Laboratorio di Cultura Digitale

Università di Pisa

PIONIERI TRA DUE CULTURE

INFORMATICA UMANISTICA A PISA IN ONORE DI MARIA SIMI

a cura di
Enrica Salvatori, Susanna Pelagatti e Chiara Mannari

QUADERNI DI CULTURA DIGITALE 3

Simonelli

a cura di
ENRICA SALVATORI
SUSANNA PELAGATTI
CHIARA MANNARI

PIONIERI TRA DUE CULTURE

Informatica Umanistica a Pisa
in onore di Maria Simi

UNIVERSITÀ DI PISA

Laboratorio di Cultura Digitale

QUADERNI DI CULTURA DIGITALE 3

SIMONELLI

Pionieri fra due culture
Informatica umanistica a Pisa in onore di Maria Simi

a cura di
Enrica Salvatori, Susanna Pelagatti e Chiara Mannari

Realizzazione grafica copertina
Theo van Boxel

Realizzazione tipografica
Nicoletta Salvatori e Mario Valori

Collana
Quaderni di Cultura Digitale LabCD - Università di Pisa

Se Book
Simonelli Electronic Book www.simonel.com — ed@simonel.com

Direttore del LabCD
Susanna Pelagatti

ISBN: 9788893203166

Le monografie del LabCD

La nuova collana di ebook “Quaderni di cultura digitale” realizzata a cura del Laboratorio di Cultura Digitale dell’Università di Pisa (http://www.labcd.unipi.it) ed edita da Simonelli editore (ebooksitalia.com) ospita brevi monografie sugli strumenti e le ricerche nell’ambito dell’informatica umanistica, emerse dal lavoro di docenti e studenti che collaborano con il Laboratorio stesso. Si propone da un lato di sostenere una più larga diffusione della cultura digitale, intesa come il campo che vede interagire e collaborare in maniera complementare le discipline umanistiche e alcuni settori dell’informatica, e dall’altro di valorizzare le competenze che sono emerse e che continuano a fiorire entro e attorno al corso di laurea di Informatica Umanistica dell’ateneo pisano.

Le monografie sono dedicate a temi a forte carattere interdisciplinare e presentano testi adatti a essere adottati in corsi universitari come anche a fornire materiale di apprendimento a pubblici interessati.

Il Laboratorio di Cultura Digitale è un Centro interdipartimentale di formazione e ricerca dell’Università di Pisa, nato nel 2011 a seguito dello sviluppo delle ricerche e dell’attività didattica del corso di laurea di Informatica Umanistica (https://infouma.fileli.unipi.it/). Al suo interno collaborano docenti di cinque dipartimenti e numerosi studiosi indipendenti, riuniti dal desiderio di sviluppare progetti, strumenti e conoscenze che uniscano l’ambito umanistico e quello informatico.

Il Laboratorio funziona come una sorta di bottega digitale, un DigiCraft, nel senso che lavora a progetti digitali integrando le

diverse capacità e competenze di docenti, esperti indipendenti, dottorandi, laureandi e tirocinanti. Intende promuovere ricerche in grado di rispondere ai mutamenti che la continua innovazione tecnologica porta nel mondo della cultura, elaborando modelli e linguaggi capaci di interpretare un mondo in cambiamento.

Indice

1. Colophon
2. Le monografie del LabCD
3. Indice
4. Prefazione
5. Nota d’uso
6. Che cosa mi ha insegnato Informatica umanistica (Mirko Tavoni)
7. 2002: Odissea nell’Informatica Umanistica (Alessandro Lenci)
8. L’informatica umanistica e gli strumenti della conoscenza (Chiara Mannari)
9. Costruendo un Digicraft: storia, ambizioni e sfide del Laboratorio di Cultura Digitale (Enrica Salvatori)
10. Unità e molteplicità nella didattica dell’Informatica umanistica (Mirko Tavosanis)
11. Le Digital Humanities come raccordo tra discipline, contesti e approcci diversi (Simona Turbanti)
12. La collaborazione del CoPhiLab (CNR-ILC) con l’Università di Pisa nell’ambito DH (Mario Angelo Del Grosso, Federico Boschetti)
13. Filologia dei requisiti: dalle Humanities al Software Engineering (Vincenzo Gervasi)
14. Challenges and Perspectives of Data Science for Digital Humanities (Angelica Lo Duca)
15. Sull’insegnamento di Introduzione all’Intelligenza Artificiale (Alessio Micheli)
16. Modellare e simulare il rapporto tra persone e tecnologia (Giovanna Broccia, Lucia Nasti, Paolo Milazzo)
17. Nel cantiere aperto delle treebank per l’italiano: dialogo e integrazione di prospettive diverse (Simonetta Montremagni, Chiara Alzetta, Felice Dell’Orletta, Giulia Venturi)

18. Per una rivoluzione digitale nella cura delle persone con disabilità cognitiva: una proposta di linee guida per progetti digitali in ambito sanitario (Susanna Pelagatti)
19. Le pratiche di Pair programming: il caso di GitHub Copilot (Lukasz Szczygiel)
20. Il peso dell'umanista informatico nella formazione per gli iscritti dell'AIB e nella digitalizzazione dei procedimenti in un istituto comprensivo statale (Maria Accarino)
21. Il metaverso come strumento di didattica aumentata: il caso di Second Life (Marco Bani)
22. Fotografia: il lungo e travagliato percorso dall'analogico al digitale, tra Guerra Fredda, innovazioni tecnologiche e sviluppi commerciali (Marco Capovilla)
23. Ritorno al passato - su due riviste open access (Cristina Cassina)
24. Informatica Umanistica e Medioevo. Metodologie di georeferenziazione della Storia (Alessandro Cignoni, Laura Galoppini)
25. The LexEcon project six months after: perspectives and problems of a research on the lexicon of economics (Marco Guidi)
26. Valorizzazione e comunicazione negli archivi storici: qualche esempio e alcuni spunti di riflessione (Cristina Moro)
27. La geografia dei testi letterari (Paolo Rossi)
28. Artificial Intelligence in Education (Daniela Rotelli)
29. Autoritratto multimediale (Elvira Todaro)
30. Elenco degli autori

Nel cantiere aperto delle treebank per l'italiano: dialogo e integrazione di prospettive diverse

- Simonetta Montemagni (CNR-ILC)
- Chiara Alzetta (CNR-ILC)
- Felice Dell'Orletta (CNR-ILC)
- Giulia Venturi (CNR-ILC)

Questo contributo si propone di ripercorrere le principali tappe della collaborazione dell'Istituto di Linguistica Computazionale "Antonio Zampolli" del Consiglio Nazionale delle Ricerche con il Dipartimento di Informatica dell'Università di Pisa, in particolare con Maria Simi: la collaborazione ha avuto origine più di 15 anni fa, con l'obiettivo di dotare la lingua italiana di risorse per il trattamento automatico della lingua, ed è tuttora in corso. Con gli anni, il gruppo di lavoro si è progressivamente allargato, coinvolgendo fino dall'inizio il Dipartimento di Filologia, Letteratura e Linguistica della stessa università (Alessandro Lenci), poi l'Università di Torino (Cristina Bosco) e, più recentemente, l'Università di Bologna (Fabio Tamburini).

