Risorse linguistiche di varietà storiche di italiano: il progetto TrAVaSI[1]

Manuel Favaro
Istituto di Linguistica Computazionale "A. Zampolli" - CNR
manuel.favaro@ilc.cnr.it

Marco Biffi
Università di Firenze
Accademia della Crusca
marco.biffi@unifi.it

Simonetta Montemagni
Istituto di Linguistica Computazionale "A. Zampolli" - CNR
simonetta.montemagni@ilc.cnr.it

Abstract

Italiano Questo contributo si propone di presentare il progetto TrAVaSI (*Trattamento Automatico di Varietà Storiche di Italiano*), il cui obiettivo è la creazione di risorse per il trattamento automatico di varietà storiche della lingua italiana, in particolare lessici diacronici e corpora arricchiti con annotazione linguistica da utilizzare per lo sviluppo e/o la specializzazione di strumenti di annotazione. Il contributo illustra gli obiettivi, i primi risultati conseguiti e le prospettive di sviluppo.

English *The paper is aimed at illustrating the TrAVaSI project, whose aim is the design and development of language resources for the automatic processing of historical varieties of the Italian language, in particular diachronic lexicons and corpora enriched with linguistic annotation to be used for the development and/or specialization of annotation tools. The results achieved so far are reported together with current and future directions of research.*

1 Introduzione

Sono ormai numerosi gli archivi testuali digitali che testimoniano varietà storiche dell'italiano. Tuttavia, l'accesso e l'interrogazione dei testi sono spesso elementari, per lo più limitati alle stringhe di caratteri che costituiscono il testo; ciò complica il lavoro degli studiosi e rende pressoché impossibile l'utilizzo delle risorse da parte degli utenti non addetti ai lavori. Questa situazione mostra che, nonostante i recenti progressi nel settore delle Digital Humanities, l'accesso e l'interrogazione di testi che testimoniano varietà storiche di italiano, più o meno lontane nel tempo, rappresentano ancora oggi una sfida.

Il progetto TrAVaSI (*Trattamento Automatico delle Varietà Storiche di Italiano*), nato dalla collaborazione tra l'Istituto di Linguistica Computazionale "Antonio Zampolli" e l'Accademia della Crusca e finanziato dalla Regione Toscana, si propone di affrontare questa sfida, creando i presupposti per la navigazione e l'interrogazione sistematica di fonti che documentano le varietà storiche della lingua.

Il punto di partenza del progetto è pragmatico: potenziare due strumenti realizzati dall'Accademia della Crusca all'interno di altri progetti. Gli strumenti in questione sono la versione elettronica del *Grande Dizionario della lingua italiana* (*GDLI*)[2] e la banca dati del *Vocabolario Dinamico dell'Italiano Moderno*, che è più precisamente l'oggetto specifico di indagine del contributo che qui presentiamo. TrAVaSI ha come obiettivo principale quello di massimizzare le implementazioni pratiche, ma proiettandole nel quadro dello sviluppo di strumenti di riferimento per banche dati diacronicamente connotate e i dizionari storici in versione elettronica. In particolar modo si tratta di mettere a frutto l'occasione di lavorare in contemporanea (e quindi di intersecare ricerche, risultati e prodotti parziali) da un lato sulla strutturazione e marcatura di un dizionario storico come il

[2] Un prototipo del *GDLI* è attualmente consultabile negli Scaffali digitali del sito web dell'Accademia o direttamente all'indirizzo <http://www.dgli.it>.

GDLI nella sua versione informatizzata (mettendo a punto procedure di collazione semi-automatica del testo ottenuto con l'OCR e di marcatura dei campi strutturali identificati; vedi Sassolini *et alii* 2019, Biffi e Sassolini 2020), e dall'altro sulla creazione di lessici computazionali differenziati in diacronia per sopperire finalmente all'indebolimento – anche fino al grado zero – dell'efficacia di strumenti di annotazione linguistica quando ci si allontani dalle condizioni ideali dell'italiano contemporaneo scritto di quelle varietà sostanzialmente riconducibili al campione del *Lessico di frequenza dell'italiano contemporaneo* (LIF). E così determinare un allargamento dell'efficacia degli strumenti di annotazione linguistica in diacronia, in diamesia e in diafasia.

