Nove Anni di jTEI: *What's New?*

Federico Boschetti[1,2] **Gabriella Pardelli**[1] **Giulia Venturi**[1]
1 Istituto di Linguistica Computazionale "A. Zampolli", CNR / Pisa
2 Digital and Public Humanities Center – Università Ca' Foscari / Venezia
{nome.cognome}@ilc.cnr.it

Abstract

English. This paper illustrates methods and tools to study the development of research topics in the TEI community across the years. For this purpose, automatic terminology extraction technologies were exploited.

Italiano. *Questo contributo illustra metodi e strumenti per studiare il cambiamento diacronico degli interessi di ricerca della comunità TEI grazie all'uso di metodi di estrazione automatica della terminologia da corpora di dominio.*[1]

1 Introduzione

Questo contributo nasce dall'intento di studiare con metodi di *distant reading* jTEI: il Journal of the Text Encoding Initiative (https://journals.openedition.org/jtei), perché è una rivista che rappresenta un ponte interessante fra la comunità delle Digital Humanities e la comunità della Linguistica Computazionale.

Come indicato da Schreibman (2011), jTEI nasce nel 2011 dopo tre anni di gestazione con l'intento di pubblicare *selected papers* dei convegni annuali (i volumi 1-2, 4, 6, 8-10) e numeri monotematici su argomenti di rilevanza per la comunità TEI (il volume 3 dedicato alla linguistica e il volume 5 dedicato alle infrastrutture). Schreibman (2014) dichiara inoltre che il volume 7, il primo frutto di una *open call*, tocca "contemporary meta concerns within the community".

Un tassello del settore delle Digital Humanities viene rilevato in questo studio attraverso l'analisi diacronica di termini estratti dagli articoli pubblicati in jTEI dal 2011 al 2019. Lo scopo è quello di andare a identificare termini mono- e polirematici tipici del dominio, spia dell'orientamento tematico delle attività di ricerca della comunità TEI. Oggi lo studio delle comunità sta diventando infatti centrale per comprendere e interpretare per i vari domini la direzione scientifica nonché il genere, gli stakeholder e le possibili connessioni tra comunità. Solo per fare un esempio, dalla lettura degli indici dell'estrazione del jTEI Corpus, la comunità scientifica che ruota intorno a TEI sembra non voglia usare il sostantivo *computer* e l'aggettivo *computational*, preferendo usare invece l'aggettivo *digital* combinato con una miriade di sostantivi (come ad es. *editions*, *humanities*, *text*, *resources*, *age*, *archive*, *objects*, *facsimile*, *library*, *tools*) in linea con gli usi della più ampia comunità delle Digital Humanities, ma non della Linguistica Computazionale.

2 Background

Questo contributo prosegue sulla linea degli studi dedicati a riviste e comunità con interessi interdisciplinari di informatica e discipline linguistiche, storico-filologiche o letterarie. In particolare, per lo studio dell'evoluzione terminologica nelle Scienze Umane e Sociali si veda Tuzzi (2018); per lo studio delle comunità della Linguistica Computazionale e delle Digital Humanities si veda Sprugnoli et al. (2019) e Pardelli et al. (2019); per lo studio della comunità della Tecnologia della Lingua e delle Risorse Linguistiche si vedano Mariani et al. (2014), Francopoulo et al. (2016), Soria et al. (2014), Bartolini et al. (2018) e Del gratta et al. (2018); per lo studio della comunità internazionale di Grey Literature si veda Pardelli et al. (2017).

Le soluzioni sin ad oggi messe a punto nell'ambito dell'estrazione automatica di terminologia da corpora di dominio sono molteplici e di diversa natura. Sebbene differiscano rispetto alle metriche utilizzate, alcuni obiettivi sono condivisi e riguar-