Le risorse linguistiche su cui questo articolo si focalizza sono costituite da corpora testuali arricchiti con annotazioni relative alla struttura linguistica sottostante, in particolare morfo-sintattica e sintattica, denominate "treebank". In numerose occasioni, ci siamo trovati ad assimilare la costruzione di "treebank" a quella delle cattedrali medievali. Molteplici sono le analogie su cui si fonda la metafora proposta. Nel Medioevo, la costruzione di una cattedrale è un lavoro collettivo che impegna l'intera comunità, è un evento fuori dalla norma per le sue dimensioni, per lo sforzo economico, per le tecnologie e i mezzi necessari, nonché per i tempi richiesti. Il cantiere della cattedrale è una realtà complessa in cui dialogano diverse figure professionali, ciascuna specializzata in una diversa mansione e spesso provenienti da luoghi

differenti, che diventa una piazza speciale per lo scambio di tecniche e conoscenze nonché luogo di sperimentazioni e innovazioni. Ricchi benefattori, come i nobili, il clero e i mercanti più agiati, partecipavano in maniera finanziariamente cospicua alla costruzione della cattedrale.
Analogamente, la costruzione di una treebank rappresenta intrinsecamente un'impresa collettiva, in cui competenze ed esperienze diverse si intrecciano, finalizzate alla realizzazione di un'iniziativa che, per i suoi sforzi obiettivi, non può che avere un carattere corale. La costruzione di corpora annotati richiede il dialogo e l'integrazione tra competenze diverse, tipicamente di area informatica e di area linguistica, e la collaborazione di gruppi di ricerca diversi. Il finanziamento richiesto per imprese come questa è ingente: in Italia, un primo nucleo della treebank ha origine nell'ambito del progetto nazionale SI-TAL ("Sistema Integrato per il Trattamento Automatico del Linguaggio"), finanziato dal Ministero dell'università e della ricerca (allora denominato MURST) e finalizzato alla creazione di risorse per l'italiano, e che ha riunito le attività di alcuni tra i principali protagonisti della ricerca nazionale nel settore (Zampolli, 2001). Le successive evoluzioni delle treebank per la lingua italiana sono state supportate da finanziamenti di importanti industrie interessate a usare la risorsa sviluppata all'interno dei propri prodotti, tra le quali Google e Linkedin.
Il contributo è organizzato come segue. Dopo aver illustrato in breve cosa è una treebank, a cosa serve e come viene costruita, vengono introdotte le treebank per l'italiano. Le due sezioni successive ripercorrono l'azione corale della comunità nazionale per la costruzione di risorse sempre più ampie e armonizzate, e verso risorse iterativamente migliorate ed estese.

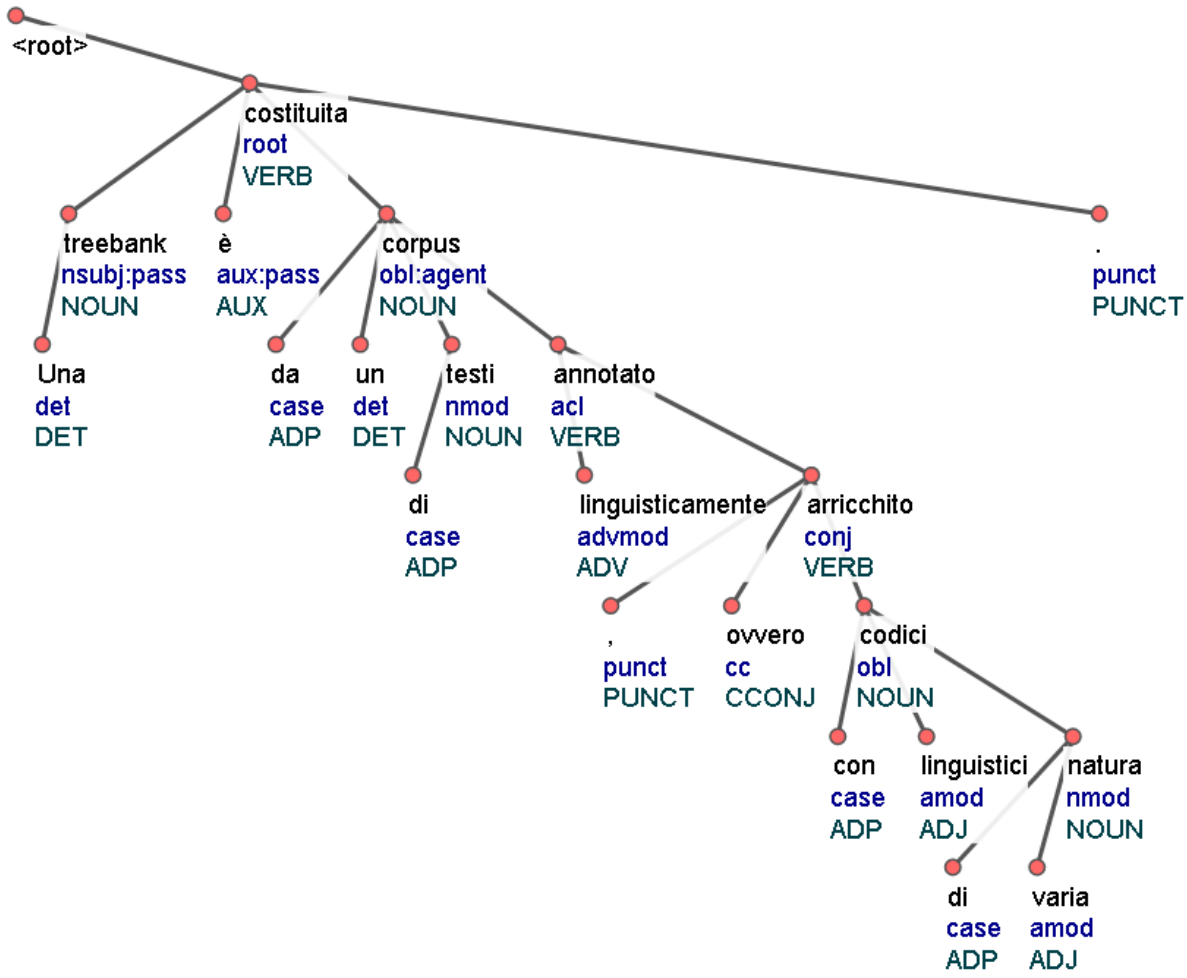
Treebank: perché, cosa e come

La costruzione di corpora testuali arricchiti con annotazione linguistica di varia natura ha una lunga storia. Alle origini, i corpora annotati sono costruiti dai linguisti, che cercano esempi (o controesempi) per una determinata teoria o ipotesi all'interno di studi basati su corpora, o dagli psicolinguisti,

