responsabile ILC: Emiliano Giovannetti

- **Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa (INDIRE)**
  Studio e sviluppo prototipale di applicazioni di gestione della conoscenza nel settore documentale educativo basate su tecnologie di trattamento automatico del linguaggio.
  Responsabile ILC: Felice Dell'Orletta

- **Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa (INDIRE) - Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG-CNR)**
  Integrazione delle competenze scientifiche e delle risorse intellettuali e tecnico-strumentali esistenti presso INDIRE, ILC e ITTIG sul tema dello sviluppo e dell'implementazione di tecnologie della conoscenza per l'innovazione nel settore educativo e formativo.
  Responsabili ILC: Simonetta Montemagni

- **Liceo Classico di San Marco dei Cavoti, Università del Sannio - Dip. di Scienze e Tecnologie** e **Associazione culturale "Provenza…Mino"**
  Collaborazione nell'ambito del Piano Nazionale Scuola Digitale per il Progetto Bibliothéke, creatività e innovazione tra gli "scaffali".
  Responsabile ILC: Angelo Mario Del Grosso

- **Museo Archeologico di Zagabria e Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" - Consiglio Nazionale delle Ricerche (ISTI-CNR)**
  Sviluppo congiunto di ricerche nei settori della linguistica computazionale e della computer graphics applicati ai beni culturali, archeologia e epigrafia digitali, informatica umanistica.
  Responsabile ILC: Federico Boschetti

- **Museo Galileo - Istituto e Museo di Storia della Scienza (IMSS)**
  Le attività di ricerca hanno avuto per oggetto lo studio e la realizzazione di parser automatici per il mapping in formato XML TEI dei testi digitali, ormai divenuti obsoleti, relativi al corpus dei testi galileiani. Si tratta di oltre 500 opere che Antonio Favaro, il massimo studioso di Galileo, ha individuato come appartenute alla biblioteca privata del grande scienziato. Le attività di standardizzazione dei materiali hanno riguardato soprattutto la conservazione di tutte le annotazioni, sia di tipo linguistico sia di tipo strutturale, che erano state oggetto di storici progetti di studio e ricerca del passato. Il corpus recuperato consiste di tutte le opere di Galileo, quelle di

scienziati contemporanei che lui stesso ha commentato e un immenso carteggio (più di 4.000 lettere) riguardante la corrispondenza che Galileo ha intrattenuto con le più grandi personalità del suo tempo.
Responsabili ILC: Simonetta Montemagni, Eva Sassolini

- **Palazzo Reale di Genova**
  Valorizzazione dei beni culturali attraverso l'individuazione di metodologie per migliorarne la fruizione da parte degli utenti. In particolare, attività di ricerca per lo studio di metodi e procedure per la creazione di ambienti virtuali e piattaforme software che facilitino la fruizione dei beni culturali e l'esplorazione del testo.
  Responsabile ILC: Paola Cutugno

- **Polo Museale della Liguria**
  Valorizzazione dei beni culturali attraverso l'individuazione di metodologie per migliorarne la fruizione da parte degli utenti. In particolare, attività di ricerca per lo studio di metodi e procedure per la creazione di ambienti virtuali e piattaforme software che facilitino la fruizione dei beni culturali e l'esplorazione del testo.
  Responsabile ILC: Paola Cutugno

- **Scuola Superiore Sant'Anna - Istituto di Management, Laboratorio Management e Sanità**
  Attività di collaborazione sul tema della comunicazione sanitaria, con particolare attenzione all'apporto delle tecnologie della lingua per la verifica dell'accessibilità dei testi e la loro eventuale semplificazione.
  Responsabile ILC: Felice Dell'Orletta

- **Sociedad Cubana de Derecho e Informática (SCDI)**
  Studio e realizzazione di moduli software di analisi testuale per gestire e interrogare corpora testuali di dominio.
  Responsabili ILC: Eva Sassolini

- **Universidad Nacional de Educación a Distancia (UNED)**
  Scambio di competenze scientifiche nell'ambito della gestione di materiale testuale di valore storico-culturale. La collaborazione è finalizzata alla conversione in formato XML-TEI dei testi e alla loro interrogazione con un sistema di analisi testuale per testi lemmatizzati. In particolare, nell'ambito del progetto lessicografico LENESO (LÉxico Nautico Español del Siglo de Oro), studio e normalizzazione delle annotazioni lessicali, mapping nel formato standard XML, realizzazione di un formato degli archivi compatibile con il sistema di indicizzazione e interrogazione DBT, realizzazione di un'applicazione web per la consultazione online di alcuni dei testi lemmatizzati che compongono il corpus oggetto del progetto.
  Responsabile ILC: Eva Sassolini

- **Università degli Studi di Firenze - Dip. di Lettere e Filosofia (DILEF)**
  Collaborazione scientifica finalizzata da un lato al recupero di materiale testuale digitale e alla sua conversione in formato standard di rappresentazione (XML TEI), dall'altro allo sviluppo e integrazione di moduli e procedure software per l'analisi testuale. In una prima fase la collaborazione ha riguardato il recupero di 58 testi tratti da Manifesti futuristi e del corpus di scritti teorici prodotti dai più significativi movimenti delle avanguardie iberiche in lingua spagnola, catalana e portoghese. La collaborazione è stata poi rinnovata per nuove attività di ricerca legate alla predisposizione di strategie e metodi per l'estrazione ragionata di campi significati dalle fonti dell'archivio dell'Opera della cattedrale di Firenze per gli anni (1417-1436). L'archivio, sul quale definire l'estrazione, contiene tutta la documentazione che l'Opera di Santa Maria del Fiore di Firenze ha conservato in merito alla costruzione della Cupola di Brunelleschi
  Responsabile ILC: Eva Sassolini

- **Università degli Studi di Genova - Dip. di Architettura e Design (UNIGE-DAD)**
  Studio e realizzazione di moduli software per accedere, gestire ed estrarre informazioni dai dati relativi alla ricerca: Censimento e schedatura di complessi di architettura moderna e contemporanea in Liguria
  Nel 2017 sono proseguite le attività di ricerca per l'analisi, l'organizzazione e l'indicizzazione dei dati testuali forniti dal DAD, relativi agli edifici individuati nel territorio ligure, con strumenti di analisi linguistica. Per l'estrazione delle terminologie di dominio si è fatto riferimento ai dizionari per l'urbanistica e i beni culturali disponibili presso ILC. I materiali testuali forniti dal DAD sono stati analizzati con strumenti linguistici ed è stata effettuata l'estrazione terminologica.

Responsabili ILC: Paola Cutugno, Lucia Marconi

- **Università degli Studi di Napoli Federico II - Centro Interdipartimentale L.U.P.T. (Laboratorio di Urbanistica e di Pianificazione del Territorio "Raffaele d'Ambrosio")**
  Azioni comuni di ricerca e formazione: monitoraggio e linee guida per la costruzione, uso e comprensione delle forme della comunicazione pubblica sincrona e asincrona in rete; corsi di formazione nell'ambito della pragmatica della comunicazione e dell'educazione digitale; approfondimenti scientifici e pianificazione di interventi formativi nell'ambito della comunicazione asincrona con riferimento alla Computer Mediated Communication e alla Keyboard-to-screen-communication; attività di collaborazione con centri deputati a sostegno delle specifiche disabilità linguistiche.
  Responsabile ILC: Simonetta Montemagni

- **Università degli Studi di Pavia - Dip. di Studi Umanistici**
  Attività di collaborazione sulle seguenti tematiche di ricerca: progettazione e sviluppo di risorse linguistiche mono e multilingui (corpora annotati e non, lessici, ontologie), relative a lingue antiche e/o moderne; utilizzo di risorse linguistiche prodotte dalla collaborazione o già liberamente disponibili presso le Parti per fini di ricerca linguistica a vari livelli di analisi (morfologia, sintassi, semantica, lessico).
  Responsabile ILC: Vito Pirrelli

- **Università degli Studi di Roma "La Sapienza" - Dip. di Scienze Giuridiche**
  La sincronizzazione del testo latino e greco con la traduzione in italiano dei Digestae giustinianei, avviata nell'ambito del programma di ricerca PRIN 2008 "Traduzione dei 50 libri dei Digesta di Giustiniano: Lessico giuridico storia e dogmatica", con il Dip. di Storia e Teoria del Diritto dell'Università di Roma Tor Vergata, è poi proseguito con la collaborazione del Dip. di Scienze Giuridiche dell'Università degli Studi di Roma La Sapienza. Nel 2017 l'ILC ha supportato il lavoro dei traduttori dell'Università degli Studi di Roma, realizzando l'allineamento automatico delle versioni latino e italiano e mettendo a disposizione degli studiosi un sistema di consultazione online del corpus bilingue, con diverse modalità di accesso.
  Responsabili ILC: Eva Sassolini

- **Università degli Studi Suor Orsola Benincasa (UNISOB)**
  Attività di ricerca sull'acquisizione di testi mediante procedure di OCR e studio delle strategie di lettura, comprensione e correzione degli stessi tramite tecnologie di eye-tracking.
  Responsabili ILC: Federico Boschetti

- **Università di Pisa - Dip. di Informatica e Università degli Studi di Torino - Dip. di Informatica**
  Costruzione di una Treebank dell'italiano con annotazione sintattica a dipendenze secondo lo schema "Universal Dependencies"
  Responsabile ILC: Simonetta Montemagni

- **Università di Pisa - Dip. di Filologia, Letteratura e Linguistica**, **Università di Perugia - Dip. di Lettere e Fondazione Camillo Caetani**
  Acquisizione di collezioni documentali mediante tecniche avanzate di OCR e sviluppo di un sistema per l'archiviazione, la gestione e l'interrogazione dei testi.
  Responsabili ILC: Federico Boschetti

- **Université Paris-Sorbonne - Observatoire de la vie littéraire (OBVIL)**
  Attività di ricerca finalizzata allo studio e produzione di edizioni digitali di manoscritti di autori moderni e contemporanei.
  Responsabile ILC: Simonetta Montemagni

- **Université Sidi Mohammed Ben Abdellah**
  Sviluppo di una collaborazione nei settori dell'istruzione e della ricerca in Scienze Umane e Sociali, in particolare per quanto riguarda il trattamento automatico della lingua e l'elaborazione digitale dei manoscritti, della collazione e dell'edizione critica.
  Responsabile ILC: Vito Pirrelli

- **University of Göttingen**
  DiTMAO "An XML-based Information System for Old Occitan Medical Terminology": creazione di un sistema di lessicografica digitale contenente termini medico-botanici in occitano antico derivanti da testi medievali in latino e in caratteri ebraici.
  Responsabile ILC: Emiliano Giovannetti
- **University of Patras (UPatras)**
  Erasmus+-Erasmus Charter for Higher Education (ECHE).
  Referente ILC: Vito Pirrelli
- Nell'ambito del progetto DLDP, sono stati attivati rapporti di collaborazione formale per promuovere la presenza digitale delle lingue minoritarie con:

  Lenguas Indigenas, https://rising.globalvoices.org/lenguas, Eddie Avila
  Conradh na Gaeilge, http://www.cnag.ie, Padraig O Tiarnaig
  Indigenous Tweets, http://indigenoustweets.com, Kevin Scannell
  Centro Interdisciplinar de Documentaçao Linguistica e Social, http://www.cidles.eu, Vera Ferreira
  Language Technologies Unit, University of Bangor, http://techiaith.bangor.ac.uk, Delyth Prys
  Anveatsa Armaneashti, http://anveatsaarmaneashti.com/invata-online, Dumitru Tosa/Florentina Costea
  Wikimedia France, http://www.wikimedia.fr, Rémy Gerbet
  Academiy du Galo, http://www.academie-du-gallo.bzh, Gwenaelle Lefeuvre
- Il rapporto di collaborazione con **Unione delle comunità ebraiche italiane - Collegio Rabbinico Italiano (UCEI-CRI)** nell'ambito del Progetto Talmud è stato consolidato grazie alla stipula di una nuova convenzione operativa, che vede l'Istituto (in particolare il gruppo di Literary Computing) impegnato nel progetto per altri 5 anni.

2.5.3 Altre collaborazioni

Nel corso del 2017 sono state rafforzate le collaborazioni dell'ILC con la comunità scientifica italiana tramite:
- la partecipazione a numerosi bandi regionali, nazionali e internazionali;
- la stipula di dichiarazioni di intenti, protocolli di intesa, accordi quadro e convenzioni di varia natura con università, enti di ricerca e istituzioni culturali;
- attività di alta formazione in università italiane;
- la collaborazione con altri istituti CNR per lo sviluppo di progetti multidisciplinari.

Oltre alle collaborazioni che sono state oggetto di un accordo formale (cfr. *supra*), nel corso del 2017 l'ILC ha ampliato la sua rete di contatti attraverso collaborazioni informali, sia con la rete scientifica CNR sia con la comunità scientifica nazionale e internazionale. Il personale dell'ILC collabora non solo con altri gruppi di ricerca specializzati nel settore della linguistica computazionale, ma anche con studiosi di altre discipline umanistiche e di altri settori che possono beneficiare delle tecnologie del linguaggio.

Nel 2017 l'ILC ha consolidato la sua rete di collaborazioni mediante:

- visite di studiosi stranieri di chiara fama presso la propria sede; stipula di accordi bilaterali di collaborazione scientifica con prestigiose istituzioni straniere; avvio o la prosecuzione di collaborazioni scientifiche internazionali extra-progettuali con università, enti di ricerca e istituzioni culturali internazionali, così come con l'UNESCO e il Consiglio d'Europa; organizzazione congiunta di conferenze e workshop internazionali.
- la posizione come referente tecnologico nazionale nell'ambito dell'azione ELRC - European Language Resource Coordination e come coordinatore nazionale di CLARIN-IT che ha portato all'estensione della rete di relazioni internazionali sia nell'ambito della Linguistica Computazionale sia delle SSH.
- la partecipazione a numerose proposte progettuali sottomesse in risposta a bandi regionali, nazionali, internazionali e di fondazioni private, da parte sia di singoli gruppi di ricerca sia di più gruppi congiuntamente. Quest'ultima opzione è stata fortemente incoraggiata per potenziare sinergie interne a partire dalla definizione di

proposte progettuali. Tra queste, va menzionata la collaborazione con i seguenti Istituti CNR per lo sviluppo di progetti interdisciplinari:

DIPARTIMENTO DI SCIENZE DEL SISTEMA TERRA E TECNOLOGIE PER L'AMBIENTE (DTA)
- Istituto di Scienze Marine (ISMAR)

DIPARTIMENTO DI SCIENZE FISICHE E TECNOLOGIE DELLA MATERIA (DSFTM)
- Istituto di Biofisica (IBF)

DIPARTIMENTO DI SCIENZE UMANE E SOCIALI, PATRIMONIO CULTURALE (DSU)
- Istituto per il Lessico Intellettuale Europeo e Storia delle Idee (ILIESI)
- Istituto di Scienze e Tecnologie della Cognizione (ISTC)
- Istituto di Teoria e Tecniche dell'Informazione Giuridica (ITTIG)

DIPARTIMENTO DI SCIENZE BIOMEDICHE (DSB)
- Istituto di Fisiologia Clinica (IFC)

DIPARTIMENTO INGEGNERIA, ICT E TECNOLOGIE PER L'ENERGIA E I TRASPORTI (DIITET)
- Istituto di Elettronica e di Ingegneria dell'informazione e delle Telecomunicazioni (IEIIT)
- Istituto di Informatica e Telematica (IIT)
- Istituto di Matematica Applicata e Tecnologie Informatiche "Enrico Magenes" (IMATI)
- Istituto dei Materiali per l'Elettronica ed il Magnetismo (IMEM)
- Istituto di Studi sui Sistemi Intelligenti per l'Automazione (ISSIA)
- Istituto di Scienza e Tecnologie dell'Informazione “Alessandro Faedo” (ISTI).

- la partecipazione attiva alla vita di associazioni scientifiche nazionali operanti nei settori della linguistica computazionale e dell’informatica umanistica, in particolare: l'*Associazione Italiana di Linguistica Computazionale* (AILC, http://www.ai-lc.it/it/), all'interno della quale Simonetta Montemagni svolge il ruolo di vice-presidente e un ricercatore dell’istituto, Felice Dell’Orletta, è membro del Direttivo; l’*Associazione per l’informatica umanistica e la cultura digitale* (AIUCD, http://www.aiucd.it/associazione/), all’interno della quale Federico Boschetti, ricercatore dell’Istituto, è membro del Direttivo. Tali azioni sono promosse nella convinzione che l’ILC debba rappresentare un partner attivo e trainante nell’ambito di iniziative associative. Entrambe le associazioni, infatti, svolgono un ruolo strategico nella creazione di comunità scientifiche nazionali dei settori della Linguistica Computazionale e delle Digital Humanities, con particolare attenzione alla promozione delle attività scientifiche e formative all’interno dei due settori, al consolidamento dei legami con altre iniziative, nazionali, europee e internazionali che operano in questi ambiti e alla loro promozione nell'ambito della politica nazionale, in particolare per quanto riguarda l'università e la ricerca scientifica.
- il ruolo di rappresentante del consorzio italiano di CLARIN: tramite il Coordinatore Nazionale per l’Italia dell’Infrastruttura di ricerca per le Scienze Umane e Sociali CLARIN-ERIC, ruolo assegnato a Monica Monachini, ricercatrice dell’ILC, l’ILC ha condotto una capillare opera di disseminazione presso la comunità dei potenziali utenti nell’ambito sia della Linguistica Computazionale sia delle Scienze Umane e Sociali. Per quanto il consorzio CLARIN-IT sia tuttora in fieri, sono stati già stipulati accordi formali e/o firmate manifestazioni di interesse da parte dei partner.