Sul versante degli studi umanistici, gli sviluppi di questi strumenti hanno ricadute notevoli, sia nell'ottica di consegnare agli studiosi metodi di indagine sempre più potenti per le loro ricerche linguistiche, sia in quella, cara a un'istituzione come l'Accademia della Crusca, di raggiungere un'utenza più vasta. La digitalizzazione dei testi e la loro disponibilità in rete – infatti –non sono sufficienti per avvicinare un'utenza non circoscritta agli addetti ai lavori al patrimonio culturale tramandato. Per rendere fruibili da una vasta e variegata utenza i contenuti di istituzioni culturali, è necessario offrire modalità di accesso e navigazione dotate di "intelligenza linguistica".

Ad oggi, è ampiamente riconosciuto che l'applicazione di metodi e delle tecniche per il Trattamento Automatico della Lingua (TAL) a varietà storiche di una lingua presenta innumerevoli ostacoli, poiché gli strumenti sviluppati per le lingue moderne necessitano di specializzazioni a vario livello (lessicale, morfologico e sintattico) per essere utilizzati con successo nel trattamento di fonti primarie che rappresentano la base degli studi umanistici. Per la lingua italiana, Pennacchiotti e Zanzotto (2008) riportano i risultati di uno studio esplorativo sulle difficoltà derivanti dal trattamento automatico di varietà storiche della lingua italiana basato su un corpus diacronico che raccoglie testi italiani letterari (sia in prosa che in poesia) dal 1200 a fine Ottocento: le analisi, focalizzate sulla composizione del vocabolario e sull'annotazione morfologica e morfo-sintattica, confermano che il trattamento automatico di varietà storiche della lingua italiana è una sfida aperta. Una delle sfide che il progetto TrAVaSI intende affrontare. In questa area, l'azione del progetto ruota attorno a due direttrici principali, riguardanti la costruzione di corpora arricchiti con informazione linguistica da usarsi per la valutazione e/o addestramento di strumenti di annotazione, e di lessici computazionali diacronici in grado di supportare il processo di lemmatizzazione del testo. Al momento, l'annotazione linguistica dei testi riguarda il livello morfo-sintattico. Le risorse sviluppate su entrambi i fronti verranno integrate all'interno di infrastrutture di ricerca preesistenti (per esempio CLARIN-IT), al fine di incrementarne la visibilità, l'accessibilità e l'interoperabilità con risorse simili.

Sul versante linguistico-computazionale, la progettazione e costruzione di risorse e strumenti per varietà d'uso della lingua che si discostano in diversa misura dall'italiano contemporaneo scritto su cui gli strumenti di TAL sono tipicamente addestrati costituisce un fertile terreno di sperimentazione per lo sviluppo e il raffinamento di tecnologie innovative di TAL.

L'articolo illustra la metodologia definita per la costruzione delle risorse e i primi risultati raggiunti. La sezione 2 descrive il corpus diacronico selezionato per l'annotazione, e la sezione 3 illustra il metodo seguito per l'annotazione del corpus. La sezione 4 riporta e discute i risultati dei primi esperimenti di annotazione semi-automatica e della strategia di revisione messa a punto. La sezione 5 illustra i risultati di annotazione automatica conseguiti con sistemi esistenti di annotazione per la lingua italiana contemporanea; infine, la sezione 6 delinea gli sviluppi in corso e futuri.

2 Corpus

La costruzione delle risorse si avvarrà in primo luogo dei testi presenti nel corpus per la realizzazione del *Vocabolario Dinamico dell'Italiano Moderno* (*VoDIM*, Marazzini e Maconi, 2018). Il corpus è frutto di due

progetti nazionali: il PRIN 2012 (*Corpus di riferimento per un Nuovo Vocabolario dell'Italiano moderno e contemporaneo. Fonti documentarie, retrodatazioni, innovazioni*, diretto da Claudio Marazzini) e il PRIN 2015 (*Vocabolario dinamico dell'italiano post-unitario*, sempre sotto la direzione di Claudio Marazzini); i due progetti, l'uno la prosecuzione dell'altro, hanno coinvolto diverse università (Catania, Firenze, Genova, Milano, Napoli, Piemonte Orientale e Torino) e l'Istituto di Teorie e Tecniche dell'Informazione Giuridica (ITTIG) del CNR di Firenze.