dano principalmente due aspetti legati alla difficoltà di definire strategie per: *i)* risolvere il problema legato al fatto che il confine tra terminologia di dominio e lingua comune non sempre è così netto (Cabré, 1999) e *ii)* delineare dei criteri comuni nella definizione di unità terminologica polirematica (Ramisch, 2015), dal momento che esse rappresentano più della metà del vocabolario di un madre-lingua (Jackendoff, 1997). La metodologia proposta in questo contributo suggerisce una strategia per superare entrambi tali aspetti problematici. Come descritto in Bonin et al. (2010), la soluzione proposta si basa su di una originale combinazione di filtri linguistici e statistici che permettono di *i)* discriminare la terminologia di dominio dalla lingua comune impiegando metriche statistiche che pesano la rilevanza dei termini estratti all'interno del corpus di acquisizione (corpus di dominio) rispetto ad un corpus di riferimento (corpus rappresentativo della lingua comune, tipicamente una collezione di articoli di giornale); *ii)* estrarre unità polirematiche anche nei casi in cui la corrispondente testa lessicale non sia stata precedentemente individuata come unità monorematica specifica del dominio. L'intuizione è di considerarle come elementi 'unici' costituiti da sequenze di categorie morfosintattiche (vedi Sezione 3.2). Ciò permette di suggerire una risposta all'osservazione che "non sempre la settorialità di un LC [lessema complesso] è connessa con l'esistenza di accezioni speciali dei membri componenti, ma può derivare dal fatto che il LC assume in determinati contesti un significato globale speciale" (De Mauro and Voghera, 1996).

3 Metodo

3.1 Descrizione e preparazione del corpus

Gli articoli della rivista sono reperibili online sia in .pdf che in .xhtml e, per i numeri più recenti, anche in .xml (TEI-XML). Il corpus su cui si basa la nostra indagine parte dall'estrazione del *plain text* dall'XHTML, una volta escluso il contenuto metatestuale e paratestuale. La Tabella 1 mostra la composizione del corpus.

Volume	#Articoli	#Parole	Lungh. media
1	6	21,480	4,198 parole
2	8	26,469	3,308 parole
3	7	38,327	5,475 parole
4	8	29,431	3,678 parole
5	7	24,921	3,560 parole
6	6	21,681	3,613 parole
7	5	26,528	5,305 parole
8	16	70,025	4,376 parole
9	6	23,897	3,982 parole
10	6	31,992	5,332 parole
TOT.	**75**	**314,751**	

Tabella 1: Composizione del corpus e lunghezza media degli articoli.

3.2 Estrazione terminologica

Per studiare la variazione terminologica avvenuta nel corso degli anni di pubblicazione della rivista abbiamo adottato due metodi complementari: il primo basato sull'indicizzazione del corpus tramite la terminologia estratta in modo non supervisionato e il secondo basato sull'indicizzazione dello stesso corpus tramite parole chiave fornite dagli autori come metadati degli articoli.

Il processo di estrazione terminologica non supervisionata è stato realizzato grazie a *Text-to-Knowledge (T2K)* (Dell'Orletta et al., 2014), piattaforma di estrazione e organizzazione della conoscenza da corpora multilingui di dominio basata su tecnologie di Natural Language Processing sviluppata da ILC-CNR e ampiamente validata in diversi contesti applicativi[2]. T2K, costruito su di un'originale combinazione di sistemi a regole e algoritmi basati su metodi di apprendimento automatico, consente di estrarre da una collezione di testi linguisticamente annotati entità rilevanti anche quando esse non sono presenti in una risorsa semantico-lessicale di dominio a disposizione. Ciò permette di far fronte e superare il tradizionale collo di bottiglia che si incontra in ogni compito di analisi semantica del testo, quello cioè di rendere esplicito il collegamento tra la realizzazione linguistica dell'informazione e la rappresentazione esplicita dell'informazione stessa.

Allo scopo pertanto di individuare ed estrarre elementi informativi nuovi rispetto a quelli presenti nel repertorio delle parole chiave a disposizione, il corpus è stato linguisticamente annotato a diversi livelli di analisi. A partire dal testo annotato a livello morfosintattico grazie al Parts-Of-Speech tagger descritto in Dell'Orletta (2009), sono state individuate le unità terminologiche candidate all'estrazione. La metodologia, descritta in Bonin et al. (2010), consente di individuare potenziali unità monorematiche e polirematiche impiegando una combinazione di filtri linguistici e statistici configurabili rispetto agli