Altre collaborazioni informali, nazionali e internazionali, riguardano:

- **Centro de Lingüística Aplicada (CLA)**
  Studi congiunti sullo spagnolo di Cuba
  Referenti: Paola Cutugno; Lucia Marconi

- **Cloud GARR**
  Collaborazione nell’ambito di CLARIN: strategie per il deposito, la archiviazione e la preservazione di grandi risorse linguistiche
  Referente: Monica Monachini

- **Istituto di Fisiologia Clinica (IFC)**
  Nell'ambito del laboratorio ComPhys, disegno e sviluppo di una piattaforma interattiva basata sull'utilizzo di un tablet, finalizzata alla valutazione dell'efficienza di lettura in alunni della scuola elementare
  Referente: Vito Pirrelli
- **Russian Committee for UNESCO Information for All Programme (IFAP)/ Interregional Library Cooperation Centre**
  Scambio di informazioni e know-how per la promozione del multilinguismo sul web
  Referente*:* Claudia Soria
- **Social Sciences and Humanities Research Council of Canada**
  "Words of the World" - SSHRC Partnered Research Training Initiative
  Referente: Vito Pirrelli, Responsabile per la collaborazione scientifica dell'unità italiana
- **Scuola Internazionale di Studi Avanzati (SISSA)**
  Language, Learning and Reading Lab: confronto tra i dati di finger tracking (progetto TABLET) e eye-tracking
  Referenti: Marzi, Marcello Ferro, Vito Pirrelli
- **Scuola Normale Superiore (SNS)**
  Collaborazione nell'ambito di CLARIN: inclusione del portale Grafo (http://grafo.sns.it)
- **UCLA, Luskin School of Public Affairs**
  Collaborazione con la professoressa Laura Wray-Lake sull'analisi delle motivazioni della protesta Women's March 2017 espresse su Twitter
  Referente: Irene Russo
- **Università di Pisa, Laboratorio di Antropologia del Mondo Antico (LAMA)**
  Collaborazione con il LAMA tramite seminari, implementazione di software (Euporia 2.0) per l'annotazione di termini rilevanti per lo studio dei riti in tragedia, stesura di articoli scientifici, partecipazione a convegni, tutoraggio per gli stage curricolari
  Referente: Federico Boschetti
- **Università al-Qarawiyyin di Fez**
  Avvio del progetto di digitalizzazione della biblioteca al-Qarawiyyin
  Referente: Vito Pirrelli
- **Università di Pisa**
  Sviluppo congiunto del software di visualizzazione web EVT nell'ambito del progetto "Rotulo di San Teobaldo"
  Referente: Simone Marchi
- **Università di Pisa, Dip. di Filologia, letteratura e linguistica**
  Collaborazione con la professoressa Daria Coppola sull'uso di risorse e strumenti digitali per l'insegnamento delle lingue seconde nella scuola secondaria
  Referente: Irene Russo
- **Università di Pisa, Dip. di Ingegneria dell'Informazione**
  Analisi di misure di arousal e valence per parole italiane somministrate in isolamento a soggetti, da mettere a confronto con dati autonomici
  Referente: Claudia Marzi
- **Università Guglielmo Marconi**
  Collaborazione con il dott. Simone Pisano sull'annotazione e l'estrazione automatica di fenomeni linguistici rilevanti dai romanzi di Grazia Deledda
  Referente: Irene Russo
- **Université Sidi Mohammed Ben Abdellah - Laboratory for Information Management System (LIMS)** (Marocco)
  Condivisione della piattaforma READLET per la valutazione dell'efficienza di lettura in decodifica e comprensione:
  Referente: Vito Pirrelli
  Altro personale: Marcello Ferro, Claudia Marzi, Ouafae Nahli

- **Universitat Autònoma de Barcelona - Computer Vision Center (CVC)**
  Collaborazione con il professore Jordi Gonzàlez sull'analisi in parallelo di immagini e testo dai social media
  Referente: Irene Russo
- **University of Alberta - Dept. of Psychology** (Canada)
  Definizione di un modello esplicativo "memory-based" dell'effetto di latenza nel compito di digitazione di parole composte inglesi in corrispondenza del confine di morfema, che delinea un effetto combinato di specializzazione predittiva e competizione inibitoria tra composti e loro costituenti memorizzati contestualmente
  Referente*:* Vito Pirrelli
- **University of Strathclyde** (Glasgow, UK)
  Progetto volto a ridurre i tempi di ricerca all'interno di spazi metrici sui quali viene costruito un indice di ricerca.
  Referente: Franco Alberto Cardillo

A testimonianza della sempre crescente capacità di "fare rete" dell'Istituto, nel corso del 2017 ricercatori dell'ILC hanno avuto contatti per collaborazioni in attività scientifiche anche con altri soggetti, tra cui: Fondazione per le scienze religiose Giovanni XXIII; Università della Calabria; Università Cattolica del Sacro Cuore; Università degli Studi di Torino; Università Roma Tre; Universität Leipzig; University Savoie Mont-Blanc; Université Paul Valéry.

2.6 Valutazione della ricerca dell'ILC

A febbraio 2017 l'ANVUR (Agenzia Nazionale di Valutazione del Sistema Universitario e della Ricerca) ha pubblicato il Rapporto "Valutazione della Qualità della Ricerca 2011-2014 (VQR 2011-2014)" relativo ai risultati della ricerca del CNR (scaricabile all'indirizzo http://www.anvur.org/rapporto-2016/files/Enti/99.CNR.pdf).
Per quanto concerne l'ILC, nella VQR 2011-2014 sono stati valutati in totale 45 prodotti della ricerca: 42 di area umanistica (L-LIN/01) e 3 di area informatica (INF/01). Il 66% di questi prodotti ha ricevuto un giudizio Eccellente (punteggio 1) o Elevato (0,7), con un voto medio di 0,64. Per quanto riguarda l'indicatore finale di qualità della ricerca dell'istituzione, ovvero IRFD, l'ILC è risultato avere un peso quali-quantitativo superiore alla quota di prodotti attesi: si colloca infatti tra i 6 istituti CNR con IRFD > n/N, e in particolare risulta essere quello che registra lo scarto maggiore (0,53 vs 0,35).

2.7 Premi

- **Conferenza BEA 2017**
  Andrea Cimino e Felice Dell'Orletta (ItaliaNLP Lab)
  1ª posizione - Native Language Identification Shared Task del 12th Workshop on Innovative Use of NLP for Building Educational Applications
- **Premio AILC (Associazione Italiana di Linguistica Computazionale) per la migliore tesi di Laurea Magistrale**
  Titolo della tesi: *Definizione di modelli computazionali per lo studio dell'evoluzione delle abilità di scrittura a partire da un corpus di produzioni scritte di apprendenti della scuola secondaria di primo grado*. Autore: Alessio Miaschi, Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica. Relatore: Felice Dell'Orletta.
  Premiazione avvenuta in occasione della IV edizione della Conferenza Italiana sulla Linguistica Computazionale (CLiC-it 2017).
  Roma, Sede centrale del CNR, 11-13 dicembre 2017

3 Attività di ricerca

In conformità con quanto indicato nelle "Linee Guida per la gestione integrata del Ciclo della Performance degli Enti Pubblici di Ricerca" dell'ANVUR, le attività condotte nel corso del 2017 sono qui di seguito suddivise in *ricerca scientifica* e *ricerca istituzionale*.

3.1 Ricerca scientifica

Nel 2017 le attività di ricerca scientifica hanno permesso di conseguire risultati apprezzabili, consolidando ed estendendo la visibilità nazionale e internazionale delle linee di attività.
Per tutte le aree di attività, sono stati pubblicati numerosi articoli su riviste e negli atti delle principali conferenze nazionali e internazionali del settore. Inoltre, è stata estesa la rete di contatti scientifici e collaborazioni di ricerca e sono state moltiplicate le iniziative di formazione specialistica e di coordinamento delle attività scientifiche.

Qui di seguito i risultati ottenuti nelle diverse aree di ricerca nel corso del 2017.

DH - Digital Humanities

Per quanto riguarda l'area di ricerca Digital Humanities nel corso del 2017 sono state sviluppate le seguenti attività:

CoPhiLab

I risultati scientifici più significativi raggiunti si articolano su più fronti: elaborazione di metodi per migliorare l'accuratezza dell'OCR applicato a testi di interesse storico e letterario; sviluppo di metodi e strumenti per l'annotazione di testi tramite Domain Specific Languages e per la riorganizzazione di Personomies/Folksonomies in ontologie; teorizzazione del ruolo dell'editore nello Scholarly Editing.

Gruppo Literary Computing

Sono stati ulteriormente sviluppati i modelli e gli strumenti per la traduzione e lo studio di testi adottati nell'ambito di numerosi progetti. La solidità e la flessibilità dei principi di modellazione e progettazione adottati hanno trovato conferma nella relativa semplicità con la quale è stato possibile declinare, nei progetti avviati nel 2017, quanto già sviluppato. Nuove importanti funzionalità sono state aggiunte alla piattaforma Omega di gestione e analisi per lo studio del testo: quelle di descrizione delle risorse digitali utilizzando lo standard di metadatazione Dublin Core (DCMI); quelle per la presentazione gerarchica delle risorse analoga alla classica navigazione del file system; quelle di accesso ai dati tramite Web-API con specifiche RESTful e identificazione dei testi tramite URI/URL, compatibili anche con l'architettura CITE-CTS. La piattaforma collaborativa Omega poggia su un'architettura a microkernel, concepita per favorire l'estensibilità e la flessibilità delle sue varie componenti. L'approccio di modellazione adottato impiega alcuni tra i più noti Design Pattern di progettazione e programmazione orientata agli oggetti (Factory, Visitor, Composite Component, ecc.) per l'impiego di soluzioni generiche a problemi ricorrenti nel dominio dello studio del testo. Questo ambiente digitale ha l'ambizione di rispondere alle esigenze proprie degli studiosi di documenti testuali. È stata inoltre avviata l'integrazione del software web EVT (Edition Visualization Technology) all'interno della piattaforma Omega. EVT è un'applicazione Web per la visualizzazione di edizioni digitali (diplomatiche, interpretative e critiche) sviluppato presso il Laboratorio di Cultura Digitale dell'Università di Pisa. Contestualmente, sono stati realizzati alcuni prototipi da integrare in EVT per la visualizzazione di immagini facsimile ad alta risoluzione e per l'ancoraggio parallelo tra testo e immagine.
Sul versante della CAT (Computer-Assisted Translation), è stata messa a punto una tecnica innovativa volta a velocizzare il lavoro di revisione del testo tradotto attraverso l'applicazione di tecniche di analisi stilistica, e sono stati sviluppati strumenti e risorse per il trattamento automatico dell'Ebraico Mishnaico. Un'ulteriore strategica linea di ricerca ha riguardato la realizzazione di un sistema collaborativo per la costruzione di lessici e risorse termino-ontologiche (LexO). Tale strumento ha dato prova di particolare flessibilità, consentendo di convertire nel formato Lemon-OWL altre risorse lessicali, tra cui la risorsa diacronica della terminologia saussuriana sviluppata in seno al progetto PRIN 2008 "Per un'edizione digitale dei manoscritti di Ferdinand de Saussure", la risorsa Clavius, una risorsa bilingue sviluppata nel progetto Totus Mundus e una parte di lessico della lingua araba medievale compilato a partire dal dizionario al-Qāmūs al-Muḥīṭ.

Gruppo Text Analysis and Mining

Le attività di ricerca e di studio hanno come centro di interesse il testo e tutte le problematiche ad esso connesse: l'acquisizione, il trattamento, l'annotazione, la consultazione e l'analisi testuale, sino alla conservazione a lungo termine e alla valorizzazione dei contenuti. Nell'ambito di queste attività è stato implementato un protocollo di recupero, conservazione e valorizzazione di testi e corpora digitali interessati da problemi di obsolescenza tecnologica. Le strategie di salvaguardia adottate si spingono oltre il salvataggio e la conservazione in un formato di rappresentazione in linea con gli standard internazionali (XML TEI). L'intento ultimo riguarda la valorizzazione di questo patrimonio recuperato attraverso nuove modalità di fruizione dei contenuti. Nel contesto offerto dal grande sviluppo delle tecnologie informatiche le classiche funzionalità di analisi testuale sono ripensate per il web, sia come modalità di interazione con l'utente sia come tipologia di risposte da proporre agli utenti, combinando tecniche di *distant reading* e funzionalità classiche di Information Retrieval. A questo scopo sono state studiate nuove organizzazioni dei materiali, da un lato realizzando estrazioni ragionate dei contenuti che possono produrre viste sintetiche, dall'altro implementando tecniche di *responsive web design* per rendere gli strumenti di ricerca adattabili ai diversi dispositivi. In questo scenario di flessibilità e riorganizzazione degli accessi sono state realizzate nuove modalità grafiche e visuali di fruizione dei dati che applicano tecniche di *visual analytics*.

LRI - Risorse, standard e infrastrutture di ricerca

Nel 2017, lo sviluppo del data center nazionale dell'infrastruttura CLARIN ha rappresentato il focus delle attività, con l'implementazione del repository e dei servizi di accesso, deposito e conservazione di dati linguistici, e l'erogazione di un insieme di servizi per il Trattamento Automatico del Linguaggio (TAL), alcuni dei quali sono stati integrati negli ambienti di annotazione disponibili all'interno dell'infrastruttura. La qualità dei servizi offerti ha permesso al centro ILC di avviare le procedure di certificazione internazionali, Data Seal of Approval e il certificato di livello B (secondo la certificazione CLARIN). Sono inoltre proseguite le attività tese all'identificazione dei requisiti per la gestione dei dati e degli strumenti linguistici, e alla redazione e implementazione di linee-guida per la comunità relative al ciclo di vita delle risorse linguistiche, in armonia con le pratiche nell'ambito del Patrimonio Culturale, in vista di un Data Management Plan (DMP). Sono stati intensificati gli studi sui metadati per l'accesso alle risorse del settore da parte della comunità e le iniziative di standardizzazione a livello nazionale (UNI) e internazionale (ISO e W3C), con l'applicazione dei formati volti a facilitare la pubblicazione e l'accesso ai dati aperti. Ciò ha portato al rilascio e all'integrazione di risorse nella rete Linked Open Data.

Sono proseguite le attività mirate allo studio di metodi e strumenti per la protezione e valorizzazione del patrimonio culturale immateriale. Le diverse attività svolte corrispondono ad altrettanti temi che si collocano all'interno di questa direzione di ricerca: una riflessione e analisi dell'importanza del mantenimento della diversità linguistica a livello globale, per preservare il portato culturale delle comunità; l'elaborazione di nuovi modelli per la valutazione del grado di vitalità linguistica delle lingue regionali e minoritarie; l'elaborazione della nozione di "diversità linguistica digitale" e di un modello teorico di "vitalità linguistica digitale" corredato da un insieme di indicatori; la promozione e valorizzazione delle lingue minoritarie, con particolare attenzione al ruolo delle tecnologie linguistiche e digitali per la loro preservazione e rivitalizzazione. Sono continuate, inoltre, le attività di ricerca nel campo della terminologia e della letteratura grigia, in particolare: studio e applicazione di tecniche di visualizzazione grafica per la rilevazione di metadati bibliografici allo scopo di monitorare attività documentaria e comunità di riferimento; studio della terminologia di dominio su un corpus costituito dai titoli delle presentazioni contenute nei proceedings dell'International Conference on Grey Literature nell'arco temporale 2003-2014; studio comparato tra la terminologia documentale usata nella comunità della letteratura grigia e quella adottata dall'infrastruttura europea CLARIN.

Si è provveduto, infine, allo studio e all'analisi di metodologie per la disambiguazione automatica dei verbi d'azione, all'elaborazione di schemi di annotazione per l'analisi automatica e in parallelo di testi e immagini dai social media e alla definizione di uno schema di annotazione e conseguente annotazione sull'uso del doppio codice nei romanzi di Grazie Deledda.