Il *VoDIM* riunisce testi, scritti e orali, attinenti all'italiano post-unitario che riguardano diversi domini dello spazio linguistico dell'italiano: arte, canzone, diritto, economia, gastronomia, poesia, politica, prosa giornalistica, letteraria, paraletteraria e scientifica. Il corpus, diacronicamente bilanciato, ha una estensione di circa 20 milioni di parole; nel prossimo futuro il progetto prevede che il corpus iniziale sia affiancato da un corpus sincronico molto più ampio (circa 2 miliardi di parole), composto da testi ricavati dalla rete (Biffi, 2016, Biffi, 2018, Biffi, 2020, Biffi e Ferrari, 2020). Inoltre, il *VoDIM* è stato di recente inserito all'interno della *Stazione Lessicografica*, consultabile tramite gli *Scaffali digitali* del sito dell'Accademia della Crusca (Biffi 2020: 360-362).[3]

3 Metodo

La specializzazione di strumenti di annotazione linguistica rispetto a varietà d'uso della lingua diverse da quelle testimoniate nel corpus di addestramento richiede – innanzitutto – la disponibilità di risorse (lessici e corpora annotati) rappresentative della varietà di lingua da trattare. In primo luogo per verificare il livello di accuratezza degli strumenti di annotazione disponibili, e – successivamente – per la loro specializzazione. Il corpus alla base del *VoDIM* si pone come ottima "palestra" da questo punto di vista, in quanto i testi al suo interno si distribuiscono in un vasto arco cronologico (dall'Unità a oggi), presentano una notevole differenziazione diamesica (scritto, parlato, parlato-scritto e scritto-parlato) e appartengono a un'ampia varietà di generi e di tipologie testuali.

Per l'arricchimento dei testi con annotazioni linguistiche di varia natura ci si avvarrà, dal punto di vista metodologico, dell'esperienza maturata nel corso del progetto *Voci della Grande Guerra* (VGG, De Felice *et al*., 2018, Lenci *et al*. 2020). In particolare, i metodi adottati per risolvere i problemi di segmentazione delle forme e di lemmatizzazione emersi durante la fase di trattamento automatico (De Felice *et al*., 2018: 161-162) sono un fondamentale punto di partenza per ottenere risultati efficienti, come nel caso del corpus *VGG*.

Il lavoro è stato articolato nelle seguenti fasi operative:

1. selezione delle fonti a partire dal corpus *VoDIM*;
2. definizione di un sottocorpus rappresentativo delle varietà del *VoDIM*; i testi scelti appartengono a sette dei domini del corpus (arte, cucina, diritto, giornali, letteratura, paraletteratura, scienze) e sono stati bilanciati in diacronia per avere la massima copertura possibile; i campioni di ogni dominio sono tra i 2.600 e i 3.000 token, per un'estensione totale del sottocorpus – ad oggi - di circa 19.000 token;
3. annotazione morfo-sintattica e lemmatizzato del sottocorpus di cui al punto 2. Lo schema di annotazione adottato è quello sviluppato all'interno dell'iniziativa internazionale *Universal Dependencies* (Nivre, 2015), che rappresenta ad oggi uno standard *de facto* per l'annotazione morfo-sintattica e sintattica a dipendenze di testi, incluse varietà storiche di alcune delle lingue trattate. Il sottocorpus selezionato è stato annotato automaticamente con *UDPipe* (Straka e Straková, 2017), addestrato sulla *Italian Universal Dependency Treebank* (IUDT, Bosco et *al.*, 2013). A ciò, ha fatto seguito una fase di revisione manuale degli errori riscontrati nei testi annotati automaticamente.

[3] http://www.stazionelessicografica.it

Il campione testuale, frutto delle prime fasi di lavoro, è di fatto il primo nucleo di corpus annotato per la validazione e, in prospettiva, l'addestramento e/o specializzazione degli strumenti di trattamento automatico di varietà storiche dell'italiano.

4 Primi risultati

4.1 Analisi quantitativa degli errori

L'analisi degli errori di annotazione morfosintattica e di lemmatizzazione riscontrati nei testi annotati automaticamente con *UDPipe* ha fornito dati importanti sulla mole di errori presenti in ogni testo; nella Tabella 1 sono riportate le percentuali medie di errore nei testi, distribuiti su sei intervalli cronologici che ricalcano grossomodo la suddivisione in periodi proposta per il *DiaCORIS* (Onelli *et al.*, 2006) e successivamente per il *LIS Lessico Italiano Scritto*[4], a cui viene aggiunto un sesto periodo comprendente i testi attinenti all'italiano contemporaneo:

intervalli temporali		n. testi	n. token	% media errori
1	1861-1900	6	4300	8,2%
2	1901-1922	5	3400	8,6%
3	1923-1945	4	3300	5,3%
4	1946-1967	1	600	8,3%
5	1968-2001	7	3800	5,3%
6	2002-oggi	18[5]	4100	3,8%

Tabella 1. Percentuale di errori per intervalli cronologici.