interessati a calcolare la frequenza di fenomeni specifici e a confrontarla con i giudizi umani. All'epoca, per annotare i testi anche con i fenomeni linguistici più semplici era necessario un notevole dispendio di tempo e di energie e i corpora annotati disponibili per lo studio erano inevitabilmente di dimensioni ridotte. L'utilizzo di treebank da parte dei linguisti computazionali è più recente, risale agli anni '90. Negli ultimi tre decenni, l'evoluzione delle capacità di calcolo e memorizzazione dei computer, insieme allo sviluppo di metodi robusti per l'annotazione automatica, hanno reso i dati annotati linguisticamente disponibili in quantità sempre più crescenti. Queste risorse svolgono oggi un ruolo centrale nel settore del trattamento automatico della lingua (o NLP, "Natural Language Processing"): oggi il paradigma dominante si basa su corpora di testo e vocali annotati linguisticamente, usati sia per valutare nuove tecnologie della lingua sia per sviluppare modelli statistici affidabili per l'addestramento di queste tecnologie. Negli ultimi anni si è assistito a una notevole impennata dell'attività di annotazione linguistica, che si è estesa a un'ampia varietà di fenomeni linguistici e di lingue, anche tipologicamente molto lontane. Nonostante le evoluzioni tecnologiche abbiano facilitato la costruzione di treebank di sempre più vaste dimensioni, si tratta sempre di imprese notevolmente impegnative, in termini di tempo e risorse richieste.

Una treebank è costituita da un corpus di testi annotato linguisticamente, ovvero arricchito con codici linguistici di varia natura che rappresentano in modo esplicito la struttura sottostante al testo. Il termine "treebank" allude al fatto che il modo più comune di rappresentare l'analisi grammaticale è attraverso una struttura ad albero (tipicamente, a costituenti o a dipendenze). Nell'uso corrente, il termine non si limita più ai corpora che contengono rappresentazioni ad albero, ma si applica anche a corpora che contengono altri tipi di rappresentazioni (ad esempio, di tipo categoriale come le parti del discorso o categorie morfo-sintattiche, o classificazioni di natura semantica). Nella figura che segue è riportata la rappresentazione a livello morfo-sintattico e sintattico a dipendenze della frase *Una treebank è costituita da un corpus di testi annotato linguisticamente, ovvero arricchito con codici linguistici di varia natura* dove a ogni parola (o "token") del testo è

associata l'informazione relativa alla categoria morfo-sintattica nel contesto specifico (ad esempio, `NOUN`=nome, `ADJ`=aggettivo, `VERB`=verbo, `ADP`=preposizione, `DET`=determinante) e la relazione di dipendenza che lo lega alla relativa testa sintattica (ad esempio, `nsubj:pass`=soggetto nominale in costruzione passiva, `obl:agent`=complemento di agente, `amod`=modificatore aggettivale).



Tipicamente, il termine "treebank" designa corpora la cui annotazione è stata oggetto di revisione manuale. La costruzione di una treebank può essere condotta in diversi modi, ossia attraverso un processo di: annotazione completamente manuale; annotazione semi-automatica, ottenuta tramite la revisione manuale dell'output automatico di strumenti di annotazione; conversione (semi-)automatica da risorse preesistenti. Se l'annotazione

completamente manuale richiede tempo, è costosa ed è soggetta a incongruenze anche da parte di un singolo annotatore (Fort et al., 2012), l'annotazione semi-automatica è più veloce, meno incline alle incongruenze derivanti da decisioni arbitrarie del singolo annotatore, ma è soggetta ai cosiddetti effetti di "ancoraggio", in base ai quali le decisioni umane sono influenzate da valori preesistenti che possono anche includere errori dello strumento di annotazione automatica (Berzak et al., 2016). Recentemente, le risorse disponibili sono sempre più spesso il risultato di un processo di conversione che sfrutta corpora annotati preesistenti: a seconda che la conversione sia eseguita all'interno dello stesso paradigma di rappresentazione sintattica o meno, gli approcci possono riguardare la conversione da costituenti a dipendenze o viceversa, oppure operare all'interno della stessa famiglia di rappresentazioni. La conversione può essere combinata anche con il processo di integrazione e armonizzazione di risorse diverse (Bosco et al., 2012): Nivre e Megyesi (2007) si riferiscono a questo caso come "Cross-Corpus Harmonization and Annotation Projection".

Le treebank per l'italiano

In questo contributo ci focalizziamo sull'annotazione sintattica dell'italiano. Possiamo identificare due principali approcci alla rappresentazione della struttura sintattica:

- una rappresentazione a costituenti, basata sull'identificazione di costituenti sintattici (ad es. sintagmi nominali, sintagmi verbali, sintagmi preposizionali, ecc.) e delle loro relazioni di incassamento gerarchico
- una rappresentazione a dipendenze o funzionale, che fornisce una descrizione della frase in termini di relazioni binarie di dipendenza tra parole come soggetto, oggetto diretto, modificatore, etc.