TAL - Trattamento automatico del linguaggio naturale ed estrazione di conoscenza

Le attività si sono focalizzate sullo studio di modelli probabilistici del linguaggio e sullo sviluppo di algoritmi di apprendimento automatico per il trattamento automatico del linguaggio e l'estrazione di conoscenza. I risultati più significativi includono lo sviluppo di:

- algoritmi per l'annotazione linguistica automatica, con particolare attenzione all'adattamento a diverse varietà d'uso della lingua;

- metodi e strumenti per l'estrazione di informazione da testi (sono stati sviluppati sistemi finalizzati a: l'identificazione automatica di testimoni oculari a partire dalle descrizioni in linguaggio naturale sui social media; Sentiment Analysis e Hate Recognition; l'identificazione e classificazione delle entità nominate e l'estrazione dei concetti rilevanti all'interno di collezioni documentali di dominio);
- metodi e strumenti per l'estrazione di conoscenza linguistica, in particolare: profilazione delle caratteristiche linguistiche di collezioni di testi; valutazione della leggibilità di un testo, rispetto alle competenze linguistiche dei destinatari e al dispositivo di lettura usato; identificazione automatica della L1 a partire dall'analisi di produzioni L2;
- sistemi per l'analisi automatica della documentazione tecnica (sviluppo di sistemi per l'estrazione di informazione dai brevetti e per l'analisi dei requisiti scritti in linguaggio naturale e sviluppo di metodi per lo studio diacronico della documentazione tecnica allo scopo di identificare la tecnologia emergente);
- metodi per la semplificazione automatica del testo (sviluppo di metodi semi-automatici per la creazione di grandi risorse monolingua parallele di frasi semplici e complesse; sviluppo di sistemi di semplificazione automatica del testo attraverso metodi di generazione basati sulle reti neurali);
- studio della complessità sintattica del testo (studio di funzioni in grado di identificare e pesare i parametri che rendono complesso un testo sia per l'uomo sia per la macchina; analisi dell'impatto del genere testuale rispetto ai parametri di complessità; studio delle correlazioni tra complessità per l'uomo e per la macchina; studio dell'interazione tra la complessità sintattica e quella lessicale);
- modelli in grado di misurare e predire l'apprendimento delle abilità di scrittura dell'italiano dei bambini in età scolare (definizione di schemi di annotazione per il marcamento e la classificazione degli errori nelle produzioni scritte di apprendenti della lingua; creazione di corpora di produzioni di apprendenti della lingua italiana in età scolastica marcati con gli errori (sia produzioni scritte sia trascrizioni del parlato); sviluppo di algoritmi in grado di predire l'apprendimento dei bambini e di identificare i parametri linguistici che variano in relazione all'apprendimento);
- sistemi per il trattamento automatico del linguaggio naturale basati sul paradigma di apprendimento supervisionato chiamato Deep Learning.

Sono state sviluppate anche metodologie di classificazione, strutturazione e indicizzazione semantica dei contenuti all'interno di corpora documentali di dominio. Si è provveduto anche alla realizzazione di applicazioni Web per l'analisi testuale e per la consultazione contrastiva di testi bilingui e alla progettazione e sviluppo di funzionalità di ricerca testuale finalizzate all'integrazione in piattaforme Web modulari per l'archiviazione, la gestione e l'interrogazione di corpora testuali. Per quanto concerne la conversione di testi in formati internazionali di rappresentazione, sono state condotte attività specifiche per lo studio e la definizione di modelli di codifica per la rappresentazione del contenuto del testo e dei metadati, in grado di rappresentare in modo omogeneo, attraverso un linguaggio condiviso, le caratteristiche da codificare.

MIND - Modelli (bio-)computazionali dell'uso linguistico

Le attività svolte sono state così articolate:

- sviluppo di una piattaforma neuro-computazionale per la simulazione di processi di elaborazione seriale
  Le attività hanno riguardato: la validazione di architetture neurali ricorrenti basate su blocchi di Long Short Term Memories (LSTM) per la simulazione di processi di produzione lessicale testati su un compito di completamento di paradigmi parzialmente attestati (cell-filling problem); lo studio e l'implementazione di modelli di deep learning (reti neurali ricorrenti LSTM) per la percezione della struttura morfologica in forme verbali in compiti di generalizzazione di forme flesse in italiano e tedesco; l'analisi comparativa dei processi di auto-organizzazione lessicale in sistemi flessionali che presentano livelli di complessità paradigmatica differenti (italiano, tedesco, arabo moderno standard, greco moderno, inglese e spagnolo); la sperimentazione del porting delle attuali architetture su sistemi hardware distribuiti per l'elaborazione parallela; il disegno e lo sviluppo di una piattaforma interattiva orientata alla valutazione dell'efficienza di lettura in alunni della scuola elementare e prime sperimentazioni in scuole italiane e marocchine.

- studio di modelli teorici e computazionali di acquisizione lessicale in contesti mono- e multi-lingui
  Le principali attività di ricerca hanno riguardato: la definizione di un modello di memorizzazione e accesso lessicale di forme morfologicamente complesse, con particolare attenzione al ruolo di fattori quali frequenza, lunghezza, gradiente di regolarità morfologica, complessità della struttura superficiale per diverse lingue; simulazioni incrementali in chiave bilingue che hanno portato ad approfondire il comportamento competitivo tra l'aspettativa per suoni della prima lingua (L1) e la percezione dei fonemi esclusivi della L2, con una prima sperimentazione con dati omogenei di italiano/spagnolo e spagnolo/tedesco, volta ad evidenziare come la maggiore percezione di similarità tra la L1 e l'input della L2 favorisca il riconoscimento, la memorizzazione e acquisizione della L2.

Il lavoro di ricerca ha consentito di indagare metodi e modelli utili ad estendere e migliorare le funzionalità del modello Temporal Self Organising Map (TSOM): versioni auto-organizzanti delle architetture deep learning quali Long-Short-Term-Memory (LSTM) e ClockWork-Recurrent-Neural-Network (CWRNN) ben si presterebbero ad estendere l'architettura TSOM così da farla operare su più livelli paralleli di rappresentazione lessicale e semantica.

3.2 Ricerca Istituzionale

L'ILC partecipa attivamente a due importanti infrastrutture di ricerca di livello europeo, con ruoli diversi.

CLARIN (Common Language Resources and Technology Infrastructure)
A seguito dell'adesione dell'Italia a CLARIN-ERIC nel 2015, l'ILC in virtù del ruolo di Istituto Esecutore, ha attivato il primo centro italiano, ILC4CLARIN, che offre a tutto il consorzio servizi di catalogazione degli strumenti e delle risorse linguistiche italiane esistenti, secondo i metadati e i protocolli stabiliti da CLARIN-ERIC. Il centro è in procinto di ottenere la certificazione CLARIN di tipo B (erogazione di servizi e Data Seal of Approval). Monica Monachini, Coordinatore Nazionale per l'Italia, sul fronte delle attività di networking, sta continuando l'opera di creazione della comunità, che attualmente comprende alcuni dei principali Atenei e Istituti di ricerca che si occupano di studi linguistici e letterari, filologia e informatica, così come gruppi principalmente coinvolti nel campo delle scienze umane digitali e nei settori della documentazione, della digitalizzazione e delle tecnologie linguistiche per le scienze umane.
Sul versante della disseminazione, ha condotto capillare opera di diffusione della missione e dei vantaggi di CLARIN presso la comunità dei potenziali utenti, attraverso i vari canali di distribuzione: interventi alle Conferenze del settore, relazioni ad invito, scuole estive presso Università e Istituzioni, interviste per le riviste specialistiche.
La partecipazione di ILC a CLARIN-ERIC ha evidenti ricadute sulla internazionalizzazione della ricerca italiana e reca una serie di benefici alla rete scientifica italiana. Prima su tutti, la collaborazione con gli altri paesi membri attorno a temi condivisi dalla comunità di studiosi consente di incrementare su scala europea il valore del nostro patrimonio culturale veicolato dalla lingua. L'aumentata visibilità e accessibilità ai dati della ricerca offre il vantaggio di posizionare gli studi del linguaggio in un contesto internazionale di eccellenza. Inoltre la partecipazione nei consessi gestionali e politici di CLARIN (la Assemblea Generale ed il Forum dei Coordinatori Nazionali) consente all'Italia (e all'ILC) di portare a livello europeo le direzioni di ricerca dei settori delle scienze del linguaggio e delle scienze umane e sociali, nonché di indirizzare le linee scientifico-strategiche all'interno della federazione europea. Oltre all'eccellenza e alla ricerca congiunta al di là dei confini e delle lingue, ulteriori elementi di internazionalizzazione consistono nel continuo allineamento alle politiche per il riuso dei dati, verso la visione di una Scienza Aperta.
Tra le molteplici attività svolte nel corso 2017 vi sono anche la descrizione e l'inserimento nel repository ILC4CLARIN di varie versioni di IWN, nonché la conversione e l'adattamento di ItalWordNet al nuovo formato LMF richiesto dall'iniziativa Open Multilingual WordNet per la pubblicazione open access di IWN.

DARIAH (Digital Research Infrastructure for the Arts and Humanities)
L'ILC è partner di DARIAH-IT, il nodo Italiano di DARIAH-EU (Digital Research Infrastructure for the Arts and Humanities) che mira a fornire servizi avanzati basati sull'uso delle tecnologie dell'informazione e della comunicazione (ICT) per la ricerca nel campo delle Arti e Scienze Umane. L'ILC, in quanto centro esecutore per il CNR del nodo italiano dell'infrastruttura CLARIN, dedicata alle risorse e le tecnologie linguistiche per le Scienze Umane e Sociali e in quanto partner del progetto infrastrutturale Parthenos, promuove la comunicazione, la collaborazione e le sinergie all'interno di DARIAH-IT e CLARIN-IT, nonché la diffusione e la condivisione di esperienze e strumenti nel campo delle Digital Humanities per la trasmissione e l'utilizzo del patrimonio culturale digitale. L'Istituto, in particolare, fornisce contenuti digitali con particolare attenzione all'interoperabilità, alla conservazione e alla sicurezza dei dati stessi. Inoltre, mette a disposizione sotto forma di servizi web le funzionalità di software legacy prodotto presso l'Istituto per l'accesso e la consultazione on-line di contenuti digitali di interesse per la comunità; fornisce strumenti per l'analisi delle lingue classiche e per varietà storiche della lingua, l'estrazione di informazione da testi e il collegamento a thesauri e ontologie. Offre inoltre moduli e catene di annotazione e acquisizione automatica di conoscenza linguistica, nonché competenze nel settore degli standard di rappresentazione e dei formati di annotazione dei dati linguistici.
Ricercatori dell'ILC (Federico Boschetti, Monica Monachini e Simonetta Montemagni) contribuiscono alla governance di DARIAH-IT mediante la partecipazione a diversi Working Groups.

4 Pubblicazioni

4.1.1 Contributi in rivista

Bartolini R., Pardelli G., Goggi S., Giannini S. e Biagioni S., A terminological "journey" in the Grey Literature domain, 2017

Basili R., Montemagni S., Nota Editoriale, Italian Journal of Computational Linguistics, Vol. 3, n. 1, giugno 2017

Bompolas S., Ferro M., Marzi C., Cardillo F. A. e Pirrelli V., For a performance-oriented notion of regularity in inflection: the case of Modern Greek conjugation, 2017

Robertson B. e Boschetti F., Large-Scale Optical Character Recognition of Ancient Greek, 2017

Branco A., Cohen K. B., Vossen P., Ide N. e Calzolari N., Replicability and reproducibility of research results for human language technology: introducing an LRE special section, 2017

Connor R., Cardillo F. A., Vadicamo L. e Rabitti F., Hilbert exclusion: improved metric search through finite isometric embeddings, 2017

Ferrari A., Dell'Orletta F., Esuli A., Gervasi V. e Gnesi S., Natural language requirements processing: a 4D vision, 2017

Giannini S., Biagioni S., Goggi S. e Pardelli G., Grey Literature Citations in the age of Digital Repositories and Open Access, 2017

Giovannetti E., Albanesi D., Bellandi A. e Benotto G., Traduco: A collaborative web-based CAT environment for the interpretation and translation of texts, 2017

Marzi C., Ferro M. e Nahli O., Arabic word processing and morphology induction through adaptive memory self-organisation strategies, 2017

Pirrelli V. e Arsalane Z., Arabic Natural Language Processing: Models, systems and applications, 2017

Pretorius L. e Soria C., Introduction to the Special Issue, 2017

Venturi G., Dell'Orletta F., Montemagni S., Flore E. e Bellandi T., La qualità dei consensi informati. Un'analisi linguistico-computazionale della leggibilità dei testi, 2017

4.1.2 Contributi in volume

Boschetti F., Del Gratta R. e Del Grosso A. M., The role of digital scholarly editors in the design of components for cooperative philology, 2017

Coppola D., Moretti R., Russo I. e Tranchida F., In quante lingue mangi? Tecniche glottodidattiche e language testing in classi plurilingui e ad abilità differenziata, 2017

Manzella M. R. G., Bartolini R., Bustaffa F., D'Angelo P., De Mattei M., Frontini F., Maltese M., Medone D., Monachini M., Novellino A. e Spada A., Semantic Search Engine for Data Management and Sustainable

Development: Marine Planning Service Platform, 2017

Montemagni S. e Wieling M., Exploring the role of extra-linguistic factors in defining dialectal variation patterns through cluster comparison, 2017

Sassi M., Cuestiones pertinentes e impertinentes de los Diccionarios Temáticos, 2017

Soria C., What is Digital Language Diversity and why should we care?, 2017

4.1.3 Contributi in atti di convegno

Cardillo F. A., Ferro M., Marzi C. e Pirrelli V., How "deep" is learning word inflection?, 2017

Bartolini R., Goggi S., Pardelli G., Russo I., Farace D. e Frantzen J., Data Visualization of a Grey Literature Community: A Cooperative Project, 2017

Bartolini R., Pardelli G., Goggi S., Giannini S. e Biagioni S., A terminological "journey" in the Grey Literature domain, 2017

Bellandi A., Boschetti F., Khan A. F., Del Grosso A. M. e Monachini M., Provando e riprovando modelli di dizionario storico digitale: collegare voci, citazioni, interpretazioni, 2017

Brunato D. e Dell'Orletta F., On the order of words in Italian: a study on genre vs complexity, 2017

Cimino A., Wieling M., Dell'Orletta F., Montemagni S. e Venturi G., Identifying predictive features for textual genre classification: The key role of syntax, 2017

Del Gratta R., (Re)Using OpeNER and PANACEA Web Services in the CLARIN Research Infrastructure, 2017

Del Grosso A. M., Giovannetti E. e Marchi S., Il modello a microkernel di Omega nello sviluppo di strumenti per lo studio dei testi: dagli ADT alle API, 2017

Del Grosso A. M., Giovannetti E. e Marchi S., Thinking like the "Modern Operating Systems": The Omega architecture and the Clavius on the Web project, 2017

Del Vigna F., Petrocchi M., Tesconi M., Cimino A. e Dell'Orletta F., Hate me, hate me not: Hate speech detection on Facebook, 2017

Goggi S., Pardelli G., Russo I., Bartolini R. e Monachini M., Providing Access to Grey Literature: The CLARIN Infrastructure, 2017

Monachini M., Discipline umanistiche: vantaggi, opportunità e benefici dell'Infrastruttura di Ricerca CLARIN e del nodo nazionale CLARIN-IT per la comunità italiana, 2017

Monachini M., Infrastrutture di Ricerca e Studi Classici. CLARIN-IT: opportunità e prospettive, 2017

Monachini M., Nicolosi A. e Stefanini A., Digital Classics: A Survey on the Needs of Ancient Greek Scholars in Italy, 2017

Monachini M., Nuove tecnologie e nuovi sviluppi di indagine: CLARIN-IT e alcuni esempi di applicazione allo

studio del greco antico, 2017

Montemagni S., Nivre J., Preface, "Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017), September 18-20, 2017, Università di Pisa, Italy", pagg. III-IV, 2017

Nicolas L., Konig A., Monachini M., Del Gratta R., Calamai S., Abel A., Enea A., Biliotti F. e Quochi V., CLARIN-IT: State of Affairs, Challenges and Opportunities, 2017

Pardelli G., Giannini S., Boschetti F. e Del Gratta R., AIUCD e CLiC-it : citazioni bibliografiche a confronto, 2017

Pardelli G., Goggi S., Bartolini R., Russo I. e Monachini M., A Geographical Visualization of GL Communities: A Snapshot, 2017

Piccini S., Marchi S. e Giovannetti E., Étudier le structuralisme par le structuralisme : expériences de sémantique distributionnelle dans la construction d'un lexique électronique de la terminologie saussurienne, 2017

Pirrelli V., Marzi C., Ferro M. e Cardillo F. A., Paradigm Relative Entropy and Discriminative Learning, 2017

Pirrelli V., Storage vs. Processing in Models of Word Inflection. A Neuro-computational Hebbian Perspective, 2017

Russo I. e Soria C., Digital Language Diversity on New Media: the DLDP Survey about European Minority Languages Speakers, 2017

Sassolini E. e Cinini A., Approcci grafici all'analisi di corpora testuali, 2017

Sassolini E., Cucurullo S. e Cinini A., I corpora digitali: dall'obsolescenza tecnologica, alla salvaguardia e alla condivisione, 2017

Soria C., Alliances for digital linguistic diversity, 2017

Soria C., Inquiring current digital use and usability of regional and minority languages: the DLDP survey, 2017

Soria C., Language policies and speakers' attitudes: evaluating the impact of official recognition on some of Italy's regional languages, 2017

Soria C., The digital language vitality scale: a model for assessing digital vitality of languages, 2017.