Come ci si aspetterebbe, la percentuale di errore tende a scendere, con il passare degli anni, dall'8 al 3%; l'unica eccezione riguarda il periodo 1946-1967, testimoniato però dal campione di circa 600 token estratto da un solo testo (un saggio di medicina di G. Brotzu, *Ricerche su di un nuovo antibiotico* del 1948); tale valore è dunque poco attendibile – se si considera, tra l'altro, che è il testo del dominio "scienze" con una percentuale di errore tra le più alte (cfr. *infra*).

Più interessanti i dati riportati nella Tabella 2, in cui sono presenti le percentuali medie di errore per ogni dominio:

domini	intervalli temporali		n. testi	n. token	% media errori
arte	2-6	1902-2009[6]	4	2600	6,3%
cucina	1-3	1871-1927	4	2700	9,9%
diritto	5-6	2000-2016	17	3000	3,7%
giornali	1-5	1867-1996	4	3000	5,5%
letteratura	1-5	1881-1982	4	2800	8,5%
paraletteratura	1-3	1892-1939	4	2600	5,9%
scienze	1-6	1864-2015	4	2700	5,6%

Tabella 2. Percentuale di errori per domini.

Osservando i dati si nota che le percentuali più alte riguardano i testi gastronomici e letterari, quelle più basse i testi giuridici; gli altri domini si attestano su una media del 5-6%. Una prima spiegazione di questa diffrazione risiede certamente nelle date di pubblicazione dei testi interessati: i libri di gastronomia da cui sono stati estratti i campioni per la cucina sono stati pubblicati tra il 1871 (l'anonimo ricettario *Il cuoco sapiente*) e il 1927 (il famosissimo *Talismano della felicità* di Ada Boni), mentre i testi del dominio "diritto" sono usciti

[4] Il *LIS* corrisponde alla sezione *DIACORIS* del CILTA di Bologna (<http://corpora.dslo.unibo.it/coris_ita.html>), sviluppata all'interno del portale *Vivit Vivi Italiano. Il portale dell'italiano nel mondo* (<http://www.viv-it.org/schede/archivi-digitali>), e costituisce una rielaborazione informatica per omogeneizzare la banca dati *LIT Lessico Italiano Televisivo* e *LIR Lessico Italiano Radiofonico* (anch'essi presenti nel portale) in modo che possano essere interrogate contemporaneamente da un metamotore (come avviene appunto, affiancando la singola consultazione delle tre banche dati, nella sezione "Archivio Digitale del Vivit: <http://www.viv-it.org/schede/archivi-digitali>). Il *LIS* comprende complessivamente 25 milioni di occorrenze, distribuite equamente su cinque sezioni cronologiche bilanciate (per approfondimenti cfr. Biffi, 2016, pp. 276-77 e nota 24).

[5] Il numero è così elevato perché comprende per la maggior parte i testi del dominio "diritto", al cui interno si trovano abstract di saggi e di opere specialistiche; i testi sono perciò molto brevi e, a differenza degli altri domini, è stato necessario attingere a molti più campioni testuali per raggiungere la quota prestabilita di token (cfr. Tabella 2).

[6] Le date presenti nella colonna si riferiscono alla pubblicazione rispettivamente del primo e dell'ultimo testo del dominio.

tra il 2000 e il 2016. Tuttavia, va considerato che i testi paraletterari, con percentuali di errore nella media, sono stati anch'essi pubblicati tra la fine dell'Ottocento e l'inizio del Novecento; e che, d'altro canto, le opere di letteratura hanno valori elevati, nonostante la presenza di romanzi abbastanza recenti come *Se non ora, quando?* di Primo Levi (1982). Probabilmente lo scarto elevato dalla media è da ricondursi, per la cucina, alle particolari tipologie testuali (come ad esempio le ricette) "nuove" nel panorama dei campioni su cui sono stati testati finora gli strumenti di annotazione. Analoghe considerazioni valgono per la letteratura. Sebbene i corpora letterari siano stati oggetto di analisi automatiche (es. lemmatizzazione) finalizzate all'interrogazione, secondo la prospettiva adottata in questo studio essi appaiono caratterizzati da tratti morfologici quantitativamente significativi su cui gli analizzatori finora utilizzati non sono adeguatamente addestrati (es. forme verbali di prima e seconda persona).

