ISSN 2499-4553

IJCoL

Italian Journal of Computational Linguistics

Rivista Italiana di Linguistica Computazionale

Volume 5, Number 2
december 2019

aA ccademia university press

Registrazione presso il Tribunale di Trento n. 14/16 del 6 luglio 2016

Rivista Semestrale dell'Associazione Italiana di Linguistica Computazionale (AILC)

Associazione Italiana di Linguistica Computazionale

direttore responsabile
Michele Arnese

isbn 979-12-80136-06-0

Accademia University Press
via Carlo Alberto 55
I-10123 Torino
info@aAccademia.it
www.aAccademia.it/IJCoL_5_2

IJCoL

Volume 5, Number 2
december 2019

CONTENTS

In Memory of Emanuele Pianta's Contribution to Computational Linguistics

Bernardo Magnini*
Fondazione Bruno Kessler

Rodolfo Delmonte**
Università di Venezia

Sara Tonelli†
Fondazione Bruno Kessler

Almost eight years after his untimely death, the scientific contribution of Emanuele Pianta still appears significant to us, in particular for the variety of the topics he dealt with and for his capacity to move cross-disciplinarily between different areas of computational linguistics. Today, retracing the steps of Emanuele's scientific carrier has the meaning of rediscovering an important part of the scientific challenges that the Italian research community has faced over a period of more than twenty years. In recognition of the role he played, the Italian Association of Computational Linguistics entitled to Emanuele Pianta the annual award assigned to the best master's degree thesis in the context of Computational Linguistics, discussed in an Italian University.

1. Il percorso scientifico di Emanuele Pianta

Emanuele Pianta, scomparso nel novembre 2012 a causa di un incidente stradale, è stato uno dei ricercatori che maggiormente hanno contribuito alla crescita della Linguistica Computazionale in Italia, muovendosi con grande competenza in vari settori, dalla semantica lessicale, allo sviluppo di risorse linguistiche, all'analisi sintattica della frase, all'estrazione di informazioni da testi, in particolare entità nominate e concetti-chiave, agli algoritmi di semplificazione del testo, all'interpretazione semantica della frase, e infine aprendo nuove strade nel settore delle Digital Humanities.

Durante gli anni dell'Università al Dipartimento di Linguistica della Facoltà di Lettere e Filosofia di Padova, Emanuele matura i suoi interessi per la Linguistica Computazionale, in particolare per la generazione in linguaggio naturale. Dopo essersi laureato nel 1990 con una tesi su "Rilevanza e Rappresentazione - Preliminari Teorici a un Sistema per la Generazione Automatica del Linguaggio Naturale" presso l'Università di Padova (relatori i Professori Rodolfo Delmonte e Gianluigi Borgato), Emanuele ha collaborato con il laboratorio di Linguistica Computazionale alla Ca' Foscari di Venezia, diretto da Rodolfo Delmonte, e con l'azienda ICON di Verona, per poi passare all'Irst di Trento nel 1994, chiamato da Oliviero Stock.

L'attività scientifica di Emanuele Pianta ha attraversato circa due decenni, nel corso dei quali si è confrontato e ha dato un contributo importante allo sviluppo della

* Fondazione Bruno Kessler - Via Sommarive, 18, 38123 Povo TN, Italy. E-mail: magnini@fbk.eu
** Department of Linguistic Studies, Ca' Bembo - Dorsoduro 1075 30123 Venezia, Italy. E-mail: delmont@unive.it
† Fondazione Bruno Kessler - Via Sommarive, 18, 38123 Povo TN, Italy. E-mail: tonelli@fbk.eu

Linguistica Computazionale in Italia, in particolare sui temi delle risorse linguistiche, dell'analisi morfo-sintattica della frase, dell'estrazione di informazione da testo, della generazione a partire da contenuti strutturati, e infine, dei metodi di valutazione delle tecnologie del linguaggio. Visti i suoi numerosi interessi, durante i suoi anni di attività Emanuele ha stabilito numerose relazioni con altri ricercatori e gruppi di ricerca, sia in Italia che all'estero, ponendo le basi per progetti di ricerca che durano ancora nel tempo. Vogliamo sottolineare come Emanuele abbia sempre avuto una attitudine interdisciplinare alla ricerca, portandolo a coniugare la sua formazione linguistica con una forte attenzione ai metodi computazionali, in particolare quelli basati sull'apprendimento da dati linguistici, e anche con una non comune capacità di tradurre la ricerca in tecnologia e applicazioni.

