Francigena

11 (2025)

Metodi, sfide e prospettive per il trattamento automatico di varietà ibride medievali

Manuel Favaro
(Cnr-Istituto di Linguistica Computazionale “Antonio Zampolli” – Pisa)

Università degli Studi di Padova

Francigena *is an international peer-reviewed journal with an accompanying monograph series entitled "Quaderni di* Francigena*"*

ISSN 2420-9767

Dipartimento di Studi Linguistici e Letterari
Via E. Vendramini, 13
35137 PADOVA

info@francigena-unipd.com

INDICE

https://doi.org/10.25430/2420-9767/V11-008
DOI: 10.25430/2420-9767/V11-008

Metodi, sfide e prospettive per il trattamento automatico di varietà ibride medievali

Manuel Favaro
manuel.favaro@ilc.cnr.it

(Cnr-Istituto di Linguistica Computazionale "Antonio Zampolli" – Pisa)

ABSTRACT:
Il presente contributo si propone di riflettere attorno al trattamento automatico delle varietà medievali ibride. L'articolo si concentrerà dapprima sui problemi di processamento delle varietà non *standard* e diacroniche, poi sugli esperimenti computazionali sul 'francese d'Italia', di cui vengono presentati anche gli obiettivi finali e gli sviluppi in corso.

This paper aims to discuss *NLP* techniques on hybrid medieval varieties. The article will first focus on the problems of processing non-*standard* and diachronic varieties, then on computational experiments on the 'French of Italy'. Final aims and ongoing developments are also presented.

PAROLE CHIAVE: Varietà medievali ibride – varietà storiche – varietà non *standard* – *NLP* – francese d'Italia.
KEYWORDS: Hybrid Medieval Varieties – Historical Varieties – Non-*standard* Varieties – *NLP* – French of Italy.

1. *Introduzione*

Il trattamento automatico delle varietà linguistiche ibride di epoca medievale rappresenta un campo di ricerca affascinante, ma particolarmente complicato, poiché l'analisi computazionale di tali varietà pone numerose sfide, a più livelli. Le varietà ibride medievali, infatti, non solo riflettono la mancanza di standardizzazione linguistica dell'epoca, ma includono anche elementi lessicali, morfologici e sintattici altamente variabili, che spesso non vengono elaborati correttamente dai moderni strumenti di *Natural Language Processing* (*NLP*).

I testi appartenenti a varietà di questo genere testimoniano cambiamenti diacronici e sincronici che coinvolgono più dimensioni linguistiche. Dal punto di vista ortografico, presentano una varietà di grafie coesistenti, spesso dovute all'assenza di norme codificate; dal punto di vista morfologico, lessicale e semantico, offrono un repertorio ricco di forme che evidenziano un'evoluzione continua del significato e dell'uso delle parole. Inoltre, anche le strutture sintattiche e testuali sono spesso più complesse e meno regolari rispetto alle varietà linguistiche contemporanee.

Per questa ragione, si richiedono approcci innovativi che vadano oltre le tecniche utilizzate per l'elaborazione automatica delle lingue moderne. Questo articolo esplora le principali sfide che emergono nell'analisi computazionale di testi medievali, descrivendo i metodi attualmente disponibili per affrontarle e

discutendo le prospettive future per migliorare il trattamento automatico di queste testimonianze linguistiche. In particolare, l'articolo si sofferma sull'importanza della creazione di *corpora* annotati, sull'adattamento dei modelli *NLP* alle specificità dei testi storici e sull'adozione di approcci interdisciplinari che combinino linguistica, filologia e informatica.