4.2 Analisi qualitativa degli errori

Sembrerebbe essere molto forte, dunque, il legame tra errore, dominio e, soprattutto, genere testuale[7]. Per esempio, i ricettari, così come la manualistica in generale, sono costellati da verbi all'imperativo di seconda plurale (per esempio *mescolate*) che *UDPipe* tende ad assimilare con le forme del participio passato; i testi letterari, dal canto loro, sono colmi di parti dialogiche che presentano al loro interno una serie di tratti, in primo luogo interpunzioni enfatiche (punti "misti" come l'accumulo di punto esclamativo e puntini di sospensione) e interiezioni, che mettono in difficoltà l'analizzatore, sia nell'annotazione morfo-sintattica, sia nella tokenizzazione; analogamente, i testi giuridici e scientifici presentano diversi errori legati alla corposa presenza di abbreviazioni, sigle e simboli.

È stata inoltre eseguita una valutazione sulle tipologie di errore, che per circa un terzo dei casi – oltre il 27% – coinvolgono soltanto il lemma, e sono senz'altro dovuti al fatto che *UDPipe*, nella versione di base, utilizza un dizionario costruito a partire dal corpus di addestramento integrato da euristiche di analisi morfologica utilizzate per trattare forme sconosciute (anch'esse automaticamente derivate dal corpus di addestramento). Seguono due esempi di lemmatizzazione basata su tali euristiche:

27 lavano **lare** VERB V Mood=Ind|Number=Plur|Person=3|Tense=Imp|VerbForm=Fin _
43 illustri **illustro** ADJ A Gender=Masc|Number=Plur _

In altri casi – circa il 20% – gli errori riguardano l'assegnazione della categoria grammaticale (sia *coarse-grained* sia *fine-grained*); tipici esempi sono il *che* pronome a cui viene attribuito il valore di congiunzione (a), e lo scambio tra aggettivo e sostantivo nella coppia nominale (b):

a.
24 Pagliarini Pagliarini PROPN SP _
25 che che **SCONJ CS** _
26 per per ADP E _
27-28 farsi _
27 far fare VERB V VerbForm=Inf _
28 si si PRON PC Clitic=Yes|Person=3|PronType=Prs _
29 ascoltare ascoltare VERB V VerbForm=Inf _
30 ha avere AUX VA Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin _
31 parlato parlare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part _
32 in in ADP E _
33 piedi piede NOUN S Gender=Masc|Number=Plur _ SpaceAfter=No
34 . . PUNCT FS _

b.
47 massimi massimo **NOUN S** Gender=Masc|Number=Plur _
48 esponenti esponente **ADJ A** Number=Plur _

Circa il 6% degli errori concerne invece solo le caratteristiche morfologiche associate alla forma (*Universal Dependencies features*); appartengono a questa tipologia gli scambi sopra menzionati tra imperativo e participio:

[7] A tal proposito, si veda il progetto EvaLatin 2020 (Sprugnoli *et al*, 2020), che mira a implementare lo studio della portabilità degli strumenti NLP per il latino attraverso la costituzione di tre *baseline* a cui appartengono diversi generi e diversi periodi cronologici; analogamente a quanto si osserva nel presente elaborato, i primi risultati di EvaLatin 2020 hanno mostrato l'impatto sia delle caratteristiche diacroniche sia di genere testuale sull'accuratezza dell'annotazione.

3 digrassate digrassare VERB V **Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part** _

Vi sono poi diversi casi – circa il 7% – di errata segmentazione dei token molto simili a quelli riscontrati nel Corpus *VGG* (De Felice *et al.*, 2018: 161), ossia il mancato riconoscimento di forme verbali enclitiche rare o desuete nell'italiano contemporaneo:

13 intendesi intendese ADJ A Number=Plur _
19 siasi siasi ADJ A Number=Sing _

Altri errori di iposegmentazione si osservano in massima parte nelle sequenze preposizionali diacronicamente marcate del tipo *pella*, *collo* ecc.:

7 co' coco DET RD Definite=Def|Number=Sing|PronType=Art _
4 colla collare VERB V Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin _
7 pei peo ADJ A Gender=Masc|Number=Plur _

Analogamente, *UDPipe* tende a non riconoscere altre combinazioni di clitici:

27 spogliatela spogliatelare NOUN S Gender=Fem|Number=Sing _
8 toglietene toglietena SCONJ CS _

I restanti errori riguardano tutte le altre possibili combinazioni presenti nella catena di annotazione:

1. lemma + categoria grammaticale + tratti morfo-sintattici (21% circa): l'annotazione automatica risulta essere totalmente sbagliata; errori di questo tipo si osservano comunemente quando, in presenza di una iniziale maiuscola, la forma viene assimilata a un nome proprio:

   3 Pigliate pigliate PROPN SP _

2. categoria grammaticale + tratti morfo-sintattici (20% circa): la forma viene correttamente lemmatizzata, ma la categoria grammaticale e i tratti associati sono errati; un caso tipico, al contrario di quanto osservato sopra, è il *che* congiunzione a cui viene attribuito il valore di pronome:

   1 Avuta avere ADJ A Gender=Fem|Number=Sing _
   2 notizia notizia NOUN S Gender=Fem|Number=Sing _
   3 che che **PRON PR PronType=Rel** _
   4 era essere AUX VA Mood=Ind|Number=Sing|Person=3|Tense=Imp|VerbForm=Fin _
   5 già già ADV B _
   6 cominciato cominciare VERB V Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part _

3. lemma + tratti morfo-sintattici (3% circa)**:** soltanto la categoria grammaticale è corretta:

   20 investe **investa** VERB V **Gender=Fem|Number=Plur|Tense=Past|VerbForm=Part** _

La revisione ha comportato non soltanto la correzione meccanica degli errori, ma anche l'adattamento di talune annotazioni ai criteri definiti per la *IUDT treebank*, al fine di garantire la compatibilità delle annotazioni dei diversi UD corpora presenti per la lingua italiana. La selezione di tali criteri è stata una delle imprese più ardue: si è optato per una strategia conservativa e un continuo confronto con i corpora di UD per ottenere il minor rumore possibile e una massima sistematizzazione del *tagset*, soprattutto per quanto concerne avverbi e congiunzioni: per esempio *come*, nei corpora UD consultati, viene sempre classificato funzionalmente con il valore di preposizione di fronte ai sintagmi normali, di congiunzione in presenza di un'altra congiunzione (*come se*), di avverbio in tutti gli altri contesti; i casi in cui sono state introdotte delle innovazioni hanno per la maggior parte riguardato incoerenze nell'annotazione: *po'*, generalmente avverbio, è stato taggato come nome in contesti quali *un po' di pane*, poiché la forma piena *poco* veniva già riconosciuta con lo stesso valore nei medesimi contesti; analogamente, gli infiniti sostantivati, a volte riconosciuti a volte no, sono stati etichettati sempre come nome.

5 *Baseline* di confronto

La scelta di *UDPipe* come strumento per la preannotazione del corpus (cfr. sezione 3) ha molteplici motivazioni, che vanno dallo schema di annotazione adottato (UD) alla possibilità di riaddestramento, che ne fanno uno strumento adeguato per la costruzione dei corpora annotati sviluppati nel progetto TrAVaSI.

In vista del riaddestramento della catena di analisi per il trattamento di varietà storiche di italiano, i risultati ottenuti con *UDPipe* sono stati confrontati con l'output di strumenti di annotazione morfo-sintattica e lemmatizzazione sviluppati in ambito DH (per esempio l'annotatore morfologico del PiSystem, Picchi, 2003) o già testati all'interno di applicazioni di DH come LinguA (Attardi e Dell'Orletta, 2009, Attardi *et al*., 2009, Dell'Orletta, 2009). I dati emersi mostrano che la baseline costituita dall'annotatore PiMorfo, sviluppato primariamente in funzione dell'interrogazione di vasti archivi testuali, presenta una percentuale media di errore assai più elevata di *UDPipe* – oltre il 9% –, ma è stato riscontrato che una parte sostanziale – all'incirca il 13% – riguarda alcuni errori sistematici, ad esempio il mancato riconoscimento della preposizione con vocale elisa *d'* o del valore di forme ambigue quali *ancora*, che in tutte le occorrenze in funzione di avverbio/congiunzione è stato confuso con la terza persona del presente indicativo di *ancorare*.

Per quanto riguarda LinguA, si è rilevato invece un indice di errore indubbiamente più basso rispetto a *UDPipe* – con una media intorno al 4% – e oltre il 40% degli errori riguarda soltanto sei tipologie, tra cui spicca l'assegnazione scorretta del tag SP (nome proprio, cfr. sezione 4.2), che viene associato a quasi tutte le occorrenze delle forme con l'iniziale maiuscola.

6 Conclusioni e sviluppi in corso

Abbiamo illustrato brevemente i passaggi che il progetto TrAVaSI sta seguendo per costruire risorse per il trattamento automatico di varietà storiche di italiano, differenziate anche in diamesia, diafasia e a livello di genere testuale. In questa fase preliminare abbiamo cercato di dare risposte a quesiti irrisolti sull'annotazione e la lemmatizzazione dell'italiano in diacronia, per esempio sul modo di trattare alcune forme (voci diverbate, infiniti sostantivati, varianti regionali, per citare alcuni esempi) e sui criteri di lemmatizzazione delle varianti fono-morfologiche dello stesso lemma. Le risorse sviluppate saranno sfruttate per testare strumenti esistenti di annotazione morfo-sintattica e di lemmatizzazione e per lo sviluppo e/o specializzazione di componenti software per il trattamento di varietà storiche della lingua italiana.