Quando Emanuele è prematuramente scomparso l'Associazione Italiana di Linguistica Computazionale (AILC) ancora non esisteva, essendo stata fondata nel settembre del 2015. Ci sembra oggi che la sua attenzione alla ricerca multidisciplinare rappresenti bene lo spirito di AILC, nata con la missione di includere sotto un'unica iniziativa le diverse anime della Linguistica Computazionale in Italia. Intitolando a lui il premio per la miglior tesi magistrale, AILC riconosce a Emanuele Pianta l'importante ruolo svolto nell'avviare tematiche di ricerca basate sia sullo studio linguistico dei fenomeni sia sulla loro modellazione computazionale, realizzando soluzioni ancora oggi apprezzate.

Di seguito menzioniamo i principali progetti di ricerca in cui Emanuele è stato coinvolto.

Semantica Situazionale. Uno dei primi interessi di Emanuele è stata l'interpretazione semantica tramite linguaggi logici per la rappresentazione della frase in forma simbolica. Nel periodo a Venezia Emanuele ha lavorato al componente di Semantica Situazionale del sistema di analisi della lingua italiana GETARUNS (Delmonte, Bianchi, and Pianta 1992), contribuendo alla implementazione del modulo che trasferisce il contenuto del DAG (grafo diretto aciclico) con l'informazione sintattica alla Forma Logica, dopo aver elaborato la risoluzione delle referenze pronominali.

MultiWordNet. Insieme ad alcuni colleghi dell'IRST di Trento (ora Fondazione Bruno Kessler) Emanuele ha contribuito alla progettazione e alla realizzazione di MultiWordNet, la versione italiana di WordNet, fin dal suo inizio (Magnini et al. 1994b), (Magnini et al. 1994a), nel 1994. Negli anni successivi Emanuele divenne il riferimento per una serie di attività di ricerca legate alle metodologie di sviluppo di wordnet multilingui allineati, includendo studi sui "lexical gap" e sulla possibilità di trasferire annotazioni semantiche da una lingua ad altre (e.g. MultiSemcor (Bentivogli and Pianta 2005)). La metodologia sperimentata con MultiWordNet è stata adottata per diverse lingue, inclusa una originale versione per il latino, e la risorsa è stata distribuita in diverse migliaia di licenze d'uso.

Traduzione speech to speech. Alla fine degli anni '90 Emanuele ha svolto un ruolo importante all'interno del progetto NESPOLE, portando le proprie competenze di linguistica computazionale in un contesto di collaborazioni internazionali sul tema della traduzione automatica speech-to-speech. Uno dei risultati di rilievo è stato un dataset multilingua (Mana et al. 2004), che raccoglie dialoghi parlati nei domini del turismo e della medicina, con le loro trascrizioni e annotazioni a livello di interlingua.

CELCT. Per il periodo da giugno 2009 a novembre 2012 Emanuele ha assunto la direzione scientifica di CELCT, il "Centro per la valutazione delle tecnologie del linguag-

gio e della comunicazione" di Trento, subentrando a Amedeo Cappelli, che ne era stato direttore dal 2003. Fondamentali sono stati i contributi di Emanuele per lo sviluppo di una serie di benchmark per la lingua italiana, tra cui I-CAB (Magnini et al. 2006), ancora oggi utilizzato come data set di addestramento per task di estrazione di informazione da testi, e la versione italiana di Time-ML (Caselli et al. 2011).

Evalita. Sotto la direzione di Emanuele, CELCT ha contribuito in particolare all'organizzazione di Evalita 2011, la campagna di valutazione delle tecnologie del linguaggio scritto e parlato, per la lingua italiana, di cui fu co-coordinatore scientifico (Magnini et al. 2012). Si devono in buona parte al contributo di CELCT i task su Named Entity Recognition on Transcribed Broadcast News e Cross-document Coreference Resolution of Named Person Entities in quella edizione di Evalita. In quanto direttore del Centro, Emanuele fu anche responsabile dei numerosi progetti che hanno coinvolto CELCT, e che hanno fatto di Emanuele una figura molto conosciuta e apprezzata anche a livello internazionale.

TextPro. Uno dei maggiori risultati tecnologici raggiunto da Emanuele è stata l'ideazione e la realizzazione della piattaforma TextPro (Pianta, Girardi, and Zanoli 2008) per l'annotazione di informazione su testi. TextPro è stato progettato come una cascata di annotatori indipendenti (tokenizzatore, post tagging, analizzatore morfo-sintattico, riconoscitore di entità nominate, ecc.) raggruppati in una unica piattaforma. Progettata inizialmente per l'italiano, TextPro è stato in seguito esteso all'inglese, e il piano iniziale arricchito con ulteriori moduli di annotazione. La gran parte dei progetti applicativi nel campo delle tecnologie del linguaggio portati avanti da FBK, per anni si è avvalsa di TextPro come strumento di estrazione di informazioni da testi scritti.