All'interno di questi due approcci, lo spettro di variazione nella definizione degli schemi di annotazione sintattica è molto ampio. In questa sede, ci concentriamo sulla rappresentazione a dipendenze, che risulta particolarmente adeguata in rapporto a lingue come l'italiano caratterizzate da una forte variabilità nell'ordine dei costituenti e la possibilità di omettere il soggetto al livello della frase principale (Lenci et al., 2009).
Tre sono le principali treebank per la lingua italiana originariamente basate su una rappresentazione a dipendenze:

- Italian Semantic Syntactic Treebank (ISST), sviluppata nell'ambito del progetto nazionale SI-TAL (1999-2001) che ha riunito le attività di alcuni tra i principali protagonisti della ricerca nazionale nel settore del trattamento automatico della lingua, Montemagni et al. (2000, 2003)
- Turin University Treebank (TUT), sviluppata dal gruppo di NLP dell'Università di Torino (2000), Bosco et al. (2000, 2004)
- Venice Italian Treebank (VIT), sviluppata dall'Università Ca' Foscari di Venezia (2007), Delmonte et al. (2007)

Per quanto le tre treebank si collochino tutte all'interno della stessa famiglia basata sulla nozione di dipendenza sintattica, si osserva un ampio spettro di variazione, riguardante: i) la dimensione e la composizione del corpus annotato; ii) lo schema di annotazione, iii) i criteri di applicazione definiti per il riconoscimento delle categorie. Per quanto riguarda la dimensione, si va dai poco più di 300.000 tokens di ISST ai 94.000 di TUT. Mentre ISST include principalmente testi di tipo giornalistico, TUT include anche altri generi, come ad esempio testi giuridici. In VIT, invece, trovano spazio anche trascrizioni di parlato riconducibili a diversi registri. Relativamente allo schema di annotazione adottato, le principali differenze riguardano la granularità del repertorio di relazioni di dipendenza utilizzate per l'annotazione (si va da 323 per TUT a 10 per ISST), così come i criteri di annotazione riguardanti in particolare l'identificazione della testa e di specifiche relazioni (v. infra). Un ultimo aspetto riguarda il formato di rappresentazione utilizzato, che alle origini era specifico per ciascuna risorsa.

Verso risorse sempre più ampie e armonizzate

Standardizzazione del formato di rappresentazione

Con l'affermarsi di algoritmi di Machine Learning per affrontare diversi compiti di trattamento automatico della lingua, le treebank diventano un componente centrale degli strumenti di annotazione linguistica: a partire da una treebank viene costruito il modello statistico per l'annotazione del testo.
Tuttavia, i formati proprietari delle tre treebank le rendevano difficilmente utilizzabili per questo scopo. Nel 2007, sia TUT che ISST vengono convertite dal formato proprietario originariamente utilizzato al formato CoNLL IOB, uno standard *de facto* progettato per essere utilizzato all'interno delle campagne di valutazione di componenti di NLP organizzate dalla comunità internazionale della Conference on Natural Language Learning (CoNLL). Si tratta di un formato tabellare ideato per consentire annotazioni multi-livello relative agli stessi dati, facilmente elaborabile da gruppi di ricerca e software diversi.
La conversione dei formati di rappresentazione proprietari verso il formato TUT-CoNLL diventa una necessità per poter includere la lingua italiana all'interno di campagne di valutazione nazionali e internazionali. TUT viene convertita nel formato CoNLL e utilizzata all'interno della campagna nazionale di valutazione incentrata sulla lingua italiana, EVALITA 2007 (Bosco et al. 2007). La conversione di ISST nasce invece come impresa congiunta tra l'Università di Pisa e ILC, ed è finalizzata a includere la lingua italiana all'interno di competizioni internazionali. ISST viene usata come punto di partenza per lo sviluppo del corpus da utilizzare nell'ambito del CoNLL-2007 Shared Task (Nivre et al., 2007). Il risultato di questo processo di conversione, denominato ISST-CoNLL , è costituito da un sottoinsieme del corpus annotato a dipendenze (79.654 tokens), ottenuto in modo semiautomatico combinando l'informazione di due livelli di annotazione, morfo-sintattica e sintattico-funzionale. Il contributo Montemagni e Simi (2007) documenta le sfide e il risultato di questo passaggio.

Grazie al formato standard di rappresentazione adottato, le treebank TUT-CoNLL e ISST-CoNLL (l'ultima evoluta successivamente in ISST-TANL) sono state utilizzate all'interno di successive campagne di valutazione: TUT-CoNLL nel Parsing Task di EVALITA 2009 (Bosco et al., 2009) e 2011 (Bosco e Mazzei, 2011); ISST–TANL nel Parsing Task di EVALITA 2009 (Bosco et al., 2009), nel Domain Adaptation Task di EVALITA 2011 (Dell'Orletta et al., 2013), e nello shared task su Dependency Parsing of Legal Text nell'ambito del Workshop su Semantic Processing of Legal Text (SPLeT@LREC '12, Dell'Orletta et al., 2012). In questa fase, le treebank sono sempre usate separatamente. Unica parziale eccezione è rappresentata da EVALITA 2009: in questa occasione, al fine di valutare l'influenza dell'impatto di diversi schemi di annotazione sull'accuratezza di un parser, la stessa porzione di corpus annotata secondo le specifiche di ISST-TANL e TUT-CoNLL è stata usata come testbed. Bosco et al. (2010) riportano i risultati di questa analisi comparativa, volta ad analizzare e quantificare l'impatto dei due schemi sull'accuratezza di sistemi automatici di annotazione. Questi risultati suggeriscono la necessità di una riflessione approfondita su questa relazione, dimostrando che il miglioramento delle tecnologie di parsing dovrebbe procedere di pari passo con lo sviluppo di rappresentazioni sintattiche più adeguate. Nonostante il formato di rappresentazione condiviso, al termine di questa fase profonde differenze rimanevano a livello di:

- granularità e inventario dei tipi di relazioni di dipendenza: ISST-CoNLL (2007) ha 21 relazioni di dipendenza diverse, ISST-TALN (2009) ne ha 29, e TUT-CoNLL 72
- criteri di selezione della testa della relazione: di natura sintattica vs semantica
- strategia di annotazione di specifiche strutture e fenomeni sintattici (ad esempio, strutture coordinate, trattamento della punteggiatura, identificazione della radice)

Nonostante le importanti evoluzioni appena delineate, le risorse sviluppate per la lingua italiana non erano competitive rispetto a quelle sviluppate per altre lingue: è infatti ampiamente risaputo che le dimensioni di una treebank

influiscono significativamente sull'accuratezza del modello addestrato su di essa. Mentre treebank coeve per la lingua inglese superavano ampiamente il milione di parole, per l'italiano le dimensioni rimanevano al di sotto dei 100.000 tokens per entrambe le treebank, TUT-CoNLL e ISST-CoNLL /TANL. All'interno di questo scenario, emerge dunque una nuova sfida, quella di dotare la lingua italiana di una treebank di maggiori dimensioni, che potesse creare i presupposti per migliorare l'accuratezza di strumenti di annotazione linguistica della lingua italiana. Per raggiungere questo obiettivo, piuttosto che costruire nuove risorse abbiamo preferito partire da risorse preesistenti, già ampiamente testate in campagne di valutazione nazionali e internazionali: ciò ha comportato l'unificazione e l'armonizzazione degli schemi di annotazione a dipendenze di TUT-CoNLL e ISST-CoNLL /TANL. Rispetto alla fase precedente si è passati da una conversione del formato di rappresentazione a un obiettivo ben più ambizioso, ovvero l'integrazione e l'unificazione dell'inventario di relazioni sintattiche e delle strategie di annotazione.