[2] http://www.italianlp.it/demo/t2k-text-to-knowledge/

obiettivi di ricerca. Allo scopo della nostra indagine, i filtri linguistici sono stati configurati in modo da individuare all'interno del corpus di acquisizione: *i)* le potenziali unità monorematiche, sulla base della categoria morfo-sintattica assegnata (tipicamente 'sostantivo'); *ii)* le potenziali unità polirematiche, sulla base di una serie di sequenze di categorie morfo-sintattiche rappresentative di diversi tipi di modificazione nominale. Ad esempio, da una sequenza come 'aggettivo+sostantivo' sono individuate polirematiche quali *critical edition, lexical entry, cultural heritage*; da una sequenza 'sostantivo+sostantivo' sono individuati potenziali termini quali *TEI standard, manuscript material, knowledge representation*; per arrivare a sequenze più complesse come 'sostantivo+preposizione+sostantivo' sulla base della quale sono stati individuati termini quali *string of text, editions of letters* o sequenze 'sostantivo+preposizione+aggettivo+sostantivo' che permette di rintracciare un termine come *DTABf for printed texts, evaluation of digital scholarship* o 'aggettivo+aggettivo+sostantivo' realizzazione linguistica di un termine come *historical financial records*. I filtri statistici, applicati alla lista di termini candidati all'estrazione, consentono di ordinare tali termini sulla base della loro rilevanza all'interno del corpus di acquisizione, attribuendo loro un valore di significatività stabilita sulla base del C-NC Value (Frantzi and Ananiadou, 1999), una delle misure più utilizzate nei sistemi di estrazione terminologica.

In linea con gli obiettivi di ricerca del nostro studio, i termini così estratti sono stati impiegati dal modulo di indicizzazione di T2K per rintracciare all'interno dell'intera collezione di articoli del *jTEI* i singoli contesti nei quali i termini compaiono. Grazie a questo processo è stato possibile condurre l'indagine sulla variazione diacronica dei termini nelle diverse annate della rivista, consentendo di studiare l'evoluzione di tendenze di ricerca e tematiche di studio.

3.3 Trattamento delle parole chiave

Sono state prese in considerazione le parole chiave che gli autori stessi hanno indicato fra i metadati. Sul totale degli articoli raccolti le parole chiave distinte sono 259.

3.4 Mann-Kendall Trend Test

Per esplorare le variazioni significative d'impiego dei termini e delle parole chiave nell'intervallo temporale osservato, è stato scelto il Mann-Kendall trend test, disponibile nel pacchetto `trend` di `R` (`https://bit.ly/30bWRkd`). Considerando il numero esiguo di dati disponibili per ciascun termine (o parola chiave) si è scelta quindi una statistica non parametrica sufficientemente affidabile anche con un numero di misurazioni inferiori a dieci. Per motivi di omogeneità dei dati, sono stati presi in considerazione soltanto i sette numeri della rivista riguardanti atti di convegni presi in successione cronologica, come si può vedere nelle Figure 3 e 4. I dati su cui si è applicato l'MK Test sono stati preparati in formato tabellare sia per i termini estratti automaticamente, sia per le parole chiave indicate dagli autori, disponendo su ciascuna riga un termine (o una parola chiave), su ciascuna colonna un numero della rivista e in ciascuna cella la relativa frequenza percentuale. L'MK Test fornisce un valore positivo per trend crescenti e un valore negativo per trend decrescenti. Per lo studio dei risultati sono stati presi in considerazione soltanto gli esiti con p-value$<$0.05.

4 Risultati

4.1 Studio dei profili degli autori

Dall'analisi dei trend terminologici i numeri della rivista non dedicati ad atti dei convegni TEI (3, 5 e 7) sono stati esclusi anche perché i profili degli autori stessi hanno carattere di eccezione. Per il monitoraggio, gli autori sono stati classificati in base alla loro presenza o meno in riviste o atti di convegno di Linguistica Computazionale (con contributi o con menzioni in bibliografia). Come si può vedere in Fig. 1, il numero dedicato a TEI e linguistica (3) e il numero aperto (7) hanno attratto un numero elevato di linguisti computazionali. Sorprendentemente invece il numero dedicato alle infrastrutture TEI (5) non ha avuto la stessa attrattiva.

4.2 Dati relativi ai termini estratti

I risultati discussi in quanto segue fanno riferimento ai primi 500 termini circa mono- e polirematici estratti, con una frequenza di occorrenza $\geq$3. La Tabella 2 riporta un estratto della lista dei primi 25 termini estratti dall'intero corpus, ordinati per rilevanza statistica e accompagnati dalla frequenza assoluta nel corpus. Per ogni termine, T2K permette di estrarre il lemma e la forma prototipica, cioè la variante linguistica più frequente del

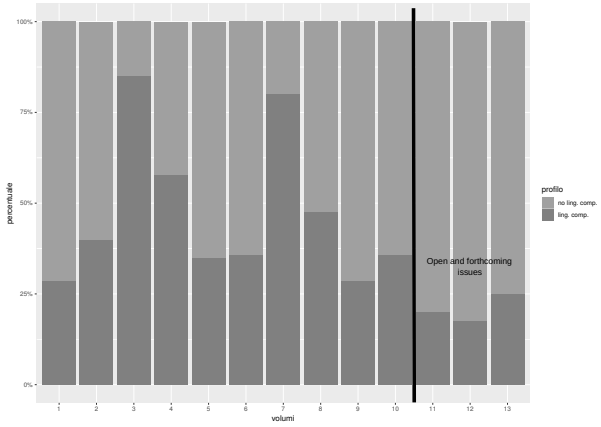

Figura 1: Autori che non hanno pubblicazioni in ambito di linguistica computazionale (no lc) e autori che ne hanno (lc)

lemma all'interno della collezione documentale di partenza.