2. *Varietà standard contemporanee* vs *varietà non standard storiche e diacroniche*

L'analisi e lo studio computazionale delle varietà ibride medievali rientrano in un panorama più ampio, che riguarda la differenza di processamento tra varietà linguistiche contemporanee e storiche, e al contempo tra varietà linguistiche *standard* e non-*standard*. In genere, infatti, gli strumenti *NLP* vengono addestrati su varietà *standard* e contemporanee, comportando numerosi ostacoli nel trattamento di tutte le altre varietà, in quanto gli strumenti sviluppati per le lingue moderne necessitano per forza di cose di specializzazioni su tutti i livelli linguistici per essere impiegati favorevolmente nell'elaborazione di fonti primarie, alla base degli studi umanistici. A tal proposito, è stato condotto più di un decennio fa uno studio esplorativo sulle difficoltà dell'elaborazione automatica delle varietà storiche, in particolare quelle riguardanti la lingua italiana[1]. Attraverso un *corpus* di testi letterari (in prosa e in poesia), composti tra il XIII e il XIX secolo, lo studio analizza la copertura lessicale dei testi in riferimento a risorse lessicografiche basate su *corpora* di testi moderni e su alcuni lessici morfologici generati automaticamente; tale copertura viene misurata in termini di percentuale di parole presenti. Secondo lo studio, in media, i testi diacronicamente marcati mostrano una copertura del 44%, inferiore del 19% rispetto ai testi giornalistici contemporanei, usati come riferimento.

Le difficoltà nel corretto trattamento delle forme lessicali possono derivare da diversi fattori, principalmente riguardanti la presenza di varianti che corrispondono per esempio a voci in disuso, forme dialettali, oppure tecnicismi e voci settoriali. A tal proposito, un altro lavoro[2], in relazione all'annotazione e all'analisi del *corpus Voci della Grande Guerra* (*VGG*), comprendente testi di diversi generi e di diversi registri linguistici, risalenti al periodo della Prima Guerra Mondiale, riporta alcuni esempi di tipi lessicali che hanno dato origine alle principali difficoltà nella lemmatizzazione da parte dei linguisti computazionali che se ne sono occupati, quali: forme di basso uso (ad es. *costì, tardanza*); regionalismi del tipo *cocuzza* e dialettalismi (p.e. *batajun*). In aggiunta, un'altra componente che gli autori registrano è la presenza di varianti sincroniche (del tipo *comperare* per

[1] Ci si riferisce al lavoro di Pennacchiotti – Zanzotto 2008.
[2] Cfr. De Felice (*et alii*) 2018.

comprare), che hanno causato delle difficoltà sia teoriche, sia applicative nelle scelte di lemmatizzazione del *corpus*.

Relativamente alla stessa raccolta testuale, un altro studio testimonia ulteriori difficoltà nel trattamento, questa volta però riguardanti il piano della sintassi[3]. Per esempio, il contributo rivela nel *corpus VGG* una maggiore lunghezza delle frasi, a differenza della media di testi contemporanei messi a confronto (25 *token* per i primi, 21 per i secondi), così come un maggiore ricorso alla subordinazione e agli ordini marcati, rispetto all'ordine basico soggetto-verbo-oggetto.

Riferito in parte allo stesso periodo temporale, il lavoro sul *corpus* per la creazione del *Vocabolario Dinamico dell'Italiano Moderno* (*VoDIM*), che raccoglie testi dell'italiano postunitario, mediante la costruzione e l'analisi di un *test corpus* per il riaddestramento dei modelli di POS-tagging e lemmatizzazione per l'italiano in diacronia si concentra anche sull'analisi degli errori di riconoscimento del lemma e della categoria grammaticale compiuti dai modelli presi in esame[4]. Nello specifico, per quanto riguarda l'analisi degli errori morfologici, nello studio emergono alcune difficoltà sistematiche nel trattamento di talune parti del discorso, in particolare i nomi propri (normalmente sovraestesi sugli altri *tag*, in quanto la presenza della maiuscola può creare delle ambiguità di attribuzione) e i verbi ausiliari, spesso assimilati ad altre tipologie di verbo. A proposito della morfologia verbale, il contributo accenna ad altre difficoltà tipiche del periodo postunitario, come la presenza di alternanze tematiche *chieggo/chiedo* e di variazioni flessive, per esempio nei participi passati, del tipo *concesso/conceduto*[5].