FrameNet. Dopo MultiWordNet, Emanuele si è dedicato alla creazione di FrameNet per l'italiano (Tonelli and Pianta 2008), una risorsa semantica per categorizzare situazioni e eventi in "frame", e i relativi partecipanti in "frame element", o ruoli semantici. Partendo da FrameNet per l'inglese, sviluppato alla fine degli anni '90 a Berkeley sulla base della "frame semantics" proposta dal linguista Charles Fillmore, Emanuele ha proposto di crearne la versione italiana, riutilizzando dove possibile tecniche di proiezione dell'annotazione già sperimentate in MultiWordNet. La risorsa annotata, da lui coordinata, è stata rilasciata alla comunità scientifica e rappresenta tutt'ora uno dei nuclei centrali di FrameNet per l'italiano (Basili et al. 2017), un progetto ancora in corso a cui collaborano diverse università.

Lessico bilingue della lingua veneta. Per un breve periodo Emanuele ha collaborato al progetto STILVEN sulla lingua veneta, finanziato dalla Regione Veneto, producendo un lessico bilingue con le forme di parola morfologiche di tutti i verbi - solo lemmi - inclusi nei dizionari già disponibili, incluse le forme cliticizzate.

Parole chiave. Gli interessi scientifici di Emanuele nascevano spesso da esigenze pratiche. Per esempio, l'idea di implementare un estrattore di concetti-chiave multilingua era stato pensato come un primo passo per arrivare alla generazione automatica di mappe concettuali, che gli studenti potessero utilizzare a scopi educativi. Anche se il tema delle mappe concettuali è rimasto purtroppo inesplorato, Emanuele ha ideato, implementato e rilasciato il sistema Keyword eXtractor (KX) (Pianta and Tonelli 2010), un estrattore di concetti-chiave configurabile a seconda del dominio, basato su criteri linguistici per il riconoscimento di espressioni polirematiche. Il tool ha dimostrato la propria efficacia in

ambiti diversi, dall'analisi di testi brevettuali (progetto Patexpert) a quella di documenti storici (progetto Alcide).

Semplificazione del testo. Un altro ambito di studio a cui Emanuele si è dedicato è stato quello della profilazione del testo finalizzata a comprendere quali aspetti di un documento potevano risultare di difficile comprensione, soprattutto per bambini con disabilità cognitive. Questo problema è stato affrontato da Emanuele con un approccio interdisciplinare che coniugava l'analisi e la generazione di linguaggio naturale con le scienze cognitive, il design di interfacce uomo-macchina e la gamification. Le tecnologie sviluppate da Emanuele nei progetti LODE e Terence sono state utilizzate con successo da bambini non udenti e da quelli con lievi disabilità cognitive, che hanno potuto giocare e fare esercizi a partire da storie semplificate con metodi automatici.

2. Il premio AILC "Emanuele Pianta" per la miglior tesi di laurea magistrale

Alla luce dei suoi numerosi contributi scientifici, il Consiglio Direttivo dell'Associazione Italiana di Linguistica Computazionale, nella seduta del 12 febbraio 2020, ha deciso all'unanimità di intitolare a Emanuele Pianta il premio annuale assegnato alla miglior tesi di laurea magistrale nell'ambito della Linguistica Computazionale, discussa in una università italiana.

Il premio AILC è stato istituito nel 2017 in corrispondenza della quarta edizione della conferenza CLiC-it, svoltasi a Roma dall'11 al 13 dicembre 2017, con l'obiettivo di promuovere e individuare eccellenze nel campo della ricerca della Linguistica Computazionale (vengono considerate le aree elencate nella call for papers della conferenza CLiC-it). Il premio viene assegnato da una giuria composta da tre membri: un membro del comitato organizzatore del convegno CLiC-it dell'anno precedente, un membro del comitato organizzatore del convegno CLiC-it dell'anno in corso (questo membro si impegna a servire nella giuria per due anni, così da garantire continuità) e un membro del Direttivo AILC. Il premio consiste in 500 euro, l'iscrizione gratuita a AILC per un anno, e l'iscrizione alla conferenza CLiC-it, dove l'autore ha la possibilità di presentare la tesi vincitrice del premio.