Armonizzazione e integrazione delle annotazioni: la prospettiva intralinguistica

L'impresa di armonizzazione degli schemi di annotazione di TUT-CoNLL e ISST-CoNLL/TANL e di integrazione dei corpora annotati all'interno di un'unica risorsa ha richiesto l'estensione del gruppo di lavoro al gruppo di ricerca dell'Università di Torino, che aveva curato la costruzione di TUT. In questa fase di armonizzazione intralinguistica degli schemi di annotazione, la sfida ha riguardato la definizione e sperimentazione di una metodologia per combinare diversi schemi di annotazione. Bosco et al. (2012) illustra questo processo, che va dall'analisi delle dimensioni di variabilità degli schemi di annotazione (cfr. sopra) alla proposta di uno schema di annotazione che funga da "interlingua" in vista della conversione delle treebank sorgenti e la loro combinazione all'interno di un'unica risorsa di maggiori dimensioni. Nasce così, MIDT, ovvero la "Merged Italian Dependency Treebank". Lo

schema di annotazione di MIDT è il risultato del bilanciamento di diversi obiettivi, non sempre convergenti:

1. la fondatezza della rappresentazione adottata dal punto di vista della teoria linguistica;
2. la replicabilità dell'annotazione in modo accurato da parte di parser a dipendenze;
3. l'utilizzo dello schema in vista di compiti di comprensione linguistica automatica, come Information Extraction;
4. infine, ma non certo per importanza, la possibilità di recuperare in modo affidabile e automatico le informazioni necessarie dalle risorse originarie.

Simi et al. (2015) confronta i risultati di esperimenti condotti con il parser stocastico DeSR (Dependency Shift Reduce; Attardi, 2006) usando il corpus MIDT rispetto alle risorse originarie in formato CoNLL (TUT e ISST–TANL): in entrambi i casi, viene riportato un incremento di accuratezza che mostra l'adeguatezza di MIDT per l'addestramento di parser a dipendenze.
La costruzione della Merged Italian Dependency Treebank rappresenta senza dubbio un'importante evoluzione della risorsa, guidata dalla necessità di dotare l'italiano di una treebank di maggiori dimensioni. Ciò ha comportato la definizione di una strategia di armonizzazione delle annotazioni e di uno schema di annotazione "ponte" così come l'integrazione dei risultati all'interno di un'unica risorsa. Tuttavia, la risultante treebank rimaneva isolata nel panorama delle treebank a dipendenze per le diverse lingue. Per quanto potesse essere utilizzata per l'addestramento di sistemi di annotazione per la lingua italiana, risultava difficile confrontare i risultati ottenuti per lingue diverse.

Armonizzazione e integrazione delle annotazioni: la prospettiva interlinguistica

L'ampia variabilità degli schemi di annotazione usati in treebank di lingue diverse rendeva difficile stabilire con esattezza se le differenze di accuratezza

dell'annotazione automatica fossero dovute a reali differenze strutturali tra le lingue oppure a differenze a livello dello schema di annotazione. Il desiderio di poter confrontare i risultati del parsing a dipendenze tra diverse lingue in modo significativo è stata una delle principali motivazioni alla base dell'iniziativa denominata "Universal Dependencies" (UD; Nivre, 2015; de Marneffe et al., 2021), volta a sviluppare treebank per un numero di lingue più ampio possibile con un'annotazione coerente dal punto di vista linguistico (morfo-sintattico e a dipendenze). L'obiettivo principale di UD era di facilitare lo sviluppo di parser a dipendenze multilingui e modalità di apprendimento interlinguistico (denominato "cross-lingual learning"), così come di condurre analisi comparative per le singole lingue e di studiare l'interazione tra algoritmi di parsing e tipologia linguistica. Iniziato come un progetto su scala ridotta nel 2014, UD è cresciuto fino a raccogliere intorno a sé una grande comunità che coinvolge centinaia di ricercatori in tutto il mondo e che, ad oggi, ha rilasciato più 243 treebanks per 138 lingue (v 2.11 di Novembre 2022). Il progetto UD si basa e sussume diverse iniziative precedenti nate con simili obiettivi (Google Universal Part-of-Speech Tags, HamleDT, Universal Dependency Treebanks e Universal Stanford Dependencies, come descritto in Nivre, 2015). La filosofia generale che guida l'iniziativa è quella di fornire un inventario universale di categorie linguistiche e linee guida per facilitare l'annotazione coerente di costruzioni simili in tutte le lingue, consentendo al contempo estensioni specifiche per le singole lingue.

Per integrare l'italiano tra le treebank di UD era necessario un passo ulteriore, riguardante la conversione dello schema di MIDT in quello definito per il progetto UD, in una prima fase denominato "Stanford Dependency Scheme" (SD, de Marneffe e Manning, 2008). La prima versione convertita è rappresentata da ISDT, ovvero la "Italian Stanford Dependency Treebank". Bosco et al. (2013) documenta le sfide e la strategia adottata per questo passaggio, che ha richiesto un'analisi comparativa degli schemi di annotazione MIDT e SD, la definizione di una complessa metodologia di conversione, nonché la localizzazione dello schema di annotazione SD per far fronte alle peculiarità della lingua italiana (a livello sia delle categorie morfo-

sintattiche sia delle dipendenze). ISDT (successivamente denominata IUDT, ovvero "Italian Universal Dependency Treebank") è parte dell'insieme delle treebank di UD fino dalla prima release, UD v1.0, rilasciata a maggio 2015. In questa fase, la metodologia per convertire e armonizzare diverse treebank già utilizzata nei passaggi precedenti è stata ulteriormente raffinata. Anche in questo caso la conversione è stata condotta in modo semi-automatico, grazie a un'ampia batteria di regole volte a trattare casi semplici come il rinominare una relazione di dipendenza, per arrivare a casi più complessi che richiedono una ristrutturazione dell'albero sintattico, condotta attraverso la ridefinizione della testa e/o la riassegnazione del tipo di dipendenza (Bosco et al. 2013; Simi et al. 2014).

ISDT è stata usata nel Dependency Parsing Task di Evalita 2014 (Bosco et al., 2014). Era la prima volta che veniva usata una treebank significativamente più grossa (di 252,965 tokens) rispetto alle precedenti campagne, e conforme agli standard *de facto* correnti a livello del formato di rappresentazione (CoNLL) e dello schema di annotazione adottato (SD). Ciò ha permesso anche l'organizzazione di un "Cross-Language Dependency Parsing" task (CLAP) con l'italiano come lingua sorgente e con diverse lingue target, anche non tipologicamente correlate.