Vadicamo L., Carrara F., Falchi F., Cimino A., Dell'Orletta F., Cresci S. e Tesconi M., Cross-media learning for image sentiment analysis in the wild, 2017

Weingart A. e Giovannetti E., From canabo to Cannabis sativa L.: Modelling Diachronic Termino-ontological Resources in the Context of DiTMAO, 2017

4.1.4 Curatele

Montemagni S., Nivre J., Proceedings of the Fourth International Conference on Dependency Linguistics

(Depling 2017), September 18-20, 2017, Università di Pisa, Italy, 2017

Soria C., Irene R. e Quochi V. (a cura di), Reports on Digital Language Diversity in Europe, 2017

Attività di curatela del volume *"Word Knowledge and Word Usage: a Cross-disciplinary Guide to the Mental Lexicon"* per la casa editrice De Gruyter (editors: Vito Pirrelli, Ingo Plag, Wolfgang U. Dressler, data di pubblicazione: 2017). Pubblicazione finanziata dall'European Science Foundation nel quadro del progetto NetWordS.

4.1.5 Rapporti tecnici e working paper

Sassolini E. e Cinini A., Digesto: nuove funzionalità e sito web, 2017

Cinini A., Cucurullo S. e Sassolini E., Rapporto Tecnico: Standardizzazione del corpus testuale del PRIN Crusca, 2017

Russo I. e Soria C., Sardinian - a digital language?, 2017

4.1.6 Altri prodotti della ricerca

Zamorani N., Featured Linguist: Nicoletta Calzolari, 2017

4.2 Comunicazioni a convegni senza pubblicazione degli atti

- **Atelier "Les manuscrits de Saussure, parmi d'autres. Problèmes, stratégies et solutions d'édition pour les archives numériques"**
  Silvia Piccini, Simone Marchi, Emiliano Giovannetti - *Étudier le structuralisme par le structuralisme: expériences de sémantique distributionnelle dans la construction d'un lexique électronique de la terminologie saussurienne*
  Ginevra (Svizzera), 9-14 gennaio 2017
- **Global Philology Open Conference**
  Angelo Mario Del Grosso, Emiliano Giovannetti, Simone Marchi - *Thinking like the "Modern Operating Systems": The Omega architecture and the Clavius on the Web project*
  Lipsia (Germania), 20-23 febbraio 2017
- **The Medieval Brain Conference**
  Anja Weingart, Emiliano Giovannetti - *From canabo to Cannabis sativa L.: Modelling Diachronic Termino-ontological Resources in the Context of DiTMAO*
  York (Gran Bretagna), University of York, 9-11 marzo 2017
- **First International Conference on Revitalization of Indigenous and Minoritized Languages**
  Claudia Soria - *Inquiring current digital use and usability of regional and minority languages: the DLDP survey*
  Panel First: *Towards the rediscovery of Italy's hidden multilingualism* (Claudia Soria, relazione su invito)
  Tavola Rotonda, Linguapax-I: Generating contexts for linguistic diversity to thrive: networks of linguistic, cultural and digital cooperation - *Alliances for digital linguistic diversity* (Claudia Soria, relazione su invito)
  Barcelona-Vic (Spagna), 19-21 aprile 2017
- **BAAL-Cambridge University Press Seminar on Minority Languages in New Media**
  Claudia Soria, Irene Russo - *Digital Language Diversity on New Media: the DLDP Survey about European Minority Languages Speakers*
  Birmingham (Gran Bretagna), 27-28 aprile 2017

- **SW4SH 2017 - Third International Workshop on Semantic Web for Scientific Heritage**
  Anja Weingart, Andrea Bellandi, Emiliano Giovannetti - *Representing Multilingualism and Multiwords in a Lemon Old Occitan Medico-Botanical Lexicon*
  Portorose (Slovenia), 29 maggio 2017

- **Wokshop "Linking Data, Ontologies and Distributional Models for the Representation of Lexical Meaning"**
  Irene Russo - *Distributional representations of concrete nouns for action verbs disambiguation*
  Firenze, Università di Firenze, 9 giugno 2017

- **XXXI Convegno AISG 2017** - **Nuovi studi sull'Ebraismo**
  Alessandra Pecchioli, Davide Albanesi, Andrea Bellandi, Emiliano Giovannetti e Simone Marchi - *Elaborazione del linguaggio naturale (NLP) in Ebraico: il caso dell'analisi linguistica automatica applicata all'ebraico mishnaico del Talmud*
  Ravenna, 4-6 settembre 2017

4.3 Altri prodotti della ricerca

Referente: Federico Boschetti (per il rilascio delle chiavi d'accesso: federico.boschetti@ilc.cnr.it)

- **CoPhiProofReader** per *Voci della Grande Guerra* (Web application)
  http://lexit.fileli.unipi.it:8080/vggWeb
- **CoPhiProofReader** per *Commerce Numérique* (Web application)
  http://cophilab.ilc.cnr.it:8080/commerceWeb
- **EuporiaWeb**: sistema di annotazione sui riti nella tragedia
  http://cophilab.ilc.cnr.it:8080/euporiaweb
- **EuporiaRAGT**: sistema di ricerca sui riti nella tragedia
  http://cophilab.ilc.cnr.it:8080/euporiaRAGThttp://cophilab.ilc.cnr.it:8080/euporiaRAGT

Referente: Riccardo Del Gratta (e-mail riccardo.delgratta@ilc.cnr.it)

- **Tokenizer-Base-Service**: Web service integrabile in WebLicht e Language Resource Swtichboard basato su un tokenizzatore in Perl sviluppato in Opener e portato in Java

Referente: Felice Dell'Orletta (e-mail felice.dellorletta@ilc.cnr.it)

- Sistema di identificazione della lingua materna attraverso l'analisi di testi scritti in inglese
- Sistema per l'identificazione e la classificazione dell'odio nei testi presenti su Facebook
- Sistemi per il monitoraggio costante di Twitter e Facebook
- **READ-IT**, un sistema per la valutazione della complessità linguistica di diverse tipologie testuali in lingua italiana
- **MONITOR-IT**, un sistema per il monitoraggio di un'ampia gamma di caratteristiche linguistiche di diverse tipologie testuali in lingua italiana
- **LINGUA**, un sistema per l'annotazione linguistica automatica del testo

Referenti: Vito Pirrelli; Marcello Ferro (e-mail vito.pirrelli |marcello.ferro @ilc.cnr.it)

- **Prototipo dello strumento READLET per il monitoraggio dell'efficienza di lettura in alunni delle scuole primarie**
  Il prototipo dello strumento READLET per la valutazione dell'efficienza di lettura (decodifica e comprensione) è stato sviluppato in collaborazione con l'istituto di Fisiologia Clinica (IFC) del CNR di Pisa.
  L'architettura consente di creare sessioni per la valutazione dell'efficienza di lettura tramite un dispositivo tablet. Viene presentato un testo, eventualmente corredato da immagini, che viene letto in modalità silente o ad alta voce. Durante la lettura il testo può essere "segnato" con il dito, così che il dispositivo possa catturare, mediante il touchscreen, le tempistiche di lettura con una elevata risoluzione temporale e spaziale. Viene successivamente proposto un questionario a risposta multipla, così da poter valutare l'effettiva comprensione del testo.
  Per quanto concerne il software sviluppato, l'architettura prevede una parte server (database, applicazione server-side, moduli per l'annotazione e l'analisi automatica della leggibilità del testo, servizi di gestione, servizi di

analisi e post-processing) e una parte client (web-app client-side) per la somministrazione del protocollo. L'interfaccia consente di: gestire operatori, utenti, testi e questionari da somministrare; somministrare il protocollo per la valutazione dell'efficienza di lettura; raccogliere informazioni sugli eventi generati dallo scorrimento del dito del bimbo durante la lettura "segnata" del testo e durante la compilazione del questionario per la verifica della comprensione. Il prototipo supporta al momento la lingua italiana, inglese, francese e araba.

Sperimentazione:
mediante un dispositivo tablet 9'', READLET è stato sperimentato con successo tra maggio ed agosto 2017 in Italia ed in Marocco. Nel territorio toscano sono state condotte sessioni di lettura ad alta voce (6 bimbi della quarta classe della scuola primaria di Montecalvoli, di cui 1 bimbo con DSA, 2 bimbi di famiglie straniere, 3 bimbi di controllo) e sessioni di lettura silente (50 bimbi della quarta e quinta classe della scuola primaria di Capalbio in Maremma (di cui 5 bimbi con DSA, 10 bimbi di famiglie straniere, 35 bimbi di controllo e per screening preliminare). In Marocco, grazie al progetto di collaborazione CNR-CNRST, READLET è stato sperimentato su 12 bimbi in compiti di lettura silente e su due diverse lingue: francese ed arabo vocalizzato. La successiva analisi dei dati ha consentito una validazione preliminare del sistema. Le informazioni di base acquisite automaticamente (tempo di lettura del testo, tempo di compilazione del questionario, accuratezza delle risposte al questionario) sono risultate già sufficienti ad individuare la presenza di problemi della lettura, in linea con il protocollo manuale "carta e cronometro" già validato dai ricercatori di IFC. Le informazioni di latenza catturate tramite il dispositivo touchscreen sono risultate ben correlate con informazioni di base quali lunghezza e frequenza. Grazie alla collaborazione con colleghi di SISSA (Trieste), è stato possibile un confronto di tali dati con informazioni catturate mediante dispositivi di eye-tracking. La valutazione di tale confronto è tutt'ora in corso d'opera.

Referente: Emiliano Giovannetti (e-mail emiliano.giovannetti@ilc.cnr.it)

- **Traduco**: Web application per il supporto alla traduzione del Talmud Babilonese
- **TagO**: Web application a supporto dell'annotazione linguistica dell'ebraico
- **LexO**: Web application per la creazione e la gestione di lessici e ontologie
- **Omega**: framework a microkernel per lo sviluppo di componenti software in Java orientati all'analisi e studio di testi

Referenti: Claudia Soria; Valeria Quochi; Irene Russo (e-mail claudia.soria | valeria.quochi | irene.russo @ilc.cnr.it)

- Digital Repository **Digital Language Diversity Project Survey Data**
  Data set che contiene le risposte originali a un questionario pubblicato nell'ambito del progetto DLDP sull'uso e l'utilizzabilità di 4 lingue regionali e minoritarie europee (basco, bretone, careliano e sardo) su supporti e dispositivi digitali.
  http://hdl.handle.net/20.500.11752/ILC-77

Referente: Roberto Bartolini

- Release ItalWordnet in formato OWN con CILI link e relative statistiche http://compling.hss.ntu.edu.sg/iliomw/omw (nell'ambito dello Special Issue su Linking, Integrating and Extending Wordnets - LiLT) focalizzato sul collegamento e l'armonizzazione coerente di wordnets in lingue diverse).
- Aggiornamento ItalWordnet all'uso del charset IsoLatin nel nome simbolico dei synset e soluzione al problema del case dei nomi dei synset sul file system NTFS e sui sistemi Unix.

4.4 Internazionalizzazione

Nel 2017 l'Istituto ha proseguito nell'azione di promozione dell'internazionalizzazione della ricerca scientifica e tecnologica nel settore della Linguistica Computazionale. Questo obiettivo è stato perseguito: partecipando a programmi di ricerca e a organismi a livello internazionale; fornendo competenze scientifiche su richiesta di autorità governative; garantendo la collaborazione con enti e istituzioni di altri Paesi nel campo scientifico-tecnologico e nella definizione della normativa tecnica.

Oltre ad essere esecutore dell'infrastruttura italiana CLARIN-IT, nonché National Representative di CLARIN-ERIC, l'Istituto è il referente tecnologico nazionale (Technology NAP) dell'azione ELRC - European Language Resource Coordination e ospita infrastrutture di ricerca al servizio della comunità per la condivisione dei risultati delle attività di ricerca, al fine di promuovere l'utilizzo degli standard e delle buone pratiche, nonché la diffusione e il riutilizzo delle risorse.

Nel corso del 2017 presso l'Istituto sono stati ospitati *visiting scholar* provenienti da altri Enti ed Università europee ed extraeuropee per attività di insegnamento e/o ricerca:

- **Yarina Amoroso Fernández, Universidad de Ciencias Informáticas (UCI)**
  Presidente della "Sociedad Cubana de Derecho e Informática de la Unión Nacional de Juristas de Cuba"
  Informatica giuridica e trattamento della lingua e, in particolare, pubblicazione digitale del libro "Recopilaciones y apuntes para una Historia Constitucional Cubana". Analisi del lessico utilizzato nei testi costituzionali.
  2 marzo - 7 maggio 2017
- **María Lourdes García-Macho Alonso de Santamaría - Universidad Nacional de Educación a Distancia (UNED)**
  Studio di una possibile conversione in formato XML-TEI delle annotazioni linguistiche inserite nei testi che fanno riferimento al corpus del progetto LENESO (LÉxico Nautico Español del Siglo de Oro).
  Referente ILC: Manuela Sassi
  2 marzo - 31 maggio 2017
- **Anja Weingart - Universität Göttingen**
  Studio delle estensioni al modello lessicale Lemon per il trattamento di fenomeni lessicali dell'Antico Occitano e analisi delle funzionalità da incorporare nell'editor LexO in lavorazione nell'ambito del progetto DiTMAO.
  Referente ILC: Emiliano Giovannetti
  14-16 febbraio 2017
- **Harry Diakoff, Responsabile dell'Alpheios Project (http://alpheios.net)**
  Incontro con i vari gruppi di lavoro per discutere possibili collaborazioni nel campo delle Digital Humanities.
  Referente ILC: Monica Monachini
  3 novembre 2017

Nel 2017 ricercatori ILC sono stati ospitati come *visiting scholars* presso Università straniere per attività di insegnamento e/o ricerca:

- **Università Sidi Mohamed Ben Abdellah e al-Qarawiyyin (Marocco)**
  Angelo Mario Del Grosso, Marcello Ferro, Ouafae Nahli, Vito Pirrelli - *Digitalizzazione dell'archivio storico della biblioteca al-Qarawiyyin*
  23-28 novembre 2017

5 Attività di alta formazione

5.1 Corsi presso Università

Laurea Triennale

Università di Pisa - Informatica Umanistica (classe L-10)
Corso: **Linguistica Computazionale**
Docente: Felice Dell'Orletta
a.a. 2017/2018

Laurea Magistrale

Università di Pisa - Informatica Umanistica (classe L-10)

Corso: **Linguistica Computazionale II**
Docenti: Simonetta Montemagni, Giulia Venturi
a.a. 2017/2018

Università di Pisa - Informatica Umanistica
Corso: **Tecnologie linguistiche per l'estrazione di informazione**
Docente: Federico Boschetti
a.a. 2017/2018

Venice International University (VIU) - Globalization Program
Corso semestrale (Fall 2017): **Digital Humanities: Web Resources, Tools and Infrastructures**
Docente: Federico Boschetti
a.a. 2017/2018

Nel corso del 2017, tramite il Coordinatore Nazionale di CLARIN, l'ILC ha condotto una capillare opera di coinvolgimento utenti, in particolare docenti e studenti, con attività di docenza presso le Università Italiane ai corsi di Laurea, Master e Specializzazione nel settore delle Digital Humanities e Scienze Umane (Roma III, Università di Parma, Università di Siena, Ca' Foscari, Venice International University).

- MOOC

La filologia si fa digitale - Università Ca' Foscari

Moduli di Federico Boschetti:

Il circolo ermeneutico nell'era digitale (https://youtu.be/3zM0nys5fiw)
Elementi di stilometria con Stylo (https://youtu.be/JDvXISlwlKE)
Strumenti integrati per lo studio del testo: l'esempio di voyant-tools (https://youtu.be/7r30WLKZ3g0)
Costruire i propri strumenti: l'esempio di Euporia (https://youtu.be/RihekCHmT80)
Conclusioni (https://youtu.be/rxfibT4HtGg)

5.2 Scuole estive

- **Strumenti digitali per umanisti**
  Scuola estiva organizzata da AIUCD (Associazione per l'Informatica Umanistica e la Cultura Digitale) in collaborazione con ILC e Università di Pisa (Laboratorio di Cultura Digitale - LABCD e Corso di Laurea in Informatica Umanistica).
  Il programma ha previsto i seguenti interventi:
  Giulia Venturi (ILC) - *Valutazione della complessità linguistica a sostegno dell'attività didattica*
  Federico Boschetti (ILC) - *Correzione collaborativa di risorse digitali acquisite tramite OCR*
  Angelo Del Grosso (ILC) e Matteo Abrate (IIT-CNR) - *Annotazioni collaborative di testi storici*
  Pisa, 12-16 giugno 2017
- **European Summer University in Digital Humanities – Culture & Technology**
  Nicoletta Calzolari Zamorani (Invited Lecturer) - *European Language Resources Initiatives – Infrastructural and Policy Issues*
  Lipsia, 24 luglio 2017
- **International Summer School LEX 2017 - Managing legal resources in the Semantic Web**
  Giulia Venturi - Natural Language processing and legal knowledge extraction
  Ravenna, Università di Bologna - Sede di Ravenna, 11-19 settembre 2017

5.3 Master

Ricercatori ILC svolgono attività didattica nei seguenti master di livello universitario:

- **La lingua del diritto. Comprensione, elaborazione e applicazioni professionali** - a.a. 2017/2018
  Master di I livello a docenza congiunta: Università di Pavia, Senato della Repubblica, CNR.