Come introdotto nella Sezione 3.2, la fase di indicizzazione ha permesso di calcolare la distribuzione dei termini all'interno dei singoli articoli mettendo in evidenza eventuali differenze nell'uso di uno stesso termine. La Figura 2 mostra ad esempio come, sul totale di occorrenze di parole poliremathiche estratte che contengono l'aggettivo *digital*, ogni volume sia caratterizzato da distribuzioni percentuali diverse. Alcuni termini possono considerarsi poco specifici come *digital age, digital form, digital resources, digital tools, digital projects, digital medium*. Non pochi termini risultano essere tuttavia puntuali e peculiari del settore, tra questi sono stati estratti nell'arco temporale *digital archive, digital critical editions, digital document, digital editions, digital Humanities, digital images, digital library, digital objects, digital scholarship, digital text*. Il grafico permette di leggere la modulazione diacronica dei termini introdotti dagli autori e riconoscibili nel settore delle Digital Humanities. Ad esempio, possiamo notare come il termine *Digital Humanities* è il termine che ha un significato più ampio e accoglie gli altri termini peculiari. Esso è pertanto sempre presente nei dieci volumi anche se la frequenza di occorrenza risulta essere altalenante. Un momento di prosperità di questo termine risulta circoscritto al volume 6 del 2013.

4.3 Distribuzione delle parole chiave nel testo

Abbiamo verificato la distribuzione delle parole chiave nel corpo degli articoli e ciò ci ha permesso di individuare, fra le complessive 259, 32 parole chiave usate esclusivamente come metadati, e quindi che non occorrono mai nel testo, come ad esempio *bibliographical standards*, *collaborative workflow*, *TEI corpora* e 227 impiegate invece anche all'interno del testo (ad esempio *forums*).

Un'assimetria degna di nota riguarda le sequenze aggettivo+sostantivo *critical edition* e *scholarly edition* (entrambe parole chiave) in composizione con *digital*. Mentre infatti gli autori hanno indicato nei metadati degli articoli *digital scholarly edition* come parola chiave autonoma, hanno tra-

Forma prototipica	Lemma	Frequenza
TEI	TEI	2597
text	text	1261
element	element	934
project	project	485
user	user	455
document	document	421
manuscript	manuscript	396
XML	XML	393
annotation	annotation	292
TEI Guidelines	TEI Guidelines	166
edition	edition	253
tools	tool	249
information	information	248
content	content	224
language	language	221
object	object	219
source	source	214
TEI P5	TEI P5	132
TEI Consortium	TEI consortium	98
TEI documents	TEI document	91
digital editions	digital edition	89
TEI XML	TEI XML	85
TEI community	TEI community	71
manuscript description	manuscript description	54
digital humanities	digital humanity	53

Tabella 2: I primi 25 termini estratti dall'intero corpus.

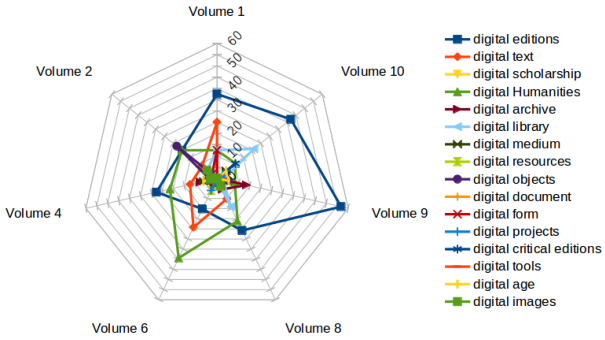

Figura 2: Distribuzione percentuale di termini poliremathici estratti che contengono l'aggettivo *digital*.

lasciato invece *digital critical edition*, benché sia termine polirematico estratto da T2K e in alcuni articoli cooccorra *digital scholarly edition*.