I primi risultati raccolti a partire dall'analisi del sottocorpus del *VoDIM* selezionato mostrano chiaramente che le dimensioni di variazione da tenere in considerazione sono molteplici e fortemente interrelate, derivanti da diversi tipi di processi, come la variazione nel tempo (variazione diacronica), la variazione correlata a variabili sociolinguistiche oppure legata al genere testuale o allo stile di chi scrive. I luoghi di variazione spaziano dall'ortografia, alla morfologia e alla sintassi (specialmente in diacronia e in testi di domini specialistici). Questo tipo di analisi è fondamentale per arrivare a definire una metodologia per la creazione di ulteriori risorse, con il fine di allargare progressivamente l'arco cronologico ma anche la tipologia di varietà d'uso della lingua per poter costruire strumenti TAL che siano applicabili a testi italiani dei secoli precedenti, riconducibili a diverse varietà d'uso della lingua. Da questa prospettiva, il corpus selezionato come punto di partenza del progetto TrAVaSI diventa quindi particolarmente significativo in quanto crea i presupposti per la definizione di un metodo per la specializzazione di strumenti di annotazione in relazione a molteplici varietà linguistiche, diacroniche ma anche diafasiche, diastratiche o corrispondenti a tipologie testuali.

Gli sviluppi in corso includono:

1. l'addestramento e la specializzazione dei modelli di *UDPipe* mediante test set costituiti dai domini del test corpus, su cui di volta in volta verrà valutato l'incremento di accuratezza dell'annotazione;
2. la costruzione di lessici diacronici a partire da corpora annotati – iniziando dal *VoDIM* e successivamente attingendo a risorse cronologicamente antecedenti – come base per il processo di lemmatizzazione;
3. sulla base della distanza tra il corpus utilizzato per l'addestramento e i testi da analizzare, l'identificazione del modello linguistico più adeguato da utilizzarsi per l'annotazione linguistica.

Ringraziamenti

Le attività di ricerca illustrate in questo articolo sono condotte nell'ambito del progetto *TRaVaSI* del programma di intervento denominato "CNR4C" di cui al progetto congiunto di alta formazione, cofinanziato dalla Regione Toscana con le risorse del POR FSE 2014-2020 – Asse A Occupazione, all'interno di "GiovaniSì", il progetto regionale Toscano per l'autonomia dei giovani.

Bibliografia

Attardi G., Dell'Orletta F. (2009), *Reverse Revision and Linear Tree Combination for Dependency Parsing*, in *NAACL-HLT 2009 – North American Chapter of the Association for Computational Linguistics – Human Language Technologies* (Boulder, Colorado, June 2009). Proceedings, Association for Computational Linguistics, pp. 261-264.

Attardi G., Dell'Orletta F., Simi M., Turian J. (2009), *Accurate Dependency Parsing with a Stacked Multilayer Perceptron*, Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italia, dicembre 2009).

Biffi M. (2016), *Progettare il corpus per il vocabolario postunitario*, in *L'italiano elettronico. Vocabolari, corpora, archivi testuali e sonori*, Atti della "Piazza delle Lingue" dell'Accademia della Crusca, edizione 2014 (Firenze, 6-8 novembre 2014), a cura di Claudio Marazzini e Ludovica Maconi, Firenze, Accademia della Crusca, pp. 259-80.

Biffi M. (2018), S*trumenti informatico-linguistici per la realizzazione di un dizionario dell'italiano post-unitario*, JADT'18. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data, a cura di Domenica Fioredistella Iezzi, Livia Celardo e Michelangelo Misuraca, Roma, Universitalia, 2018, vol. 1, pp. 99-107.

Biffi M. (2020), *La galassia lessicografica della Crusca in rete*, in *Italiano antico, italiano plurale. Testi e lessico del Medioevo nel mondo digitale.* Atti del convegno internazionale in occasione delle 40.000 voci del TLIO, Firenze, 13-14 settembre 2018, a cura di Lino Leonardi e Paolo Squillacioti, Alessandria, Edizioni dell'Orso, pp. 219-232.

Biffi M., Ferrari A. (2020), *Progettare e ideare un corpus dell'italiano nella rete: il caso del CoLIWeb*, «Studi di Lessicografia Italiana», vol. XXXVII, 2020, pp. 357-374.