Allargando il periodo di riferimento, e toccando testi diacronicamente più marcati, compresi testi di origine medievale, emergono altri problemi nel trattamento automatico, quali per esempio il corretto riconoscimento delle forme comprendenti varianti ortografiche; difficoltà di riconoscimento che coinvolgono sia la fase di lemmatizzazione, sia quella di POS-tagging. In relazione all'ortografia, infatti, si notano le differenze maggiori e più evidenti tra varietà contemporanee e varietà storiche.

Tuttavia, i problemi non emergono soltanto nella corretta elaborazione delle varianti grafiche in diacronia e sulle grafie sincroniche coesistenti. Un primo esperimento sul *corpus* dei cosiddetti 'citati', cioè gli esempi usati per esemplificare le voci del *Grande Dizionario della Lingua Italiana* (*GDLI*), rivela infatti che le difficoltà sono legate non soltanto alla marcatezza diacronica dei testi, quanto anche all'appartenenza dei singoli testi a determinati generi e tipologie testuali e alla presenza di scelte stilistiche peculiari di alcuni autori[6]. L'analisi delle presta-

[3] Cfr. Lenci (*et alii*) 2020.
[4] Cfr. Favaro (*et alii*) 2023. Sul *corpus*, cfr. Marazzini – Maconi 2018.
[5] Sull'argomento si vedano per esempio i lavori di Mengaldo 1987 e Antonelli 2003.
[6] Cfr. Favaro (*et alii*) 2022.

zioni di due sotto *corpora* tratti dal *GDLI*, l'uno comprendente citazioni in prosa (dal *Convivio* alle opere di Vasco Pratolini), l'altro invece citazioni in versi (da Petrarca a Montale), mostra come le maggiori differenze di accuratezza tra il modello di base per la lemmatizzazione, addestrato su testi contemporanei, e quello riaddestrato sui testi tratti dal *GDLI* si registrano per le citazioni riferite a Leon Battista Alberti (sotto *corpus* prosa), a Francesco Petrarca e a Vittorio Alfieri (sotto *corpus* poesia). Solo per quanto riguarda Petrarca questa distanza tra i due modelli potrebbe essere spiegata in termini di fattori diacronici: l'analisi rivela che la maggior parte degli errori compiuti dai modelli di lemmatizzazione automatica riguarda varianti storiche, come l'alternarsi di forme geminate e degeminate (*abassare* vs *abbassare*), oppure la presenza di polimorfia verbale (*fuor* vs *furono*). Questi tipi di errori, tuttavia, sono meno numerosi nel modello riaddestrato (5, rispetto ai 18 del modello di base), ma comunque significativi, poiché rappresentano il 56% del numero totale di errori (nel modello di base la percentuale di errori era del 67%). Invece, per quanto riguarda Alberti e Alfieri, le difficoltà di annotazione riguardano più probabilmente altre caratteristiche correlate a scelte linguistiche e stilistiche peculiari. Ad esempio, le citazioni di Alberti sono caratterizzate da una grafia latineggiante, che rende quindi difficile l'elaborazione anche di parole funzionali elementari come la congiunzione *et* (in luogo di *e*) o varianti grafiche piuttosto banali come *adviato* vs *avviato*. Ciò che risulta più interessante è il peso che tali forme hanno sul computo degli errori: essi, infatti, coprono ben il 30% delle errate attribuzioni del lemma compiute dal modello di base, mentre questa percentuale scende al 12% in relazione al modello riaddestrato.