  Il Master forma professionisti della scrittura giuridica. L'obiettivo è fare acquisire consapevolezza, chiarezza e precisione nella redazione dei testi giuridici. Il contributo dell'ILC è imperniato sull'utilizzo delle tecnologie informatiche per l'analisi linguistica e per la redazione e la comunicazione dei testi giuridici.

- **Pubblicità istituzionale, comunicazione multimediale e creazione di eventi** - a.a. 2017/2018
  Master di I livello presso il Dip. di Lettere e Filosofia (DILEF) dell'Università di Firenze
  Angelo Mario Del Grosso
  27 aprile 2017 - *Introduzione alle tecnologie digitali per la redazione e la pubblicazione di contenuti Web*
  5 maggio 2017 - *Introduzione all'editoria digitale e al Web Semantico*
  Firenze, Dip. di Lettere e Filologia dell'Università di Firenze (DILEF)

5.4 Supervisione di tesi di laurea e di dottorato

Supervisione di tesi di laurea triennale

Giulia Cantoni
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Titolo: Definizione di un metodo e creazione di un corpus per la valutazione di un sistema di semplificazione automatica del testo
Relatore: Felice Dell'Orletta

Sergio Castiglia
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Titolo: Creazione di una risorsa parallela annotata con regole di semplificazione. Un'analisi linguistico computazionale sul processo di semplificazione
Relatore: Felice Dell'Orletta

Pietro Dell'Oglio
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Titolo: Lessico e sintassi: studio delle variazioni tra generi e complessità
Relatore: Felice Dell'Orletta

Supervisione di tesi di laurea magistrale

Chiara Alzetta
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Titolo: Studio linguistico-computazionale per l'analisi dei tipi linguistici. Similarità e differenze nel confronto fra Universal Dependencies Treebanks
Relatore: Simonetta Montemagni
Correlatore: Giulia Venturi

Patrizia Belik
Università Politecnica di Valencia (Spagna), Dip. di Linguistica Applicata
Titolo: Modelling second language acquisition and processing with Self-Organising Maps
Supervisore: Claudia Marzi, in co-tutela con Hanna Skorczynska Sznajder

Benedetta Iavarone
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica

Titolo: Indagine multilingue sulla complessità della frase: confronto tra difficoltà percepita e analisi automatica
Relatore: Felice Dell'Orletta

Gloria Malorgio
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Titolo: Prototipicità e marcatezza: un'analisi linguistico-computazionale delle relazioni sintattiche soggetto e oggetto diretto
Relatore: Simonetta Montemagni
Correlatore: Giulia Venturi

Alessio Miaschi
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Titolo: Definizione di modelli computazionali per lo studio dell'evoluzione delle abilità di scrittura a partire da un corpus di produzioni scritte di apprendenti della scuola secondaria di primo grado
Relatore: Felice Dell'Orletta

Sabrina Rinnone
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Titolo: Analisi della leggibilità dei consensi informati: un approccio linguistico-computazionale
Relatore: Simonetta Montemagni
Correlatore: Giulia Venturi

Alberto Stefanini
Università degli Studi di Parma
Titolo: Indagine sulle pratiche d'uso di risorse e strumenti digitali nell'ambito degli studi di filologia classica
Correlatore: Monica Monachini

- *in corso nel 2017*

Giulia Cacioli
Università di Pisa, Dip. di Filologia, Letteratura e Linguistica, Corso di Laurea in Informatica Umanistica
Tema: Sviluppo di un modulo di indicizzazione e ricerca per il software EVT
Correlatore: Angelo Del Grosso

- Supervisione di tesi di dottorato

Mustafa Khalfi
Università di Sidi Mohamed Ben Abdellah (Marocco), Facoltà di Scienze e Tecnologie, Dip. di Informatica
Titolo: Acquisition du lexique medioevale arabe Al-qamuws al-muhiyt en Lemon
Correlatore: Ouafae Nahli

Marianne Reboul
Université Paris-Sorbonne (Francia), Tesi di Dottorato in "Littératures française et comparées"
Titolo: Comparison semi-automatique des traductions en langue française de l'Odyssée d'Homère (1547-1955)
Relatore: Prof. M. Jean-Yves Masson
Membro della commissione (membre du jury): Emiliano Giovannetti

5.5 Tesi di dottorato

- discusse nel 2017

Andrea Cimino
Università di Pisa - Dottorato di Ricerca in Ingegneria dell'Informazione

Titolo: Strumenti e metodologie basati su natural language processing per l'analisi automatica di documenti tecnici
Tema: Creazione di una pipeline di analisi linguistica per l'analisi automatica di brevetti focalizzata sull'estrazione di entità di tipo utenti, vantaggi e svantaggi che può essere integrata in applicazioni di alto livello. L'estrazione di informazioni dai brevetti permette di creare una serie di applicazioni come, ad esempio, strumenti per identificare trend tecnologici.
Responsabile scientifico: Felice Dell'Orletta

- *in corso nel 2017*

Ouafae Nahli
Università: Università degli Studi di Roma "La Sapienza" - Dottorato in Lingua Araba
Titolo: Verso un'ontologia della cultura araba-islamica
Tema: Sviluppo di una rete di conoscenze per la cultura islamica araba sulla base di un processo di estrazione automatica dei dati da testi classici fondamentali. La rete di conoscenze sarà costituita da una ricca ontologia formale legata ad ontologie "general-purpose" esistenti (SUMO).
Supervisore: Vito Pirrelli

5.6 Tirocini

Nel corso del 2017 sono stati attivati numerosi tirocini rivolti a studenti iscritti a corsi di laurea universitari:

Tutor: Federico Boschetti

- Università di Pisa, Filologia e Storia dell'Antichità
  Tema: *Validazione di synset di Homeric Greec WordNet*
  Tirocinante: Francesco Martinolli
- Università di Pisa, Lingue e Letterature straniere
  Tema: *Correzione dell'OCR applicato alla rivista letteraria Commerce*
  Tirocinanti: Tiziano Lavorini; Antonio Stanzione; Francesca Todde

Tutor: Felice Dell'Orletta

- Università di Pisa, Informatica Umanistica
  Tema: *Definizione di un metodo e creazione di un corpus per la valutazione di un sistema di semplificazione automatica del testo*
  Tirocinante: Giulia Cantoni
- Università di Pisa, Informatica Umanistica
  Tema: *Lessico e sintassi: studio delle variazioni tra generi e complessità*
  Tirocinante: Pietro Dell'Oglio
- Università di Pisa, Informatica Umanistica
  Tema: *Indagine multilingue sulla complessità della frase: confronto tra difficoltà percepita e analisi automatica*
  Tirocinante: Benedetta Iavarone
- Università di Pisa, Informatica Umanistica
  Tema: *Definizione di modelli computazionali per lo studio dell'evoluzione delle abilità di scrittura a partire da un corpus di produzioni scritte di apprendenti della scuola secondaria di primo grado*
  Tirocinante: Alessio Miaschi

Tutor: Simonetta Montemagni

- Università di Pisa, Informatica Umanistica
  Tema: *Revisione, estensione e armonizzazione della UD Treebank per la lingua italiana*
  Tirocinante: Chiara Alzetta

5.7 Corsi di formazione professionale erogati presso altri Enti

- **Tecnologie del linguaggio per la valutazione della lingua del diritto**
  Ciclo di tre moduli didattici tenuti nell'ambito del corso *"Cultura digitale: linguaggio naturale e linguaggio giuridico per i siti web"* presso l'Azienda Ospedaliera Universitaria Pisana, Pisa
  Docente: Giulia Venturi
  13 ottobre 2017; 10 e 27 novembre 2017

- **Le tecnologie linguistico-computazionali per la leggibilità degli atti normativi e amministrativi**
  Ciclo di due moduli didattici tenuti nell'ambito del corso *"La qualità degli atti normativi e amministrativi"* presso l'agenzia formativa Unione dei Comuni della Versilia, Querceta (LU)
  Docenti: Dominique Pierina Brunato; Giulia Venturi
  23 e 26 gennaio 2017

5.8 Formazione interna

Le attività di formazione interna sono state pianificate e organizzate tenendo conto dei risultati della rilevazione dei fabbisogni formativi del personale ILC. In particolare, tra il personale addetto alle attività di ricerca è emersa la necessità di approfondire il tema dell'analisi quantitativa del dato linguistico. Nel mese di maggio 2017 è stato quindi organizzato un corso incentrato sull'acquisizione e sull'approfondimento di competenze di analisi quantitative e statistiche indispensabili alla descrizione e alla modellazione della complessità delle informazioni relative ai dati linguistici:

2 - 5 maggio 2017
Corso **Metodi statistici e analisi quantitative del dato linguistico**
Docente: Martijn Wieling (Università di Groningen, NL)

Programma del corso:

- *Introduzione all'esplorazione del dato linguistico*
- *Test statistici: t-test, ANOVA*
- *Test statistici: alternative non-parametriche*
- *Modelli di regressione lineare*
- *Modelli di regressione multipla*
- *Modelli di regressione non-lineare*

6 Disseminazione scientifica

6.1 Workshop, conferenze, seminari

6.1.1 Workshop e conferenze organizzati e co-organizzati dall'ILC

L'Istituto è da sempre coinvolto nell'organizzazione di alcuni dei principali eventi scientifici del settore della Linguistica Computazionale. Tra gli eventi del 2017 sono da segnalare, in particolare:

- **Voci della Grande Guerra**
  Firenze, Accademia della Crusca, 10 febbraio 2017
  Il primo convegno su "Voci della Grande Guerra" è stato organizzato dall'Università di Pisa (capofila) in collaborazione con l'ILC, l'Università di Siena e l'Accademia della Crusca. Voci della Grande Guerra è un'iniziativa scientifica e culturale che ha l'obiettivo di preservare e diffondere la memoria della Prima Guerra Mondiale attraverso la realizzazione e la pubblicazione di un corpus digitale di testi opportunamente scelti da storici e linguisti in quanto rappresentativi dei diversi modi di narrare e descrivere l'Italia in guerra da parte dei suoi protagonisti.

- **Depling 2017** - **Conferenza Internazionale sulla Linguistica della Dipendenza**
  Pisa, 18-20 settembre 2017
  IV edizione della serie DepLing, che risponde alla crescente necessità di incontri linguistici dedicati ad approcci nella sintassi, nella semantica e nel lessico centrati attorno a strutture di dipendenza come nozione linguistica centrale. La conferenza è stata organizzata dall'Università di Pisa e dall'ILC.
  http://depling-iwpt2017.di.unipi.it

- **IWPT 2017 - Workshop Internazionale sulle Tecnologie di Parsing**
  Pisa, 20 - 22 settembre 2017
  Un workshop interamente dedicato all'analisi sintattica automatica delle lingue naturali.
  http://depling-iwpt2017.di.unipi.it

- **SIMPLICITAS. Semplificazione Linguistica della Comunicazione Istituzionale per facilitare l'Accessibilità ai Contenuti Informativi**
  Napoli, 29 settembre 2017
  Workshop organizzato nell'ambito del 51° Congresso Internazionale di Studi della Società di Linguistica Italiana (SLI) "Le lingue extra-Europee e l'italiano. Problemi didattici, socio-linguistici, culturali" (28-30 settembre 2017) con lo scopo di promuovere il dibattito tra gli studiosi e gli interlocutori istituzionali che si trovano quotidianamente a confrontarsi con il tema della comunicazione efficace e accessibile a tutti.

- **CLiC-it 2017 - IV Conferenza Italiana di Linguistica Computazionale**
  Ricercatori ILC con il ruolo di "Area chairs": Vito Pirrelli (*Cognitive modeling)*; Claudia Soria *(Language Resources)*; Felice dell'Orletta (*Morphology and Syntax Processing*).
  http://sag.art.uniroma2.it/clic2017/it/home/
  Roma, Sede centrale del CNR, 11-13 dicembre 2017

6.1.2 Partecipazione a comitati scientifici di conferenze

Numerosi ricercatori dell'ILC sono membri dei comitati scientifici di conferenze e workshop nazionali e internazionali in settori strategici della Linguistica Computazionale e delle Digital Humanities, tra i quali:

- ACL2018 - 56ª Edizione dell'Annual Meeting of the Association for Computational Linguistics per l'area ``Multilinguality''
- AIUCD 2017 - VI Convegno annuale dell'Associazione per l'Informatica Umanistica e le Culture Digitali
- AIUCD 2018 - VII Convegno annuale dell'Associazione per l'Informatica Umanistica e le Culture Digitali
- CLARIN 2017 - Conferenza Annuale
- CLIC-IT 2017 - IV Conferenza Italiana di Linguistica Computazionale
- CCURL2018 - Sustaining knowledge diversity in the digital age (Workshop di LREC 2018)
- CRH-2 2017 - II Workshop on Corpus-based Research in the Humanities
- DH 2017 - Conferenza internazionale Digital Humanities
- EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing
- IMM'17 - International Morphology Meeting
- ISWC 2018 - International Semantic Web Conference
- KnowRSH - Workshop ``Knowledge Resources for the Socio-Economic Sciences and Humanities''
- LOTKS 2017 - Workshop on Language, Ontology, Terminology and Knowledge Structures
- LREC 2018 - XI Conferenza sulle Risorse Linguistiche e sulla Valutazione
- WILDRE-4 2018 - IV Workshop on Indian Language Data Resource and Evaluation (under LREC 2018)
- LT-LRL 2017 - V Workshop Language Technology for Less Resourced Languages
- WLSI 2017 - International Workshop on Worldwide Language Service Infrastructure
- XV Simposio Internacional de Comunicacion Social, organizzato dal Centro de Lingüística Aplicada (CLA)

6.1.3 Comunicazioni e seminari

- **Giornata di Studi ForumTAL 2017** - **Tecnologie Vocali e del Linguaggio Naturale per i Beni Culturali**
  Simonetta Montemagni - *Risorse e tecnologie linguistiche per la conservazione, salvaguardia e valorizzazione del patrimonio culturale*
  La giornata di studi fa il punto sulle tecnologie vocali e dell'elaborazione naturale della lingua nell'ambito della fruizione dei beni culturali, della loro catalogazione, e in generale dell'uso di tali tecnologie per i beni culturali.
  Roma, Università degli Studi Roma Tre, 19 gennaio 2017

- **6th AIUCD Conference 2017 - Il telescopio inverso: big data e distant reading nelle discipline umanistiche**
  Federico Boschetti; Anas Fahad Khan – *Provando e riprovando modelli di dizionario storico digitale: collegare voci, citazioni, interpretazioni*
  Eva Sassolini - *Approcci grafici all'analisi di corpora testuali*
  Roma, 24-28 gennaio 2017

- **ODYCCEUS Kick-off Meeting**
  Nicoletta Calzolari Zamorani (Invited Speaker)
  Lipzia, 21-23 febbraio 2017

- **Università Ca' Foscari, Dip. di Lingue e Letterature Straniere**
  Federico Boschetti - *Annotare i testi senza marcarli: come creare un domain specific language per i propri interessi di studio*
  Seminario dedicato al confronto e all'impiego di tecniche avanzate di annotazione
  Venezia, 2 marzo 2017

- **CLARIN-PLUS workshop "Working with Parliamentary Records"**
  Simonetta Montemagni - *Methods and Techniques for the Analysis of Parliamentary Records: Two Case Studies on Italian*
  Sofia (Bulgaria), 27-29 marzo 2017

- **International Workshop on Asian Language Resources**
  Nicoletta Calzolari Zamorani (Invited Speaker) - *Introducing ISO/TC 37/SC 4 Language Resources Management Activities*
  Nanning (Cina), 21-24 aprile 2017

- **CLARIN PLUS workshop on Oral History & Technology**
  Riccardo Del Gratta, Monica Monachini - *About the data infrastructure in the country and how our services could fit into that & access to data, tools, metadata for the research community at large & IPR / informed consent / ethical issues*
  Arezzo, 10-12 maggio 2017