Le risorse UD hanno avuto un impatto significativo sulla ricerca in NLP, in particolare per il parsing multilingue a dipendenze, come testimoniato nelle due edizioni di campagne di valutazione CoNLL condotte sul vasto insieme delle treebank UD nel 2017 e 2018 (Zeman et al., 2017, 2018). Rispetto agli sforzi pionieristici delle campagne di valutazione CoNLL del 2006 e del 2007, il numero di treebank e di lingue è notevolmente aumentato: la campagna del 2017 riguardava 49 lingue (tra cui 4 a sorpresa), a cui si sono aggiunte 8 nuove lingue in quella del 2018. La seconda e principale differenza riguarda il fatto che, grazie alle annotazioni standardizzate garantite da UD, è ora possibile condurre analisi comparative basate su confronti interlinguistici. Questo aspetto ha reso le treebank di UD fonti affidabili per studi più orientati alla linguistica, come quelli incentrati sulla tipologia dell'ordine delle parole (si veda, ad esempio, Futrell, Mahowald e Gibson 2015; Gulordava e Merlo, 2015; Naranjo e Becker 2018). Per la lingua italiana, si rinvia a Tusa et

al. (2016) e Alzetta et al. (2019, 2020).
Recentemente, alla rappresentazione di base di UD si è aggiunto un livello di rappresentazione più profondo, inizialmente denominato “Enhanced UD” (Schuster e Manning, 2016), oggi “Deep Universal Dependencies” (Droganova e Zeman, 2019). Questo tipo di rappresentazione si basa su una struttura a grafo più ricca che rende esplicite relazioni che all’interno della rappresentazione UD di base rimangono implicite: ad esempio, tra i fenomeni trattati a questo livello si segnalano i soggetti controllati di infinitive, le relazioni predicato-argomento in costruzioni ellittiche, oppure la propagazione delle relazioni riguardanti i congiunti all’interno di strutture coordinate, tra le altre cose. L’italiano è stato tra le prime lingue per le quali è stato rilasciato questo livello di annotazione, denominato e-IUDT (Simi e Montemagni, 2018).

Verso risorse più grandi e di qualità

Da quanto delineato finora, appare chiaramente che un’impresa come la costruzione di treebank è in costante evoluzione. Siamo passati da treebank in formati e schemi proprietari a versioni standardizzate a livello del formato di rappresentazione e dello schema di annotazione: questo passaggio ha permesso alla treebank italiana di entrare a fare parte dell’iniziativa denominata Universal Dependencies, che è in continuo sviluppo e mira a migliorare l’adeguatezza dello schema per il trattamento di un numero crescente di lingue. I cambiamenti maggiori hanno riguardato il passaggio dalle versioni 1.* alle 2.*; tuttavia, anche tra le singole versioni vi sono stati cambiamenti nell’annotazione di singole relazioni e/o costruzioni. Oggi, per quanto i principi di base dello schema di annotazione UD siano da considerarsi stabili, lo schema e la teoria sottostante sono ancora destinati a evolvere nel tempo. Ne è una viva testimonianza il dialogo continuo a cui assistiamo tra esperti di lingue diverse che cercano di trovare il giusto equilibrio tra le prospettive specifiche della lingua e quelle universali riguardanti l’applicazione dello schema UD alla loro lingua. Infine, ma non

certo per importanza, la vitalità dell'iniziativa è evidente se consideriamo il fatto che alle treebank italiane storiche si aggiungono costantemente nuove risorse specializzate in relazione a domini e/o varietà d'uso della lingua diverse, e rilasciate da gruppi diversi.

Nel quadro complesso e dinamico appena tratteggiato, appare necessario verificare la coerenza delle annotazioni, innanzitutto quelle provenienti da diverse risorse ma anche quelle interne alla stessa treebank. Il processo di miglioramento della qualità di una treebank è un processo potenzialmente iterativo, volto da un lato ad accelerare il processo di annotazione, e dall'altro a rimuovere le annotazioni errate o semplicemente incoerenti rispetto alle specifiche (Dickinson e Tufis, 2017). Come sottolineato da de Marneffe et al. (2017), le treebank UD rappresentano senza dubbio un buon banco di prova per le tecniche di rilevamento degli errori. A ciò contribuiscono essenzialmente due fattori: lo schema di annotazione UD e le relative linee guida sono in continua evoluzione (vedi sopra); la maggior parte delle treebank UD risultano da un processo di conversione automatica, che potrebbe aver contribuito alla generazione di errori di annotazione.

Tra le treebank UD per la lingua italiana, IUDT – risultante dall'armonizzazione e dalla fusione di risorse più piccole con schemi di annotazione originariamente incompatibili – ha rappresentato un interessante banco di prova per individuare relazioni di dipendenza errate o semplicemente incoerenti, riconducibili a diversi stili di annotazione per la stessa costruzione, oppure a criteri di annotazione ancora troppo vaghi che lasciavano spazio a scelte arbitrarie. A tal fine è stato utilizzato un algoritmo originariamente sviluppato per misurare l'affidabilità delle relazioni di dipendenza prodotte automaticamente, LISCA (Dell'Orletta et al., 2013). Alzetta et al. (2018a) mostra che LISCA permette di restringere significativamente lo spazio di ricerca degli errori in una treebank e, cosa più importante, di identificare in modo affidabile errori sistematici che rappresentano evidenza pericolosa e fuorviante per sistemi di annotazione automatica. I risultati ottenuti dimostrano l'efficacia del metodo, che al fine di evitare possibili interferenze legate al genere testuale è stato testato sulla sezione di IUDT contenente articoli di giornale. L'impatto della strategia

incrementale per migliorare la qualità e la coerenza interna delle annotazioni di una treebank proposto da Alzetta et al. (2018a) è stato oggetto di uno studio volto a quantificare l'influenza di questa strategia di correzione degli errori sull'accuratezza di diversi parser a dipendenze (Alzetta et al., 2018b). I risultati ottenuti sono promettenti: tutti i parser mostrano un aumento delle prestazioni in termini delle misure standard LAS, UAS e di SLAS, con quest'ultima che rappresenta una misura più focalizzata sui pattern di errori rilevati.
Lo stesso metodo può essere usato per l'estensione delle risorse di una lingua ad altre varietà d'uso della lingua o anche generi testuali. Questo è l'approccio che è stato seguito per la costruzione della treebank UD ParlaMint-it, recentemente rilasciata tra le treebank dell'italiano e contenente trascrizioni di sedute parlamentari. In questo caso l'algoritmo ha permesso di identificare all'interno dell'output automatico di un parser a dipendenze le annotazioni meno affidabili da sottoporre a revisione manuale, minimizzando così lo sforzo di correzione nella costruzione della treebank.