Biffi M., Sassolini E., *Strategie e metodi per il recupero di dizionari storici*, in *La svolta inevitabile: sfide e prospettive per l'informatica umanistica*, Atti del IX Convegno Annuale dell'Associazione per l'Informatica Umanistica e la Cultura Digitale (AIUCD), Milano, Università Cattolica del Sacro Cuore, 15-17 gennaio 2020, a cura di Cristina Marras, Marco Passarotti, Greta Franzini ed Eleonora Litta, dell'Associazione per l'Informatica Umanistica e la Cultura Digitale, 2020, pp. 235-239 (pubblicazione elet-tronica in «Quaderni di Umanistica Digitale»: <http://doi.org/10.6092/unibo/amsacta/6316>).

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

De Felice I., Dell'Orletta F., Venturi F., Lenci A. Montemagni S. (2018), *Italian in the Trenches: Linguistic Annotation and Analysis of Text of the Great War*, Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-it), 10-12 dicembre 2018, Torino.

Dell'Orletta F. (2009), *Ensemble system for Part-of-Speech tagging*, Proceedings of EVALITA 2009 – Evaluation of NLP and Speech Tools for Italian 2009 (Reggio Emilia, Italy, December 2009).

GDLI = Grande dizionario della lingua italiana, di Salvatore Battaglia (poi diretto da Giorgio Bàrberi Squarotti), Torino, UTET, 1961-2002, 21 voll.; con *Supplemento 2004* e *Supplemento 2009*, diretti da Edoardo Sanguineti, Torino, UTET, 2004 e 2008, e *Indice degli autori citati nei volumi I-XXI e nel Supplemento 2004*, a cura di Giovanni Ronco, Torino, UTET, 2004.

Lenci A., Montemagni S:, Boschetti F., De Felice I., dei Rossi F., Dell'Orletta F., Di Giorgio M., Miliani M., Passaro L. C., Puddu A., Venturi G., Labanca N. (2020), *Voices of the Great War: A Richly Annotated Corpus of Italian Texts on the First World War*, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), Marsiglia, 11–16 maggio 2020, pp. 911-918.

LIF = Bortolini U., Tagliavini C., Zampolli A., *Lessico di frequenza della lingua italiana contemporanea*, IBM Italia 1971 (poi: Milano, Garzanti, 1972).

Marazzini C., Maconi L. (2018), *Il* Vocabolario dinamico dell'italiano moderno *rispetto ai linguaggi settoriali. Proposta di voce lessicografica per il redigendo VoDIM*, «Italiano digitale», VII/4, pp. 101-20.

Nivre J. (2015), *Towards a Universal Grammar for Natural Language Processing*, Computational Linguistics and Intelligent Text Processing -Proceedings of the 16th International Conference, CICLing 2015, Part I., Cairo, Egitto, pp. 3-16.

Pennacchiotti M., Zanzotto F.M. (2008). *Natural Language Processing Across Time: An Empirical Investigation on Italian*, Proceedings of GoTAL - 6th International Conference on Natural Language Processing, LNAI, volume 5221, pp. 371-382.

Picchi E. (2003), *PiSystem: sistemi integrati per l'analisi testuale*, in *Computational Linguistics in Pisa. Linguistica Computazionale*, a cura di Zampolli A., Calzorali N., Cignoni L., Special Issue, XVIII-XlX, Pisa-Roma, IEPI, 2003, pp. 597-627.

Sassolini E., Fahad Khan A., Biffi M., Monachini M., Montemagni S., *Converting and Structuring a Digital Historical Dictionary of Italian: A Case Study*, in Kosem, I., Zingano Kuhn, T., Correia, M., Ferreria, J. P., Jansen, M., Pe-reira, I., Kallas, J., Jakubíček, M., Krek, S. & Tiberius, C. (eds.), *Electronic lexicography in the 21st century: Smart lexicography*. Proceedings of the eLex 2019 conference (1-3 October 2019, Sintra, Portugal), Brno, Lexical Computing CZ, s.r.o., 2019, pp. 603-621 (pubblicazione elettronica: https://elex.link/elex2019/proceedings-download/).

Sprugnoli R., Passarotti M., Cecchini F. M., Pellegrini M. (2020), *Overview of the EvaLatin 2020 Evaluation Campaign*, Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages, LREC 2020, Marsiglia, Francia, pp. 105-110.

Straka M., Straková J. (2017), *Tokenizing, POS Tagging, Lemmatizing and Parsing UD2.0 with UDPipe*, Proceedings of the CoNLL 2017: Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, pp.88–99.