- **Université Paris-Sorbonne - Observatoire de la vie littéraire (OBVIL)**
  Emiliano Giovannetti - *Seminario di presentazione delle attività del gruppo di Literary Computing* (relazione su invito)
  Parigi (Francia), 19 maggio 2017

- **Giornata di studi**
  Federico Boschetti - *Lo studio del greco nell'Europa del XV secolo: futuro e prospettive di ricerca*
  Venezia, 26 maggio 2017

- **Workshop "Totus Mundus: Un viaggio virtuale attraverso l'Atlante di Matteo Ricci (1602)"**
  Emiliano Giovannetti - *Il ruolo del Cnr nel progetto Totus Mundus: navigare tra le lingue con i remi dell'informatica e della linguistica computazionale*
  Silvia Piccini - *Il mappamondo di Ricci e la traduzione di Pasquale d'Elia a confronto: un viaggio semantico attraverso la geografia e la cosmografia dalla Cina del XVI secolo all'Italia del XIX secolo*
  Roma, Sapienza Università di Roma, Dip. di Storia Culture Religioni, 26 maggio 2017

- **MMM 2017 - 11th Mediterranean Morphology Meeting**
  Vito Pirrelli - *Transparency and predictability in Modern Greek conjugation: Implications for models of word processing*
  Cipro, 22-25 giugno 2017

- **TEAM - Theoretical and empirical approaches to microvariation**
  Simonetta Montemagni - *Round Table: Big Data, big problems?*
  Padova, Università degli Studi di Padova, Dip. di Studi Linguistici e Letterari (DiSLL), 22-24 giugno 2017

- **Ex Nihilo - "Zero Conference" della European Academy of Religion**
  Emiliano Giovannetti - *Proponente e coordinatore del panel "The Role of Technology and Computational Linguistics in the Translation of the Babylonian Talmud in Italian"* (su invito)
  Emiliano Giovannetti - *An Introduction to Traduco*
  Andrea Bellandi - *How to help translators using artificial intelligence: The Translation Memory*
  Alessandra Pecchioli - *The role of linguistic analysis in Traduco*
  Bologna, 20 giugno 2017

- **Seminario - Università Roma 3, Dip. di Lingue, Letterature e Culture Straniere**
  Presentazione delle attività del gruppo di Literary Computing
  Roma, 5 luglio 2017

- **Workshop "Korean Flagship AI Project on Emotional Intelligence", KAIST**
  Nicoletta Calzolari Zamorani (Invited Speaker)
  Daejeon (Korea), 10-11 luglio 2017

- **International Workshop "Belt and Road Forum for Language Resources"**
  Nicoletta Calzolari Zamorani (Keynote Speaker)
  Pechino (Cina), 15-16 luglio 2017

- **Workshop "Sharing Session on High Impact Publications"**
  Nicoletta Calzolari Zamorani (Invited Speaker)
  Hong Kong, Hong Kong Polytechnic University, 1 settembre 2017

- **Convegno nazionale "La Crusca torna al vocabolario, La lessicografia dinamica dell'italiano post-unitario"**
  Sebastiana Cucurullo, Eva Sassolini, Simonetta Montemagni - *Il ruolo della linguistica computazionale nella fabbrica del "vocabolario dinamico"*
  Partecipazione al dibattito seguito alla tavola rotonda operativa sul progetto PRIN, oggetto del convegno
  Firenze, Accademia della Crusca, 11 e 12 settembre 2017

- **CLARIN-PLUS workshop "Creation and Use of Social Media Resources"**
  Andrea Cimino - *Analysis of Italian Social Media Texts: from Tools and Resources to Applications*
  Kaunas, Lituania, 18-19 settembre 2017

- **Conferenza SEPLN2017 - 33rd International Conference of the Spanish Society for Natural Language Processing**
  Nicoletta Calzolari Zamorani (Invited Speaker)
  Invited Speaker alla Tavola Rotonda *"Hoja de ruta para el desarrollo de recursos lingüísticos en España: oferta y demanda"*, Workshop "*II Taller ReTeLe: Red de Recursos para Tecnologías de la Lengua"*)
  Murcia (Spagna), 19 settembre 2017

- **Conferenza stampa presso il Lincoln Center for the Performing Arts**
  Emiliano Giovannetti – *Presentazione del proprio ruolo di responsabile scientifico nello sviluppo del software Traduco* (relazione su invito)
  New York (USA), 24 ottobre 2017

- **Giornata di studi "Digital Humanities - Le scienze umanistiche e le nuove tecnologie"**
  Approcci digitali allo studio del testo letterario: le esperienze del "Literary Computing Group"
  Angelo Mario Del Grosso - *Strumenti software per lo studio e l'analisi di risorse testuali*
  Silvia Piccini - *Lessicografia e Terminologia computazionale: metodologia, modelli ed esempi*

Firenze, Dip. di Lettere e Filologia dell'Università di Firenze (DILEF), 25 ottobre 2017

- **International Workshop on Machine Learning and Natural Language Processing**
  Vito Pirrelli, Marcello Ferro - *The Readlet project: Machine Learning and Natural Language Processing platform for detecting child reading deficiencies*
  Fez (Marocco), Sidi Mohamed Ben Abdellah University, 25 novembre 2017

- **Giornata organizzata dal FOAGE "Valorizzare l'architettura contemporanea in Liguria"**
  Lucia Marconi, Alessandra Cinini, Paola Cutugno - Presentazione dei risultati di estrazione terminologica riferiti al "Censimento e schedatura di complessi di architettura moderna e contemporanea in Liguria", in collaborazione con il DAD - UNIGE e il FOA
  Genova, 11 dicembre 2017

Da segnalare anche il contributo di carattere divulgativo sulla rivista online Engramma (150, ottobre 2017) dal titolo *"Estrarre parole dalle immagini nell'era digitale: alcune osservazioni sull'Ocr storico".* Autore: Federico Boschetti.

6.1.4 Seminari interni

L'ILC promuove attività di studio e divulgazione della ricerca scientifica nei settori d'interesse attraverso una serie di incontri di natura prevalentemente interdisciplinare. Oltre ai seminari su temi specifici e alla discussione di articoli scientifici, sono organizzate lezioni tenute da esperti esterni e brevi presentazioni interne, prevalentemente legate a progetti nazionali e internazionali in corso presso l'Istituto.

Tra le iniziative organizzate nel 2017 sono da segnalare i seguenti seminari:

Data: 15 febbraio 2017
Titolo: Un lessico per la terminologia medico-botanica in occitano antico in Lemon
Relatore: Anja Weingart - Georg-August-Universität di Göttingen

La relazione presenta l'adattamento del modello Lemon (un modello per lessici come dati RDF) per un lessico multilingue e multi-alfabetico di terminologia medico-botanica in occitano antico. Il lessico è il componente principale di un sistema informatico basato su ontologie costruito e implementato nell'ambito del progetto "Dictionnaire de Termes Médico-botaniques de l'Ancien Occitan" (DiTMAO), finanziato dalla Deutsche Forschungsgemeinschaft (DFG - Fondazione per la Ricerca Tedesca) dal 2011 al 2015. Le difficoltà per la lemmatizzazione sollevata dalle particolarità del corpus (termini in latino, ebraico e scrittura araba e termini corrispondenti in altre lingue antiche, soprattutto in ebraico e arabo) possono essere perfettamente superate estendendo le proprietà di base del modello Lemon e introducendo un vocabolario specifico di dominio.

Data: 29 marzo 2017
Titolo: Il Fattore Umano nei Sistemi di Riconoscimento Ottico dei Caratteri - OCR Proof-Reading & Eye Tracking
Relatore: Barbara Balbi, Vincenzo Broscritto, Flavia De Simone - Università Suor Orsola Benincasa (Napoli)

I sistemi di riconoscimento ottico dei caratteri (OCR) sono dei software atti alla conversione di un'immagine contenente caratteri a stampa in un testo digitale modificabile con qualsiasi tipo di editor. Non sempre l'output di questi programmi corrisponde al testo originale, in quanto alcuni fattori (quali l'integrità della carta e del testo o l'uso di particolari caratteri) possono portare questi sistemi a compiere delle valutazioni errate durante il processo di conversione. Per aumentare l'accuratezza dei risultati è quindi necessario procedere alla correzione manuale del testo. Lo studio presentato mira ad individuare le strategie visive per il riconoscimento di questi errori tramite l'impiego di un "eye tracker", uno strumento che registra i movimenti degli occhi e le fissazioni sullo schermo nel corso degli esperimenti di correzione del testo. Sono stati esposti i metodi impiegati e i primi risultati del progetto-pilota.

Data: 6 aprile 2017
Titolo: Stemmatology a bird's eye view
Relatore: Armin Hoenen - CEDIFOR, Goethe University Frankfurt

La Stemmatologia, la scienza di stabilire le relazioni genealogiche tra i testi manoscritti non ha neppure 200 anni e ha tuttavia già visto dibattiti e discussioni allettanti. La sua transizione nel digitale sfida ancora una volta queste basi vacillanti. Presentazione di questa ricca storia che influenza profondamente la nostra comprensione dei testi storici e della loro trasmissione e illustrazione delle nuove frontiere della filologia computazionale nel campo della stemmatologia.

Data: 6 giugno 2017
Titolo: Arabo e Digital Humanities: standard, applicazioni e practices
Relatore: Giuliano Lancioni - Università degli Studi Roma Tre

Presentazione dei progetti del gruppo Arabic and Digital Humanities, coordinato dalla cattedra di Lingua e Letteratura Araba dell'Università degli Studi "Roma Tre". Le linee guida del gruppo evolvono a partire da un progetto comune, ovvero l'adozione di strumenti digitali per lo studio della storia e della cultura araba attraverso strumenti linguistici. I progetti interessano Universal Dependencies, Text Encoding e lemmatizzazione applicati a corpora in arabo classico, standard e informale.

Data: 6 luglio 2017
Titolo: Bright ideas, smart colors - metaphor detection and grading through neural networks
Relatore: Yuri Bizzoni - Göteborg University

Presentazione dei risultati di due esperimenti per il riconoscimento automatico delle metafore effettuati nell'ambito di uno studio sul linguaggio figurativo. Il primo esperimento, in cui è stata utilizzata una rete neuronale interamente collegata per distinguere espressioni aggettivo-nome metaforiche dalle espressioni letterali, ha consentito di ottenere risultati che superano lo stato dell'arte. Si tratta di un approccio che consente di distinguere le metafore dalle frasi letterali con alta precisione, determinando anche un "grado di metaforicità" per una determinata espressione. Nel secondo esperimento è stata usata una profonda combinazione di reti neurali convolutive e di memoria a lungo termine per classificare possibili parafrasi di una data metafora su una scala da 1 a 5. A differenza di approcci precedenti, il metodo è stato testato su diverse categorie grammaticali e sono state introdotte variazioni sintattiche e stilistiche nelle parafrasi candidate. Il punto fondamentale dello studio è che per affrontare ragionevolmente la figuratività è necessario assumere un approccio sfumato: qualsiasi testo ha un grado di figuratività e un certo numero di interpretazioni possibili.

Data: 13 luglio 2017
Titolo: Significati e semantica dei disturbi del comportamento alimentare: uno studio su blog e colloqui clinici
Relatori: Luigi Enrico Zappa, Alessandro Chinello - Università Milano Bicocca - Fondazione Maria Bianca Corno (Monza)

I disturbi del comportamento alimentare (DCA) rappresentano un gruppo di patologie particolarmente diffuse tra le ragazze adolescenti e le giovani donne. Le tematiche centrali di questi disturbi riguardano la paura di ingrassare, le diete restrittive, una relazione distorta con la propria immagine corporea e una difficoltà nel riconoscimento delle proprie emozioni. Attraverso alcuni studi preliminari su corpus linguistici provenienti da social-app e colloqui clinici, sarà possibile esplicitare importanti vissuti, significati e pensieri che caratterizzano il vocabolario di questa popolazione clinica. Nello specifico, sono mostrati i risultati di due studi riguardanti alcune esperienze specifiche: i blog proAna e l'esperienza della gravidanza di pazienti anoressiche. Nel primo caso, sono state mostrate le relazioni tra i contenuti linguistici raccolti da diversi blog proAna, cioè analizzando corpus provenienti da spazi virtuali che contengono commenti e pensieri di pazienti con DCA o a rischio di DCA. Secondariamente, sono stati mostrati i risultati di interviste semi-strutturate riguardanti l'esperienza della gravidanza di pazienti anoressiche attraverso un approccio fenomenologico. L'incontro è stato l'occasione per una discussione comune e per valutare eventuali nuove prospettive e modalità di analisi dei corpus con lo scopo di tipizzare il vocabolario anoressico.

Data: 5 ottobre 2017
Titolo: Statistical Learning and Reading (in spite of arbitrariness?)
Relatore: Davide Crepaldi - SISSA (Trieste)

È ampiamente accettato che il linguaggio umano sia caratterizzato da arbitrarietà: non c'è nulla nelle forme dei segni che sia intrinsecamente legato ai loro significati. Nel corso del seminario è stato dimostrato come la morfologia rompa l'arbitrarietà segno-significato e stabilisca regolarità probabilistiche che il cervello cattura e usa attivamente durante l'elaborazione linguistica. Sono stati forniti i dati relativi a questa ipotesi dal 'masked priming' nelle persone bilingui, dai modelli di monitoraggio degli occhi nei lettori in via di sviluppo e dai modelli semantici basati sulla co-occorrenza delle parole.

7 Attività editoriali

Le attività editoriali dell'Istituto riguardano sia la direzione scientifica di riviste sia la partecipazione a comitati scientifici e ad attività redazionali.

Direzione scientifica di riviste:

- **ITALIAN JOURNAL OF COMPUTATIONAL LINGUISTICS (IJCOL)**

Lanciata nel 2015 come iniziativa editoriale dell'Associazione Italiana di Linguistica Computazionale (AILC), la Rivista Italiana di Linguistica Computazionale si propone come forum aggiornato di discussione attorno alla Linguistica Computazionale, con l'obiettivo di alimentare sinergie tra studi legati ad aree diverse del trattamento automatico del linguaggio. IJCOL si propone come continuazione ideale della rivista Linguistica Computazionale, fondata nel 1981 da Antonio Zampolli e non più pubblicata dal 2006. La rivista copre temi che ruotano attorno a linguaggio e computazione, affrontati da prospettive diverse, ad esempio: trattamento e apprendimento automatico del linguaggio; modelli computazionali del linguaggio, della cognizione e della variazione linguistica; acquisizione di conoscenza da testi; costruzione di risorse linguistiche; sviluppo di infrastrutture per l'interoperabilità e l'integrazione di risorse e tecnologie linguistiche; diverse applicazioni del trattamento automatico del linguaggio (Information Extraction, Question Answering, sommarizzazione automatica e traduzione automatica, ecc.).

Tipo: Rivista peer-reviewed, open access
Periodicità: semestrale
Copertura: Volume 1 (2015) - Volume 3 (2017)
ISSN: 2499-4553
Direzione Scientifica: Roberto Basili, ***Simonetta Montemagni***
Segreteria di redazione: ***Sara Goggi***
Casa editrice: Accademia University Press (www.aaccademia.it) per l'Associazione Italiana di Linguistica Computazionale (www.ai-lc.it)

- **LANGUAGE RESOURCES AND EVALUATION (LRE)**

La prima rivista internazionale dedicata all'acquisizione, alla creazione, all'annotazione e all'uso delle risorse linguistiche, insieme a metodi per la valutazione delle risorse, delle tecnologie e delle applicazioni. Le risorse linguistiche includono dati linguistici e descrizioni in forma leggibile dal computer utilizzate per assistere e incrementare le applicazioni di elaborazione linguistica, quali corpora e lessici dello scritto o del parlato, risorse multimodali, grammatiche, basi di dati e dizionari terminologici o a dominio specifico, ontologie, basi di dati multimediali ecc., come pure strumenti software di base per la loro acquisizione, la loro preparazione, la loro annotazione, la loro gestione,

il loro adattamento e il loro utilizzo. La valutazione delle tecnologie linguistiche consiste nel valutare lo stato dell'arte di una determinata tecnologia confrontando diversi approcci a un dato problema, nel valutare la disponibilità di risorse e tecnologie per una data applicazione e nel valutare l'utilizzabilità del sistema e la soddisfazione degli utenti.

Tipo: Rivista
Periodicità: Trimestrale
Copertura: Volume 1 (1967) - Volume 51 (2017)
ISSN Stampa: 1574-020X
ISSN Online: 1574-0218
Direzione Scientifica: Nancy Ide, ***Nicoletta Calzolari Zamorani***
Assistant Editor: ***Sara Goggi***
Casa editrice: Springer, Netherlands (www.springer.com)

- **LINGUE E LINGUAGGIO**

È un luogo di discussione di respiro internazionale sulla linguistica generale e teorica, con una particolare attenzione alle aree di interazione con altre discipline, dalla psicologia alle neuroscienze, dall'informatica alle scienze cognitive. Pubblica saggi in inglese e in italiano operando nel rispetto dei più alti standard qualitativi. La rivista si occupa centralmente di teoria del linguaggio, dei vari livelli dell'analisi linguistica (fonologia, morfologia, sintassi, semantica, pragmatica) e della loro interazione, ma anche di linguistica computazionale, acquisizionale, variazionale e diacronica e di storia del pensiero linguistico. Lingue e linguaggio pubblica contributi di natura sia teorica che descrittiva ed è aperta ai diversi orientamenti teorici offerti dalla linguistica contemporanea.