Conclusioni

Oggi, la varietà e la tipologia delle treebank UD per la lingua italiana è ampia. Il ventaglio delle treebank disponibili è stato esteso a diversi livelli: con l'inclusione, previa conversione e armonizzazione, della terza treebank a dipendenze per la lingua italiana (VIT), così come con risorse specializzate rispetto a particolari varietà d'uso della lingua o costruzioni specifiche. La comunità nazionale delle Universal Dependencies è attualmente vasta e diffusa, ed è impegnata non solo sul fronte della lingua italiana ma anche sul fronte delle lingue classiche.
In questo contributo, abbiamo mostrato come siamo passati da piccoli e sparsi progetti locali a un'iniziativa corale che è cresciuta nel tempo. Tornando alla metafora dalla quale siamo partiti, speriamo di essere riusciti nell'intento di mostrare come la costruzione di un ecosistema di treebank per la lingua italiana può essere assimilata a quella di una cattedrale medievale,

che ha progressivamente coinvolto gruppi di ricerca diffusi sul territorio italiano. Nella cattedrale italiana delle treebank di UD, il dialogo e l'integrazione di prospettive diverse, sia sul versante disciplinare ma anche su quello legato all'utilizzo delle risorse (applicativo vs. teorico), è stato molto stimolante e proficuo.
Maria Simi, all'interno di questa comunità, ha svolto un ruolo fondamentale. Ha stimolato, insieme a Giuseppe Attardi, la conversione di ISST verso formati di rappresentazione standard, così come ha promosso le risorse italiane che sono derivate dagli sforzi descritti in questo contributo all'interno della comunità industriale, nazionale e internazionale, ottenendo importanti finanziamenti che sono stati reinvestiti per il raffinamento delle treebank esistenti e per la costruzione di nuove. Sul versante scientifico, Maria Simi ha costituito un importante e costante punto di riferimento per la comunità nazionale, mentre a livello internazionale ha rappresentato un interlocutore privilegiato con i vertici di UD. Grazie Maria!

Bibliografia

Alzetta C., Dell'Orletta F., Montemagni S., Venturi G. (2018a) "Dangerous Relations in Dependency Treebanks", in *Proceedings of the 16th International Workshop on Treebanks and Linguistic Theories*, Praga, 23-24 gennaio 2018, pagg. 201-210.

Alzetta C., Dell'Orletta F., Montemagni S., Simi M., Venturi G. (2018b) "Assessing the Impact of Iterative Error Detection and Correction. A Case Study on the Italian Universal Dependency Treebank", in *Proceedings of the Universal Dependencies Workshop 2018 (UDW 2018)*, Brussels, 01/11/2018, pagg. 1-7.

Alzetta C., Dell'Orletta F., Montemagni S., Venturi G. (2019) "Inferring quantitative typological trends from multilingual treebanks. A case study". *Lingue e Linguaggio*, XVIII(2), pagg. 209-242.

Alzetta C., Dell'Orletta F., Montemagni S., Osenova P., Simov K., Venturi G. (2020) "Quantitative Linguistic Investigations across Universal Dependencies Treebanks", in *Proceedings of 7th Italian Conference on Computational Linguistics (CLiC-it)*, 1-3 March, 2021, Bologna, Italy.

Attardi, G. (2006) "Experiments with a multilanguage non-projective dependency parser", in *Proceedings of the CoNLL-X'06*, New York City, New York, pagg. 166–170.

Berzak Y., Huang Y., Barbu A., Korhonen A., Katz B. (2016) "Anchoring and Agreement in Syntactic Annotations", in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Austin, Texas, pagg. 2215–2224.

Bosco C., Lombardo V., Lesmo L., Vassallo D. (2000) "Building a treebank for Italian: a data-driven annotation schema", in *Proceedings of the 2nd Language Resources and Evaluation Conference (LREC'00)*, Athens, Greece, ELRA, pagg. 99–105.

Bosco C., Lombardo V. (2004) "Dependency and relational structure in treebank annotation", in *Proceedings of the Workshop on Recent Advances in Dependency Grammar at COLING 2004*, Ginevra, Svizzera.

Bosco C., Mazzei A., Lombardo V. (2007) "Evalita Parsing Task: An Analysis of the first Parsing System Contest for Italian", in *Proceedings of EVALITA* 2007.

Bosco C., Montemagni S., Mazzei A., Lombardo V., Dell'Orletta F., Lenci A. (2009) "Evalita'09 Parsing Task: comparing dependency parsers and treebanks", in *Proceedings of EVALITA* 2009.

Bosco C., Mazzei A. (2011) "The Evalita 2011 Parsing Task: the Dependency Track", in *Proceedings of EVALITA* 2011.

Bosco C., Montemagni S., Mazzei A., Lombardo V., Dell'Orletta F., Lenci A., Lesmo L., Attardi G., Simi M., Lavelli A., Hall J., Nilsson J., Nivre J. (2010)

“Comparing the Influence of Different Treebank Annotations on Dependency Parsing“, in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, 17-23 May 2010, ELRA, pagg. 1794 – 1801.

Bosco C., Montemagni S., Simi M. (2012) “Harmonization and Merging of two Italian Dependency Treebanks”, in *Proceedings of the LREC 2012 Workshop on "Language Resource Merging"*, PARIGI:ELRA, Istanbul, 22 May 2012, pagg. 23-30.

Bosco C., Montemagni S., Simi M. (2013) “Converting Italian Treebanks: Towards an Italian Stanford Dependency Treebank“, in Proceedings of the 7th Linguistic Annotation Workshop " *Interoperability with Discourse (LAW VII " ID at ACL-2013)*, Sofia, Bulgaria, August 8-9, pagg. 61-69.

Bosco C., Dell’Orletta F., Montemagni S., Sanguinetti M., Simi M. (2014) “The Evalita 2014 Dependency Parsing task“. In *Proceedings of 4th Evaluation of NLP and Speech Tools for Italian (EVALITA 2014)*, 11 December, Pisa, Italy, Pisa University Press.

Dell’Orletta F., Marchi S., Montemagni S., Plank B., Venturi G. (2012) “The SPLeT-2012 Shared Task on Dependency Parsing of Legal Texts”, in *Proceedings of the 4th Workshop on Semantic Processing of Legal Texts (SPLeT 2012)*, Istanbul, Turkey, 27 Maggio, pagg. 42-51.