3. *Il trattamento di varietà medievali ibride: caratteristiche e sfide*

Le principali caratteristiche dei testi appartenenti a varietà medievali ibride, che possono mettere a rischio il corretto trattamento automatico da parte degli strumenti di elaborazione del linguaggio naturale, sono diverse e differentemente ostiche. In primo luogo, come più volte accennato, bisogna tener conto della presenza nei testi di elementi diacronicamente marcati, a partire dal livello grafico e grafico-fonetico, arrivando fino al lessico, alla semantica e alla sintassi. In secondo luogo, può essere decisiva la presenza di elementi stilisticamente connotati, determinati dalle scelte autoriali, in particolar modo nei testi letterari (cfr. par. 2). Inoltre, sono da considerare le differenze nella forma (prosa o poesia), non soltanto in relazione alla *facies* linguistica, ma anche dovute alla conformità dei modelli di annotazione, generalmente addestrati su testi in prosa. Come poi già accennato nel paragrafo precedente, hanno un peso non indifferente le differenze nella tipologia testuale di appartenenza dei singoli testi, che comportano di conseguenza scelte linguistiche connotate, e dunque un maggiore o minore adattamento degli strumenti di analisi *NLP*. A tutto ciò si aggiunge, infine, la presenza di elementi problematici a priori, ossia di tratti linguistici comuni, sia in

diacronia sia in sincronia, generalmente ostici nel trattamento, quali le enclitiche, spesso ipo- o ipersegmentate[7], oppure la presenza di poliрematiche, idiotismi di varia natura, ecc.

Tali caratteristiche, tuttavia, sono costitutive della maggior parte dei testi appartenenti a varietà diacroniche. Le varietà medievali ibride presentano però due fattori in più che rappresentano un'ulteriore sfida per il trattamento automatico. In primo luogo, naturalmente, la presenza di elementi ibridi, su cui pesa molto il diverso livello di mescidazione dei singoli testi, oltre al piano linguistico in cui avviene l'effettiva ibridazione. In secondo luogo, tali varietà costituiscono di per sé una sfida per l'universo *NLP*, in quanto generalmente non rappresentate o sotto-rappresentate nei corpora di addestramento: si tratta, dunque, di varietà che non mai state elaborate dai modelli, e di conseguenza dagli strumenti per il trattamento automatico.

4. *Possibili metodi e strategie*

Nell'ambito dell'elaborazione automatica di testi riferiti a varietà non *standard* e storiche, come anche le varietà ibride medievali, gli studiosi di ambito computazionale hanno adottato diverse strategie.

La prima è la normalizzazione ortografica. Si tratta di un approccio che mira a ridurre la distanza tra i dati di addestramento *standard* e i testi storici normalizzando l'input testuale[8]. La normalizzazione consiste nel mappare le grafie storiche a forme contemporanee, agendo come fase di preelaborazione prima dell'applicazione di strumenti *NLP*. Tuttavia, la normalizzazione ortografica è meno efficace a causa della coesistenza di altre forme di variazione linguistica, come quelle morfologiche, lessicali e strutturali.

La seconda è l'adattamento di dominio. Anziché rendere conformi i testi diacronicamente marcati agli strumenti esistenti, questo approccio adatta gli strumenti al linguaggio presente nei testi, basandosi sul presupposto che i modelli di *machine learning* tendono ad avere una capacità di successo nell'elaborazione minore quando i dati di *test* e di addestramento hanno distribuzioni diverse. L'adattamento di dominio, in questo contesto, comporta l'addestramento di modelli su dati di un dominio e la loro successiva applicazione a un dominio diverso ma correlato, come ad esempio una varietà storica di una specifica lingua[9]. L'adattamento di dominio può essere non supervisionato, quando si utilizzano dati non etichettati del dominio *target*, o supervisionato, quando si dispone di una

[7] Su cui si vedano gli esempi di De Felice (*et alii*) 2018: 161.
[8] Cfr. Bollmann 2019.
[9] Su cui si vedano gli esperimenti sulle varietà storiche di inglese di Yang – Eisenstein 2016.

piccola quantità di dati *target* etichettati, raggiungendo in alcuni casi dei risultati piuttosto favorevoli[10].

Un'ultima strategia riguarda l'utilizzo dei nuovi strumenti di intelligenza artificiale, i quali lavorano sui testi grezzi, senza quindi l'aggiunta preliminare di informazioni linguistiche. A tal proposito, uno degli esperimenti più interessanti è *BERToldo*, uno dei modelli basati sul modello di apprendimento automatico *BERT*, addestrato da zero a partire da dati provenienti da testi di varietà storiche dell'italiano[11]. Grazie a questo metodo, i diversi modelli di trasformatori sono riusciti a raggiungere alti livelli di accuratezza sul POS-tagging; in particolare, i dati sul *corpus D(h)ante*[12] mostrano un'accuratezza compresa tra il 93% e il 96%, a seconda delle diverse versioni create.