Tipo: Rivista
Periodicità: Semestrale
Copertura: Volume I (2002) - Volume XVI (2017)
ISSN: 1720-9331
Direzione Scientifica: ***Vito Pirrelli****, Sergio Scalise*
Segreteria di redazione: ***Claudia Marzi***
Casa editrice: Società Editrice Il Mulino, Bologna (www.mulino.it)

Numerosi ricercatori dell'ILC sono membri di comitati scientifici e redazionali di riviste nazionali e internazionali, così come di collane editoriali, tra i quali:

- *ACM Computing Reviews*, ACM
- ACM *Journal on Computing and Cultural Heritage (JOCCH)*, ACM
- ACM *Transactions on Internet Technology (TOIT)*, ACM
- *AI Communications* - journal on Artificial Intelligence, IOS Press
- Collana di *Cultura Digitale*, Pacini Editore
- Collana editoriale *Language Variation*, Language Science Press
- *Computational Linguistics*, MIT Press Journals
- *DSH Digital Scholarship in the Humanities*, Oxford University Press
- *Entropy*, MDPI
- *Frontiers Human Neuroscience*, Frontiers
- *International Journal of Computer Science and Technology (IJCST)*, Springer
- *Italian Journal of Computational Linguistics (IJCOL)*, Accademia University Press
- *Italian Journal of Linguistics*, Pacini Editore
- *Journal of King Saudi University – Computer and Information Sciences*, Elsevier
- *Language Resources and Evaluation (LRE)*, Springer

- *Lexis-Poetica, retorica e comunicazione nella tradizione classica*, Adolf Hakkert Editore
- *Lingua*, Elsevier
- *Linguistica e Glottodidattica Sperimentale*, Pavia University Press
- *Materiali Linguistici*, Franco Angeli
- Rivista *Frontiers in Digital Humanities - Section Digital Literary Studies*, Frontiers
- *Sensor*, MDPI
- *The Mental Lexicon*, John Benjamins
- *Umanistica Digitale*, rivista della Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), ABIS – AlmaDL

Ricercatori ILC si sono occupati anche della curatela di Special Issues ("Linking, Integrating and Extending Wordnets" in *Linguistic Issue in Language Technology - LiLT*; "Collaboration and Computing for Under-resourced Languages" in *Language Resources and Evaluation)* e di atti di convegno (*Language Resources and Evalution Conference)*; hanno collaborato come consulenti linguistici per collane editoriali ("*Città in gioco*" di Franco Cosimo Panini Editore, una collana di guide-gioco dedicata ai più piccoli per scoprire le città italiane v. www.cittaingioco.it), come *proofreader* (*Language Science Press*), come *external reviewer* per diverse riviste (*Information - MDPI; Information Sciences - Elsevier; serie Morphological Investigations, J. Blevins, P. Milin, M. Ramscar (eds), Language Science Press*) o come revisori di libri ("*The semantic transparency of English compound nouns", di Martin Schäfer*).

8 Attività di terza missione

8.1 Partecipazione a organismi tecnico-scientifici e normativi

- **Federico Boschetti**
  Portavoce per il CNR-ILC del WG2 di DARIAH-IT
  (*I workpackages originari erano tre: WG1: Governance; WG2: Evaluation; WG3: Mission. Attualmente i work-packages sono stati ridotti a due: WG1: Governance; WG2: Mission*)
- **Nicoletta Calzolari Zamorani**
  Chair del Comitato ISO/TC37/SC4 on Language Resource Management
  Membro dell'Advisory Group ISO/TC 37/AG
- **Riccardo Del Gratta**
  Membro di Clarin Centre Assessment Committee (CAC)
  Membro di Clarin Standing Committee on CLARIN Technical Centers (SCCTC)
- **Monica Monachini**
  Incarichi in Comitati per la definizione di standard e norme tecniche:

  UNI – Ente Nazionale Italiano di Unificazione
  - Rappresentante designata dal CNR nel gruppo di lavoro UNI/CT 014 Terminologia della Commissione Tecnica Documentazione e Informazione Automatica.
  - Delegato italiano UNI con diritto di voto in rappresentanza dell'Italia all'interno del Comitato ISOTC37/SC4 Language Resource Management

  ISO TC37/SC4 – Organizzazione Internazionale di Standardizzazione
  - Co-project leader nel gruppo di lavoro ISOTC37/SC4/WG4 LMF standard per risorse lessicali
  - Esperto all'interno del gruppo di lavoro ISOTC37/SC4 MetaData e ISOCAT

  Incarichi Strategici:

  Meta-Net (www.meta-net.eu)

- Membro dell'Executive Board di Meta-net, una rete di eccellenza composta da 60 centri di ricerca appartenenti a 34 Paesi che mira a costruire le basi tecnologiche per una società europea dell'informazione multilingue

CLARIN – Common Language Resource and Technology Infrastructure

- Coordinatore Nazionale di CLARIN
- Membro del National Coordinator Forum
- Membro del Comitato dedicato allo studio della Interoperabilità in CLARIN
- Membro del Comitato Strategico sui rapporti tra il National Coordinator Forum e il Board dei Direttori di CLARIN

8.2 Partecipazione ad Associazioni e Comitati scientifici

Ricercatori dell'ILC rivestono ruoli di rilievo in associazioni e comitati scientifici di livello nazionale ed europeo:

Federico Boschetti

- Membro del Comitato Direttivo dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD)
- Membro del Comitato Scientifico del VII Convegno Annuale AIUCD "Cultural Heritage in the Digital Age. Memory, Humanities and Technologies" (http://www.aiucd2018.uniba.it)

Nicoletta Calzolari Zamorani

- Presidente Onorario dell'Associazione ELRA (European Language Resources Association)
- Membro permanente dell'ICCL (International Committee of Computational Linguistics)
- Membro del Forum TAL
- Membro del Board della UNDL Foundation (Universal Networking Digital Language Foundation)
- Membro dell'International Advisory Committee of the Korean Emotion-based Dialogue Flagship Project
- Membro dell'Advisory Board del progetto FET ODYCCEUS (Opinion Dynamics and Cultural Conflict in European Spaces)
- Vice Presidente della Associazione META-TRUST (http://www.meta-net.eu/meta-trust), "the legal person of the network of excellence META-NET and the Multilingual Europe Technology Alliance - META")
- Presidente dell'Associazione Internazionale PAROLE
- External referee per la Royal Netherlands Academy of Arts and Sciences (KNAW), Amsterdam, The Netherlands.

Felice Dell'Orletta

- Socio-fondatore e membro del Direttivo dell'Associazione Italiana di Linguistica Computazionale (AILC)
- Membro del comitato scientifico, presieduto dalla Prof.ssa Savina Raynaud, del Centro Interdisciplinare di Ricerche per la Computerizzazione dei Segni dell'Espressione (CIRCSE) dell'Università Cattolica del Sacro Cuore (quadriennio 2013/2014-2016/2017)

Monica Monachini

- CLARIN (Common Language Resources and Technology Infrastructure), Coordinatore Nazionale di CLARIN-IT

Simonetta Montemagni

- Socio-fondatore e vice-presidente dell'Associazione Italiana di Linguistica Computazionale (AILC)
- Membro del Language Resources Board (LRB) del Tender European Language Resources Coordination (ELRC, www.lr-coordination.eu) e punto di riferimento nazionale Tecnologico (Technology National Anchor Point- NAP) per l'azione ELRC
- Membro del gruppo di lavoro italiano nell'ambito dell'iniziativa internazionale Universal Dependencies (UD)
- Membro del Board dell'Associazione ELRA (European Language Resources Association)

Emiliano Giovannetti

- Membro dello Special Interest Group on Language Technologies for the Socio-Economic Sciences and Humanities (SIGHUM) dell'ACL

Claudia Soria
- Vice-presidente dell'European Language Equality Network (ELEN)
- Co-chair ed ELRA liaison representative di SIGUL (*ELRA-ISCA Special Interest Group on Under-Resourced Languages*)
- Membro dell'International Advisory Board della Foundation for Endangered Languages (FEL)
- Membro del Comitato Esecutivo di MAAYA - The World Network for Linguistic Diversity

8.3 Partecipazione a comitati di valutazione

Ricercatori ILC hanno partecipato a campagne di valutazione per progetti, workshop e conferenze nazionali e internazionali, nonché a commissioni dottorali e a comitati per la valutazione e selezione di personale di istituzioni nazionali ed estere e per la Valutazione della Qualità della Ricerca.

Nicoletta Calzolari
- Membro del Consiglio di Dottorato di Linguistica dell'Università di Pisa

Claudia Marzi
- Membro dell'Albo dei Revisori MIUR-Cineca: REPRISE (register expert peer reviewer for Italian Scientific Evaluation) – Area SH
- Valutatore di due proposte di progetto nell'ambito del finanziamento "FARE 2016" MIUR (2017)

Simonetta Montemagni
- Revisore per la VQR 2011-14, Area 10 (Scienze dell'Antichità, Filologico-Letterarie e Storico-artistiche)

Vito Pirrelli
- Revisore per la VQR 2011-14, Area 10 (Scienze dell'Antichità, Filologico-Letterarie e Storico-artistiche)
- Valutatore di progetti per Bandi interni dell'Università degli Studi di Udine
- Membro del comitato di valutazione dei seguenti workshop/conferenze:
  - ParadigMo 2017 - First Workshop on Paradigmatic Word Formation Modeling
  - ISMo 2017 - International Symposium of Morphology
  - ICALP'17 - International Conference on Arabic Language Processing
  - DeriMo 2017 - The First International Workshop on Resources and Tools for Derivational Morphology

8.4 Valorizzazione dei risultati e trasferimento tecnologico

Nel corso del 2017 sono proseguite le attività volte alla valorizzazione dei risultati della ricerca dell'ILC a beneficio della società e delle imprese. Esse includono contatti e collaborazioni sia con imprese italiane e internazionali sia con enti pubblici locali, finalizzate al trasferimento tecnologico sia delle competenze acquisite in diversi settori della linguistica computazionale sia di risorse linguistiche e tecnologie linguistico-computazionali. In tal modo è stato possibile diffondere le conoscenze e le competenze acquisite in diversi ambiti della linguistica computazionale e trasferire le risorse linguistiche e le tecnologie acquisite nel settore. Grazie alla collaborazione con specialisti di altre discipline, il personale ILC ha infatti sviluppato metodi e tecnologie innovative che possono trovare applicazione in numerosi ambiti, ad esempio: Pubblica Amministrazione; Istruzione e Formazione; Sanità; Patrimonio Culturale e Turismo; Imprese; Terzo Settore.

Alcune delle attività svolte nel corso del 2017:

- integrazione con la piattaforma software sviluppata dalla società Hyperborea di strumenti per l'analisi dei testi estratti da Twitter per: analisi linguistica; estrazione di entità rilevanti: terminologia ed entità nominate; sentiment analysis; identificazione dei testimoni; organizzazione gerarchica dei tweet rispetto al contenuto;
- l'attività di validazione della piattaforma software/hardware per il monitoraggio dell'efficienza di lettura presso la Scuola Primaria "G. Rodari" di Montecalvoli (PI), dell'Istituto Comprensivo "G. Carducci" di S. Maria a Monte (PI);

- integrazione all'interno della piattaforma di Text Mining di Lottomatica S.p.A. di strumenti per l'estrazione di testi da Twitter e Facebook, l'analisi linguistica e la sentiment analysis;
- integrazione all'interno della piattaforma sviluppata dal GruppoMETA di strumenti per l'analisi della leggibilità di testi scolastici;
- accordo di collaborazione scientifica con INDIRE per lo sviluppo di analisi del dominio educativo e formativo mediante metodi e tecniche di trattamento automatico del linguaggio
- recupero e conversione in formato standard XML/TEI di importanti archivi testuali con formati di rappresentazione obsolescenti. L'ILC ha continuato a contribuire in modo significativo alla conservazione e valorizzazione del patrimonio culturale "invisibile" italiano, in particolare, linguistico, filologico, storico e letterario.

Nell'ambito del progetto "AFTTER - Alta Formazione per il Trasferimento Tecnologico degli Enti di Ricerca", finalizzato all'individuazione degli ambiti di ricerca che possono portare ad una maggiore integrazione del sistema degli Enti Pubblici di Ricerca toscani con il sistema produttivo regionale, è stato organizzato il seminario:

Tecnologie della lingua e trasferimento tecnologico: prime sperimentazioni, potenzialità e applicazioni
Pisa, Area della Ricerca del CNR, 19 dicembre 2017

Programma:

Simonetta Montemagni - *Introduzione*
Felice Dell'Orletta - *NLP at work: algoritmi, tecnologie e scenari applicativi*
Eva Sassolini - *Trattamento automatico del testo: rassegne stampa specialistiche*
Davide Albanesi, Andrea Bellandi - *TRADUCO: ricerca e sviluppo oltre il prototipo*

8.5 Attività di Public Engagement

Nel corso del 2017 sono state condotte diverse attività di divulgazione scientifica con valore educativo, culturale e di sviluppo della società, che hanno attratto l'interesse di un vasto pubblico. Scopo principale di tali iniziative sono la disseminazione dei risultati delle ricerche e un maggior coinvolgimento di tutti gli stakeholder.

8.5.1 Rapporti con le Istituzioni

Ricercatori dell'ILC hanno partecipato a tavoli di lavoro presso il MIUR, il Senato e la Regione Toscana come esperti di trattamento automatico del testo ed estrazione di conoscenza.

8.5.2 Eventi pubblici

- **FORUMTAL 2017 - Tecnologie Vocali e del Linguaggio Naturale per i Beni Culturali**
  Giornata di studi sulle tecnologie vocali e dell'elaborazione naturale della lingua nell'ambito della fruizione dei beni culturali, della loro catalogazione e in generale dell'uso di tali tecnologie per i beni culturali.
  Roma, Università degli Studi Roma TRE, 19 gennaio 2017

- **Manifestazione "Tempo di Libri"**
  Simonetta Montemagni, Emiliano Giovannetti - *L'occhio del traduttore. Di linguaggi e dispositivi. Il progetto 'Traduco' e il Talmud* (su invito)
  Illustrazione della piattaforma Traduco sviluppata nell'ambito del Progetto Talmud ad un evento di presentazione dei volumi editi da Giuntina
  Milano Rho, 20 aprile 2017

- **Workshop "Tecnologie della Lingua per la Scuola Digitale" - Fiera Didacta Italia**
  Firenze, Fiera Didacta Italia, 28 settembre 2017
  Il workshop è stata l'occasione per illustrare come le tecnologie della lingua possano essere di aiuto nell'affrontare le nuove sfide scientifiche e tecnologiche legate alla didattica e all'apprendimento. Sono stati presentati i risultati

di sperimentazioni basate sull'utilizzo di tecnologie e risorse linguistiche condotte in scuole italiane da ricercatori dell'ILC in collaborazione con altri Istituti del CNR, Università e gruppi aziendali, quali Mondadori Education e GruppoMeta. Il workshop è stato organizzato nell'ambito della prima edizione di Fiera Didacta Italia, la versione italiana di Didacta - Die Bildungsmesse, un appuntamento fieristico dedicato all'istruzione che si tiene in Germania da oltre 50 anni. Fiera Didacta Italia, il cui partner scientifico è l'Istituto Nazionale di Documentazione, Innovazione e Ricerca Educativa (INDIRE), è rivolta agli operatori dei settori dell'istruzione e della formazione professionale. L'evento ha rappresentato un importante momento di riflessione e di scambio di esperienze per la crescita del nostro sistema educativo e ha contribuito a promuovere il dibattito fra tutti gli attori coinvolti.