4.4 Risultati dell'MK Test

Lo studio delle variazioni d'impiego dei termini al fine di identificare delle tendenze significative ha prodotto i seguenti risultati con trend crescente: *different types*, `@corresp` *attribute*, *open data*, *TEI Correspondence SIG*, *research questions*, *work in progress*, *Berlin-Brandenburg Academy of Sciences*, *bibliographic references*, *TEI model*, *TEI Simple*, *case study*, *TEI XML*; e i seguenti risultati con trend decrescente: *author's note*, *literary texts*, *manuscript material*, *TEI users*, *humanities research*, *TEI-encoded documents*.

Se si escludono termini isolati oppure legati a tecnologie specifiche o a particolari gruppi di ricerca, i dati sembrano far emergere una tendenza interessante. Come si può vedere in Fig. 3, aumenta l'impiego di termini condivisi con le altre scienze con basi sperimentali, fra cui le scienze del linguaggio di cui la Linguistica Computazionale fa parte, come *research questions*, *case study* e *open data*, mentre diminuisce l'impiego di termini specifici delle discipline umanistiche, come *literary texts*, *manuscript material* e *humanities research*.

Infine, lo studio delle variazioni d'impiego significative delle parole chiave indicate come metadati dagli autori stessi (Fig. 4) mostra il crescente interesse verso il web semantico (*sense* è largamente impiegato in contesti relativi alla codifica di ontologie) e verso progetti volti a rendere TEI maggiormente usabile come *TEI Simple* (`https://tei-c.org/2014/09/10/tei-simple`). Scende invece drasticamente l'impiego di parole chiave che esprimono tecnologie o concetti ormai assodati e condivisi, come *Unicode* e *community*, parola quest'ultima comprensibilmente dominante nel primo numero della rivista.

5 Conclusione

Recuperare un campione del trend delle attività di ricerca di un particolare settore scientifico, come quelle delle Digital Humanities attraverso il jTEI, può essere stimolante per comprendere gli ambiti indagati dai vari autori nell'arco temporale di dieci anni. In particolare la disponibilità di catturare oggi, articoli open access crea opportunità per l'analisi di comunità scientifiche che nel passato non era concepibile. Il lavoro svolto rappresenta una prima esperienza di recupero informativo e di analisi per studiare il trend della comunità scientifica delle Digital Humanities attraverso una rivista ad essa dedicata, il jTEI. Pensiamo altresì che sia fondamentale ampliare le nostre fonti con altre tipologie di riferimento: come blog, forum, atti di conferenze nazionali e internazionali e riviste. Nell'analisi uno sguardo sarà rivolto anche agli autori per comprendere connessioni e estraneità tra la linguistica computazionale e le Digital Humanities.

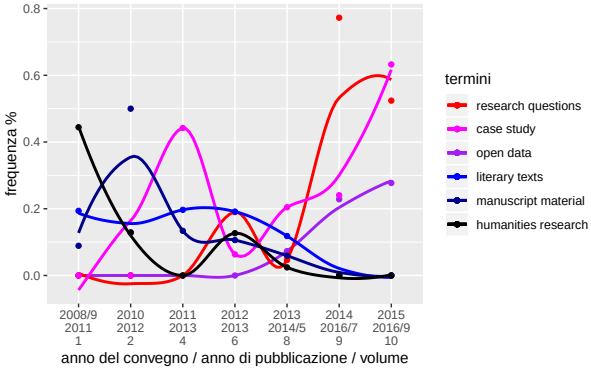

Figura 3: Trends dei termini

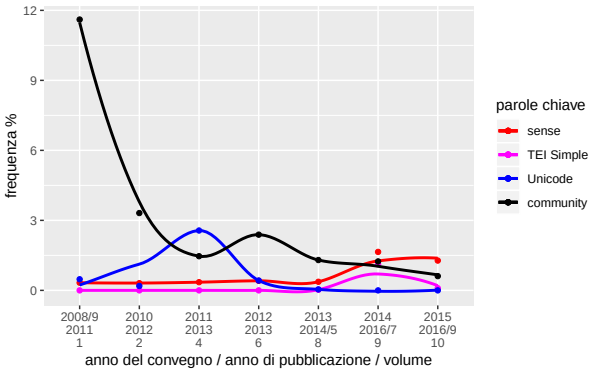

Figura 4: Trends delle parole chiave

References

R. Bartolini, S. Goggi, M. Monachini and G. Pardelli 2018. The LREC Workshops Map. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 201)*, ELRA, Paris, pp. 557-562. `https://aclweb.org/anthology/papers/L/L18/L18-1088/`

F. Bonin, F. Dell'Orletta, S. Montemagni and G. Venturi. 2010. A Contrastive Approach to Multi-word Extraction from Domain-specific Corpora. *Proceedings of 7th Edition of International Conference on*

Language Resources and Evaluation (LREC 2010), 17-23 May, Valletta, Malta.