Giunto alla terza edizione, il premio si è affermato nella comunità di ricerca italiana come un riconoscimento importante a studenti brillanti nel settore della Linguistica Computazionale. In ordine temporale, il premio è stato assegnato a Alessio Miaschi (2017 - Università di Pisa, "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"), Enrica Troiano (2018 - Università di Trento/FBK, "A Computational Study of Linguistic Exaggerations") e Ludovica Pannitto (2019 - Università di Pisa, "Event Knowledge in Compositional Distributional Semantics").

Ci piace concludere questo breve ricordo della figura di Emanuele Pianta riassumento gli aspetti che, a nostro parere, hanno maggiormente caratterizzato il suo contributo nel campo della Linguisitica Computazionale. In primo luogo l'attitudine alla ricerca multidisciplinare, con lo scopo di combinare le conoscenze acquisite in ambiti diversi, nella convinzione che questa combinazione possa portare ad una migliore comprensione della complessità sottostante all'uso del linguaggio. Poi la visione sulle direzioni della ricerca, ad esempio intuendo l'importanza di puntare sulla piattaforma TextPro, oppure sullo sviluppo di FrameNet per l'italiano. Infine, l'impatto di Emanuele nel nostro campo è stato possibile anche grazie alla sua innata capacità di comunicare,

con la quale ha coinvolto tutti, giovani studenti e ricercatori ormai affermati, in appassionate discussioni sulla Linguisitica Computazionale.

Tutto questo ha motivato AILC nella scelta di intitolare ad Emanuelel Pianta il premio per la miglior tesi magistrale in Linguistca Computazionale, e rende Emanuele un esempio per le generazioni future.

References

Basili, Roberto, Silvia Brambilla, Danilo Croce, and Fabio Tamburini. 2017. Developing a large scale FrameNet for Italian: the IFrameNet experience. In *Proceedings of the Fourth Italian Conference on Computational Linguistics (CLiC-it 2017)*, Rome, 11-13 December.

Bentivogli, Luisa and Emanuele Pianta. 2005. Exploiting parallel texts in the creation of multilingual semantically annotated resources: The MultiSemCor Corpus. *Natural Language Engineering, Special Issue on Parallel Texts*, 11(3):247–261.

Caselli, Tommaso, Valentina Bartalesi Lenzi, Rachele Sprugnoli, Emanuele Pianta, and Irina Prodanof. 2011. Annotating events, temporal expressions and relations in Italian: the It-TimeML experience for the Ita-TimeBank. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 143–151, Portland, Oregon, USA, June 23-24.

Delmonte, Rodolfo, Dario Bianchi, and Emanuele Pianta. 1992. GETA_RUN - A general text analyzer with reference understanding. In *Proceedings of the 3rd Conference on Applied Natural Language Processing, Systems Demonstrations*, pages 9–10, Trento, Italy, March.

Magnini, Bernardo, Francesco Cutugno, Mauro Falcone, and Emanuele Pianta, editors. 2012. *Evaluation of Natural Language and Speech Tools for Italian, International Workshop, EVALITA 2011, Rome, January 24-25, 2012, Revised Selected Papers*. Springer.

Magnini, Bernardo, Emanuele Pianta, Christian Girardi, Matteo Negri, Lorenza Romano, Manuela Speranza, Valentina Bartalesi Lenzi, and Rachele Sprugnoli. 2006. I-CAB: the Italian Content Annotation Bank. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, May.

Magnini, Bernardo, Carlo Strapparava, Fabio Ciravegna, and Emanuele Pianta. 1994a. Multilingual lexical knowledge bases: Applied WordNet prospects. In *Proceedings of the International Workshop on the Future of the Dictionary*, Grenoble, October.

Magnini, Bernardo, Carlo Strapparava, Fabio Ciravegna, and Emanuele Pianta. 1994b. A project for the construction of an Italian lexical knowledge base in the framework of WordNet. Technical report, IRST # 9406-15, June.

Mana, Nadia, Roldano Cattoni, Emanuele Pianta, Franca Rossi, Fabio Pianesi, and Susanne Burger. 2004. The Italian NESPOLE! corpus: a multilingual database with interlingua annotation in tourism and medical domains. In *Proceedings of the fourth International Conference on Language Resources and Evaluation (LREC-2004)*, Lisbon, Portugal, May.

Pianta, Emanuele, Christian Girardi, and Roberto Zanoli. 2008. The TextPro tool suite. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.

Pianta, Emanuele and Sara Tonelli. 2010. KX: A flexible system for keyphrase extraction. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10) at ACL 2010*, pages 170—173, Uppsala, Sweden, July.

Tonelli, Sara and Emanuele Pianta. 2008. Frame information transfer from English to Italian. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May.