Il programma:

Le tecnologie della lingua per la scuola digitale: scenari applicativi, problemi e possibili soluzioni
Simonetta Montemagni (ILC-CNR)

Come evolvono le abilità di scrittura nel primo biennio della scuola secondaria di primo grado?
Giulia Venturi (ILC-CNR), Alessia Barbagli (Università La Sapienza), Patrizia Sposetti (Università La Sapienza)

Lo studio delle lingue classiche e le tecnologie della lingua. Il ritorno del greco in Occidente.
Federico Boschetti (ILC-CNR), Paola Tomè (Liceo Classico Marco Polo), Antonella Trevisiol (Liceo Classico Marco Polo)

Tecnologie della lingua per nuovi modelli di (Semantic) Liquid Book nella scuola digitale
Paolo Ongaro (GruppoMeta), Felice Dell'Orletta (ILC-CNR), Fabio Ferri (Mondadori Education)

Monitoraggio delle abilità di lettura: leggere con Readlet
Claudia Cappa (IFC-CNR), Vito Pirrelli (ILC-CNR)

- **BRIGHT 2017 - Notte Europea dei Ricercatori**
  Pisa, Area della Ricerca di Pisa - CNR, 29 settembre 2017

  In occasione dell'evento l'Istituto ha organizzato i seguenti seminari:

  PERCORSO: SALUTE E BENESSERE; NUOVE TECNOLOGIE (ILC+IFC)
  READLET - leggere per capire: il prototipo!
  Come realizzare un'infrastruttura tablet-software per la valutazione automatica dell'efficienza di lettura (Reading Efficiency Parameter, REP) che misuri contemporaneamente velocità di lettura silente e capacità di comprensione del testo?

  PERCORSO: NUOVE TECNOLOGIE
  La linguistica computazionale ai tempi dell'intelligenza artificiale
  Come leggere un testo riconoscendone gli elementi informativi rilevanti e come scrivere un elaborato impiegando strutture linguistiche che lo rendano chiaro, semplice e comprensibile?

  PERCORSO: PATRIMONIO CULTURALE
  Vita quotidiana all'ombra delle Cupola del Brunelleschi
  Una selezione ragionata di schede estratte dall'archivio dell'Opera di Santa Maria del Fiore di Firenze, relative alla costruzione della Cupola di Brunelleschi, sulle quali sono state applicate elaborazioni software.

  Per i contenuti degli interventi cfr. http://www.ilc.cnr.it/it/content/interventi-ilc-bright-2017

- **Eventi di presentazione del progetto Traduzione Talmud Babilonese nell'ambito di iniziative nazionali e internazionali**
  Il lavoro di ricerca e sviluppo che l'Istituto ha condotto nella realizzazione del software Traduco è stato presentato a Washington (in particolare, alla Library of Congress) e a New York City nel novembre del 2017 nell'ambito di una missione del Progetto Traduzione Talmud Babilonese volta ad illustrare, in ambiti universitari e culturali, i risultati ottenuti. In tale contesto, sono stati stabiliti contatti volti all'applicazione di tecniche per la traduzione assistita di altri testi come, ad esempio, la Bibbia Ebraica per una sua traduzione in cinese.

8.5.3 Siti web e social media

Nel 2017 è proseguito lo sviluppo del sito web dell'Istituto. Il sito, disponibile in italiano e in inglese, è stato costantemente aggiornato con la segnalazione delle attività realizzate e arricchito con le informazioni relative ai progetti, alle collaborazioni, alla produzione scientifica, agli eventi e alle ultime notizie.
Per divulgare le attività svolte, sono stati costantemente aggiornati anche i siti specifici dei diversi laboratori attivi presso l'Istituto e il sito web del gruppo di ricerca "Literary Computing" (http://licolab.ilc.cnr.it).
Le informazioni relative ai progetti a cui ha preso parte l'ILC sono state rese disponibili sui relativi siti, oltre che in alcune pagine specificatamente create sui social media, ad esempio la pagina Facebook e l'account Twitter "Diversità Linguistica" e la pagina Facebook e l'account Twitter "DLDProject". Nel caso del progetto DLDP è stata prevista anche la progettazione, preparazione e diffusione di una newsletter a scopo divulgativo, con cadenza mensile.

8.5.4 Trasmissioni radiofoniche

- **Radio Aula 40 – Punto Radio**
  http://radioaula40.cnr.it
  Puntata del 23 febbraio 2017- *Arte: un bene da curare...con l'aiuto della scienza*
  Relatore: Emiliano Giovannetti

8.5.5 Iniziative di interazione con scuole

Nel corso del 2017 l'Istituto ha collaborato con diverse scuole per avvicinare gli studenti alla Linguistica Computazionale e, in particolare, ai temi delle Digital Humanities e della filologia collaborativa e cooperativa:

- **Liceo Scientifico Statale "Pietro Paleocapa" di Rovigo** (convenzione)
  Tirocinio formativo - progetto formativo individuale
  Tutor aziendale: Monica Monachini
  Periodo: 29/05/2017 e 22/06/2017 (10 ore)

- **IC "San Marco dei Cavoti" (BN)**
  Accordo di collaborazione volto a supportare le attività di disseminazione e di formazione, teorica e pratica, relative al Progetto Atelier creativo "Officina del pensiero" riguardante lo sviluppo di creatività, di capacità critiche e di analisi del pensiero logico computazionale e algoritmico, nonché l'instaurarsi di situazioni di apprendimento collaborativo con approcci metacognitivi che favoriscono, da un lato, il potenziamento di studenti con particolari attitudini disciplinari e, dall'altro, l'incremento dell'inclusività degli studenti con disabilità nel gruppo dei pari e il sostegno degli studenti stranieri.
  Referente: Angelo Del Grosso

- **Giornate di studio sul tema: "Dal vocabolario alla rete semantica: cooperare alla costruzione di Ancient Greek WordNet"**
  - Liceo Classico "Marco Polo", Venezia, 30/01/2017
  - Liceo Classico XXV Aprile, Portogruaro (VE), 03/02/2017
  - Liceo Classico E. Montale, San Donà di Piave (VE), 06/02/2017
  Relatore: Federico Boschetti

- **Liceo Tommaso Gargallo di Siracusa**
  Convenzione per attività di Alternanza Scuola-Lavoro ai sensi della legge 13/07/2017 n. 107 (art. 3/33)
  Tutor: Federico Boschetti

- **Istituto I.I.S. "Medi-Livatino"** di San Bartolomeo in Galdo (BN)
  Realizzazione di percorsi di Alternanza Scuola Lavoro; sviluppo di moduli per l'attuazione di Progetti (PNDS, PON…) e in particolare nell'area dell'innovazione tecnologica applicata alle discipline curricolari, dell'orientamento, della cittadinanza globale
  Referente: Angelo Del Grosso

8.5.6 Biblioteca

L'ILC ospita una biblioteca ad accesso libero, previo appuntamento. I cataloghi, costantemente aggiornati, sono liberamente consultabili on line in BIBLOS (http://www.biblos.cnr.it/05_ILC.html) e in MOP (http://mop.isti.cnr.it).
La Biblioteca è costituita da circa 6.400 volumi, pubblicati a partire dal secondo dopoguerra a oggi. Il patrimonio librario include circa 3.000 monografie, più di 2.000 volumi di letteratura grigia (rapporti di ricerca nazionali e internazionali, atti delle conferenze del settore della linguistica computazionale organizzate dalle principali associazioni internazionali) e una raccolta di 24 titoli tra i principali periodici italiani e stranieri del settore.
Al patrimonio librario della Biblioteca ILC si aggiunge il "Fondo Antonio Zampolli", lasciato in eredità dal fondatore dell'Istituto. Si tratta di una raccolta di inestimabile valore storico, costituita da 1.400 testi specialistici, alcuni dei quali rappresentano esemplari unici in Italia e tra i pochi nel mondo che documentano la nascita della Linguistica Computazionale.

9 English Summary

The main **AREAS OF RESEARCH** at the Institute of Computational Linguistics "Antonio Zampolli" (ILC) of the Italian National Research Council (CNR) are:

- **Natural Language Processing and Knowledge Extraction**
- **Digital Humanities**
- **Resources, Standards and Research Infrastructures**
- **(Bio-)Computational Models of language use**

ILC research is carried out in the framework of research **GROUPS** and **LABORATORIES**:

- **ComPhys Lab** - Design and development of (bio-)computational models of language behaviour with the aim to investigate the juncture of grammatical competence, language usage and neurophysiological and psycho-cognitive correlates of verbal communication and language and communication disorders (http://www.comphyslab.eu)
- **CoPhiLab** - Formal description of the entities and relations involved in the domain of collaborative philology; creation of digital resources; design and implementation of software components, in particular for classical languages (http://cophilab.ilc.cnr.it:8080/CoPhiLabPortal)
- **ItaliaNLP Lab** - Design and development of models, methods, algorithms and technologies for Natural Language Processing, with particular emphasis on the Italian language. Main lines of activity: multi-level linguistic annotation of texts; domain–specific information extraction; development of application prototypes (http://www.italianlp.it)
- **LaRI Group** (LARI -Language Resources and Infrastructures, http://lari.ilc.cnr.it), that aims to foster research in the field of language engineering and to optimize the production of language resources
- **Literary Computing Group** (http://licolab.ilc.cnr.it), that aims to apply Computational Linguistics and Knowledge Engineering to Humanistic Texts and have the following main research topics: textual models; text ontologies; computer-assisted translation; object-oriented design and development of applications for literary computing.

At the end of 2017, ILC **STAFF** included researchers, technologists and technical staff with different backgrounds distributed as follows: **28** units of permanent and temporary staff, and **12** units represented by research fellows and associated personnel.

Research activities carried out during 2017 have led to significant results, consolidating and extending the national and international visibility of the Institute. The network of scientific contacts and research collaborations has been extended as testified by numerous scientific collaboration agreements with Universities and other Institutions signed during the year.

ILC actively participates in two important European **RESEARCH INFRASTRUCTURES**, with different roles:

- **CLARIN-IT** (www.clarin-it.it), the Italian node of *CLARIN - Common Language Resources and Technology Infrastructure* (www.clarin.eu). As CLARIN ERIC member and Leading NC partner, ILC creates and provides access to digital language data collections and digital tools and expertise for researchers. ILC has set up the 1st CLARIN-IT National Centre: ILC4CLARIN (ilc4clarin.ilc.cnr.it), that assists scholars in documenting resources using harmonized metadata descriptions, depositing resources and sharing them with clear licensing policies and secure federated access to protected resources.
- **DARIAH-IT** (http://it.dariah.eu), the Italian node of DARIAH-EU (Digital Research Infrastructure for the Arts and Humanities) *DARIAH-EU - Digital Research Infrastructure for the Arts and Humanities* (www.dariah.eu). ILC is partner of this pan-european infrastructure for arts and humanities scholars working with computational methods that supports digital research as well as the teaching of digital research methods.

EXTERNALLY FUNDED RESEARCH PROJECTS

During 2017, ILC was involved in the following externally funded projects:

- **3 European projects/programmes:**

ILC Role	Project
Coordinator	• DLDP - The Digital Language Diversity Project, www.dldp.eu1
Partner	• PARTHENOS - Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies, www.parthenos-project.eu
Subcontractor	• ELRC-European Language Resource Coordination, http://lr-coordination.eu. ILC is Technology National Anchor Point for Italy and is member of the Language Resources Board of the ELRC.

- **6 national projects:** TALMUD; TOTUS MUNDUS; VOCI DELLA GRANDE GUERRA; 1 FIRB Project - Future in Research: MODELACT; 1 research project of national interest (PRIN): CHROME; 1 PON "Ricerca e Competitività" project: CITTÀ EDUCANTE
- **1 international project** as a partner: DiTMAO - Dictionary of Old Occitan medico-botanical terminology
- **1 bilateral agreement** for scientific and technological cooperation with the Sidi Mohamed Ben Abdellah University (Morocco)
- **3 regional projects**: PERFORMA ARCO CNR (Personalizzazione di pERcorsi FORMativi Avanzati); SMARTNEWS; UBIMOL
- **4 CNR projects:** CLAVIUS - Clavius on the Web; NINFA - iNtelligent Integrated Network For Aged people; SM@RTINFRA-SSHCH; ItaliaNLP – WAFI Project on Cyber Intelligence
- **5 other projects:** COMMERCE NUMÉRIQUE; ENCORE - ENgaging Content Object for Reuse and Exploitation of cultural resources; GREEK STUDIES IN XVTH CENTURY EUROPE; IL TESORO DELL'ILC; MUSEO VIRTUALE DELLA MUSICA BELLININRETE

ILC researchers have won 2 **AWARDS**: First position award for the Native Language Identification Shared Task of the 12th Workshop on Innovative Use of NLP for Building Educational Applications; AILC Master Thesis Prize 2017.

More specifically, in 2017 ILC **DISSEMINATION AND OUTREACH ACTIVITIES** included:

- **workshops, conferences, seminars.** ILC researchers have participated in scientific committees of italian and international conferences, such as the *Sixth Annual Conference of the Associazione per l'informatica umanistica e la cultura digitale (AIUCD)* and the *Italian Conference on Computational Linguistics-CLIC-IT 2017*. They organized also internal seminars and have been invited for the presentation of papers, talks and communications in many national and international conferences.
- **high-education activities**, such as: courses on Computational Linguistics and Digital Humanities at Pisa University and at Venice International University (VIU); external seminars and workshops; tutoring and co-tutoring of Bachelor's, Master's and PhD thesis in Italy and abroad; traineeships programs and hosting of visiting scholars.
- **editorial activities**: scientific direction of three journals: *Italian Journal of Computational Linguistics – IJCOL; Language Resources and Evaluation* and *Lingue e Linguaggio*; different editorial activities and responsibilities in scientific committees of other journals, e.g. *The Mental Lexicon.*
- **participation in technical-scientific and regulatory bodies**, such as UNI; ISO TC37/SC4; ISO/TC 37/AG; CLARIN), **in evalutation comittees and in associations and scientific committees** like AILC, ELRC, ELRA and UNESCO.
- **exploitation of research results and technology transfer:** ILC worked for the implementation of an Italian Treebank with syntactically annotated dependencies

- **public engagement activities**: dissemination of scientific results through press and public lectures like the Bright 2017 European Night of Researchers; constant updating of the institutional website; development of websites and web services open to the community; common activities with schools and universities.
- **archival and library services:** ILC Library contains about 6,400 volumes and the "Antonio Zampolli Collection", donated by the founder of the Institute, that is made up of 1,400 specialist texts, some of which are unique examples in Italy and among the few ones in the world representing the origins of Computational Linguistics.

The table below provides a quantitative summary of ILC Research output in 2017:

ILC Research output 2017 – Overview	
Journal article	13
Book chapter/contributions in books	6
Conference proceeding	31
Edited book/proceedings	3
Scientific Direction of journals	3
Master's degree and doctoral thesis supervision	10
Teaching in university courses	6
Teaching modules in summer schools	3
Research projects	21
Research Infrastructures	2
Honors and awards	2

10 Appendice

Personale ILC

SEDE DI PISA	Profilo		N. (aggiornato al 31/12/17)
Contratto Lavoro Diritto Privato	Direttore di Istituto	Montemagni Simonetta (*dal 1/12/17 Direttore f.f.*)	1
Personale a tempo indeterminato	Dirigente di ricerca	Pirrelli Vito	1
	I° Ricercatore	Monachini Monica	1
	Ricercatore	Bartolini Roberto Boschetti Federico Dell'Orletta Felice Ferro Marcello Giovannetti Emiliano Marchi Simone Pardelli Gabriella Quochi Valeria Soria Claudia	9
	I° Tecnologo	Enea Alessandro	1
	Tecnologo	Goggi Sara Marzi Claudia	2
	CTER	Cucurullo Sebastiana Gadducci Antonella Parrinelli Vanessa Picchi Paolo Sassolini Eva Terreni Noemi	6
	Collaboratore di amministrazione	Pieri Antonella	1
Personale a tempo determinato	Ricercatore	Del Gratta Riccardo Russo Irene *(dal 1/6/17)* Cardillo Franco Alberto *(dal 2/11/17)*	3
	Tecnologo	Baroni Paola	1
	CTER	Albanesi Davide	1
		(personale in servizio il 31/12/17)	
		Totale Personale a tempo indeterminato	**21**
		Totale Personale a tempo determinato	**5**
		Totale Contratto Lavoro Diritto Privato	**1**

SEDE DI GENOVA	Profilo		N. (aggiornato al 31/12/17)
Personale a tempo indeterminato	I° Ricercatore	Marconi Lucia	1
	CTER	Cutugno Paola	1
		(personale in servizio il 31/12/17)	**2**
		Totale Personale a tempo indeterminato	

ASSEGNISTI DI RICERCA		N. (aggiornato al 31/12/17)
Sede di Pisa	Bellandi Andrea Brunato Dominique Pierina Cardillo Franco Alberto *(fino al 14/06/17)* Carlino Michela Cimino Andrea Del Grosso Angelo Mario Khan Anas Fahad Mancini Lorenzo *(fino al 18/03/17)* Nahli Ouafae Pecchioli Alessandra *(fino al 15/05/17)* Piccini Silvia Russo Irene *(fino al 19/05/17)* Venturi Giulia	9
Sede di Genova	Cinini Alessandra Lucentini Roberta *(fino al 3/3/17)*	1
	Totale Assegnisti di ricerca (dati aggiornati al 31/12/17)	**10**

ASSOCIATI	Nome	N. (aggiornato al 31/12/17)
Sede di Pisa	Calzolari Zamorani Nicoletta Sassi Manuela	2

Redazione a cura di Michela Carlino

Febbraio 2018

http://www.ilc.cnr.it