M. T. Cabré. 1999. The terminology. Theory, methods and applications. John Benjamins Publishing Company.

R. Del Gratta, S. Goggi, G. Pardelli and N. Calzolari 2018. LREMap, a Song of Resources and Evaluation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. ELRA, Paris, pp. 1275-1281. https://www.aclweb.org/anthology/L18-1203

F. Dell'Orletta. 2009. Ensemble system for Part-of-Speech tagging. *Proceedings of Evalita'09, Evaluation of NLP and Speech Tools for Italian* , Reggio Emilia, December.

F. Dell'Orletta, G. Venturi, A. Cimino and S. Montemagni. 2014. T2K: a System for Automatically Extracting and Organizing Knowledge from Texts. *Proceedings of 9th Edition of International Conference on Language Resources and Evaluation (LREC 2014)*, 26-31 May, Reykjavik, Iceland.

T. De Mauro and M. Voghera. 1996. Scala mobile. Un punto di vista sui lessemi complessi. P. Benincà et al. (eds.), *Italiano e dialetti nel tempo. Saggi di grammatica per Giulio C. Lepschy*, Roma, Bulzoni, pp. 99-131.

G. Francopoulo, J. Mariani and P. Paroubek 2016. A Study of Reuse and Plagiarism in LREC papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, ELRA, Paris, pp. 1890-1897. https://www.aclweb.org/anthology/L16-1298

K. Frantzi and S. Ananiadou. 1999. The C-value / NC Value domain independent method for multiword term extraction. *Journal of Natural Language Processing*, 6(3):145-179.

R. Jackendoff. 1997. Twistin' the night away. *Language*, 73, pp. 534-559.

J. Mariani, P. Paroubek, G. Francopoulo and O. Hamon 2014. Rediscovering 15 Years of Discoveries in Language Resources and Evaluation: The LREC Anthology Analysis. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, ELRA, Paris, pp. 4632-4669. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1228_Paper.pdf

G. Pardelli, S. Goggi and F. Boschetti 2019. Strolling around the dawn of Digital Humanities. *Book of Abstract for the 8th Annual Conference AIUCD 2019*, pp. 261-264.

G. Pardelli, S. Goggi, R. Bartolini, I. Russo and M. Monachini 2017. A Geographical Visualization of GL Communities: A Snapshot. In *Eighteenth International Conference on Grey Literature: Leveraging Diversity in Grey Literature*, Washington, November 28-29, 2016. Edited by Dominic Farace and Jerry Frantzen, TransAtlantic-Amsterdam, 18, pp. 109-113.

T. Pohlert. 2018. *Non-Parametric Trend Tests and Change-Point Detection*, CRAN. https://bit.ly/30bWRkd,

C. Ramisch. 2015. Multiword expressions acquisition: A generic and open framework. New York: Springer.

S. Schreibman. 2011. Editorial Introduction to the First Issue. *Journal of the Text Encoding Initiative*, 1. http://journals.openedition.org/jtei/229

S. Schreibman. 2014. Editorial Introduction to Issue 7 of the Journal of the Text Encoding Initiative. *Journal of the Text Encoding Initiative*, 7. http://journals.openedition.org/jtei/1046

C. Soria, N. Calzolari, M. Monachini, V. Quochi, N. Bel, K. Choukri, M. Mariani, J. Odijk and S. Piperidis 2014. The language resource Strategic Agenda: the FLaReNet synthesis of community recommendations. *Language Resources and Evaluation*, December 2014, 48 (4), pp. 753–775. https://link.springer.com/article/10.1007/s10579-014-9279-y

R. Sprugnoli, G. Pardelli, F. Boschetti and R. Del Gratta. 2019. Un'Analisi Multidimensionale della Ricerca Italiana nel Campo delle Digital Humanities e della Linguistica Computazionale *Umanistica Digitale*, ISSN 2532-8816, 5, pp. 59-89. https://umanisticadigitale.unibo.it/article/view/8581

A. Tuzzi. 2018. Tracing the Life Cycle of Ideas in the Humanities and Social Sciences. New York: Springer.