Dell’Orletta F., Marchi S., Montemagni S., Venturi G., Agnoloni T., Francesconi E. (2013) “Domain Adaptation for Dependency Parsing at Evalita 2011”, in Magnini B., Cutugno F., Falcone M., Pianta E. (a cura di), *Evaluation of Natural Language and Speech Tool for Italian, LNCS–LNAI*, Vol. 7689, Springer–Verlag Berlin Heidelberg, pagg. 58–69.

Dell’Orletta F., Venturi G., Montemagni S. (2013) “Linguistically-driven Selection of Correct Arcs for Dependency Parsing”, *Computación y Sistemas*, 2, pagg. 125–136.

Delmonte R., Bristot A., Tonelli S. (2007) “VIT - Venice Italian Treebank: Syntactic and Quantitative features”, in *Proceedings of the Sixth International Workshop on Treebanks and Linguistic Theories*, volume 1, pagg. 43–54.

de Marneffe M., Manning C. (2008) “The Stanford typed dependencies representation”, in *Proceedings of the COLING Workshop on Cross-Framework and Cross-Domain Parser Evaluation*, Stroudsburg, PA, USA, Association for Computational Linguistics, pagg. 1–8.

de Marneffe M.C., Grioni M., Kanerva J., Ginter F. (2017) “Assessing the Annotation Consistency of the Universal Dependencies Corpora”, in *Proceedings of the 4th International Conference on Dependency Linguistics (Depling 2007)*, Pisa, Italy, pagg. 108–115.

de Marneffe M., Manning C., Nivre J., Zeman D. (2021) “Universal Dependencies”, *Computational Linguistics*, vol. 47, no. 2, pagg. 255-308.

Dickinson M., Tufis D. (2017) “Iterative enhancement”, in *Handbook of Linguistic Annotation*, Springer, Berlin, Germany, pagg. 257–276.

Droganova K., Zeman D. (2019) “Towards Deep Universal Dependencies”, in *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, Syntaxfest 2019)*, pagg. 144-152.

Fort K., Nazarenko A., Rosset S. (2012) “Modeling the Complexity of Manual Annotation Tasks: a Grid of Analysis”, in *Proceedings of COLING 2012*, pagg. 895–910.

Futrell R., Mahowald K., Gibson E. (2015) “Quantifying Word Order Freedom in Dependency Corpora”, in J. Nivre J. e E. Hajičová (a cura di), *Proceedings of the Third International Conference on Dependency Linguistics* (Depling 2015), pagg. 91–100.

Gulordava K., Merlo P. (2015) “Diachronic Trends in Word Order Freedom and Dependency Length in Dependency-Annotated Corpora of Latin and Ancient

Greek", in J. Nivre e E. Hajičová (a cura di), *Proceedings of the Third International Conference on Dependency Linguistics (DepLing 2015)*, pagg. 121–130.

Lenci A., Montemagni S., Pirrelli V. (2009) "Annotazione sintattica di corpora: aspetti metodologici", in Cecilia Andorno, Stefano Rastelli (a cura di), *Corpora di italiano L2: tecnologie, metodi, spunti teorici*, Perugia, Guerra Edizioni, pp. 25-46.

Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M.T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R. (2000) "The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation", in *Proceedings of the COLING Workshop on "Linguistically Interpreted Corpora (LINC-2000)"*, Luxembourg, 6 August 2000, pagg. 18-27.

Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Lenci A., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M.T., Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R. (2003) "Building the Italian Syntactic-Semantic Treebank", in Anne Abeillé (a cura di), *Building and using Parsed Corpora, Language and Speech series*, Kluwer, Dordrecht, pagg. 189-210.

Montemagni S., Simi M. (2007) *The Italian dependency annotated corpus developed for the CoNLL-2007 shared task. Technical report*, ILC-CNR.

Nivre J., Hall J., Kübler S., McDonald R., Nilsson J., Riedel S., Yuret D. (2007) "The CoNLL 2007 shared task on dependency parsing", in *Proceedings of the EMNLP-CoNLL*, pagg. 915–932.

Naranjo M.G. , Becker L. (2018) "Quantitative word order typology with UD", in D. Hautg, S. Oepen, L. Øvrelid, M. Candito, J. Hajic (a cura di), *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*,

December 13–14, 2018, Oslo University, Norway, Linköping University Electronic Press.

Nivre J., Megyesi B. (2007) “Bootstrapping a Swedish Treebank Using Cross-Corpus Harmonization and Annotation Projection”, in *Proceedings of the 6th International Workshop on Treebanks and Linguistic Theories (TLT)*, pagg. 97–102.

Nivre J. (2015) “Towards a Universal Grammar for Natural Language Processing”, in *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*.

Schuster S., Manning C.D. (2016) “Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks”, in *Proceedings of LREC-2016*.

Simi M., Bosco C., Montemagni S. (2014) “Less is More? Towards a Reduced Inventory of Categories for Training a Parser for the Italian Stanford Dependencies“, in *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, 26-31 May, Reykjavik, Iceland.

Simi M., Montemagni S., Bosco C. (2015) “Harmonizing and merging Italian treebanks: Towards a merged Italian dependency treebank and beyond”, in Basili R., Bosco C., Delmonte R., Moschitti A. e Simi M. (a cura di), *Harmonization and Development of Resources and Tools for Italian Natural Language Processing within the PARLI Project*, Springer International Publishing, Svizzera, vol. 589, pagg. 3-23.

Simi M., Montemagni S. (2018) “Bootstrapping Enhanced Universal Dependencies for Italian”, in *Proceedings of CLiC-it 2018*.

Tusa E., Dell’Orletta F., Montemagni S., Venturi G. (2016) “Dieci sfumature di marcatezza sintattica: verso una nozione computazionale di complessità”, in *Proceedings of the Third Italian Conference on Computational Linguistics (CLiC-it)*, Napoli, pagg. 3–16.

Zampolli A. (2001), La treebank sintattico-semantica dell'italiano nel contesto di due programmi di interesse nazionale nel settore del Trattamento Automatico della lingua parlata e scritta, in S. Montemagni, M. T. Pazienza (a cura di), *Atti del Workshop su "La Treebank sintattico-semantica dell'italiano di SI-TAL", 7° Congresso della Associazione Italiana per l'Intelligenza Artificiale (AI*AI 2001)*, Bari, 26 settembre 2001, pagg. 1-14.

Zeman D. et al. (2017) “CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies”, in *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pagg. 1-19.

Zeman D. et al. (2018) “CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies”, in *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pagg. 1-21.