5. *Esperimenti sul 'francese d'Italia'*

Come accennato nel par. 3, mancano nel panorama *NLP* dei reali modelli di riferimento per il trattamento automatico delle varietà ibride del passato, e in particolar modo delle varietà ibride medievali.

Grazie al progetto *FringE* (*The French in/of Italy: Code-MixiNG in Medieval Europe*), che si propone di intraprendere una prima indagine sistematica sulle opere scritte in 'francese d'Italia'[13], è stato possibile impostare i primi esperimenti su testi appartenenti a varietà storicamente marcate e variabilmente ibridate, prendendo come riferimento il *corpus* costituito per il *Dizionario del Franco-Italiano* (*DiFrI*), nato in seno *RIALFrI* e che comprende, com'è noto, testi rappresentativi di un *continuum* che va da opere scritte in un francese pressoché perfetto, dove il peso del substrato italiano è scarso, irrilevante e talora nullo, a un contatto tra le due lingue che ha comportato in alcuni casi a varietà caratterizzate da una elevata mescidazione[14].

Per il *corpus* si è scelto di adottare in prima battuta il *software* di annotazione e di post-correzione *Pyrrha*, sviluppato dall'École Nationale des Chartes per l'analisi del francese antico[15].

[10] L'esperimento sul *VoDIM* del prima citato lavoro di Favaro (*et alii*) 2023 mostra come un approccio di questo tipo permette di guadagnare diversi punti percentuali sull'accuratezza dell'annotazione, in particolare sulla lemmatizzazione, che su alcuni testi ha raggiunto percentuali di accuratezza superiori al 99%.

[11] Cfr. Palmero Aprosio (*et alii*) 2022.

[12] Cfr. Basile – Sangati 2016.

[13] Secondo la definizione di Gambino – Beretta 2023.

[14] Sul progetto e sul *corpus*, si veda Gambino (*et alii*) 2024.

[15] Cfr. https://pyrrha.huma-num.fr/. I criteri che hanno condotto all'adozione di *Pyrrha*, rispetto al confronto con gli strumenti per l'analisi automatica del francese antico, sono esposti in Gambino (*et alii*) 2024: 301-302.

La scelta di un *software* come *Pyrrha*, progettato per l'analisi del francese antico, ma integrato con metodi di *machine learning* volti alla correzione degli errori automatici e al riaddestramento dei modelli di lemmatizzazione e di annotazione morfologica[16], ha condotto all'adozione di una strategia di adattamento di dominio supervisionato, tra le possibili esposte nel paragrafo precedente[17]. All'interno del progetto è stato dunque costituito un *test corpus*, vale a dire un campione testuale composto o da testi interi, o da piccole porzioni di testo lemmatizzati e annotati morfologicamente. Il campione testuale, di circa 55000 *token*, è stato interamente corretto a mano e costituito nel tentativo di mantenere un buon grado di bilanciamento e di rappresentatività del *corpus* complessivo[18], ossia scegliendo opere che potessero garantire la presenza sia di testi in prosa, sia di testi in versi, oltre a una differenziazione diacronica, diatopica e relativa alla tipologia testuale e al genere di appartenenza delle opere. In tale prospettiva, la raccolta testuale si configura come *gold standard* del francese d'Italia, vale a dire un punto di riferimento sia per la valutazione dei modelli di annotazione preesistenti, sia per l'addestramento computazionale di nuovi modelli[19].

La prima fase dell'esperimento è stata appunto la valutazione delle prestazioni dei modelli preesistenti di *Pyrrha*, le cosiddette *baseline*, cioè i modelli ancora non riaddestrati sui nuovi testi; la valutazione è stata svolta mettendo a confronto la versione dei testi generati automaticamente con quella annotata e manualmente rivista, cioè con i testi comprendenti appunto il *test corpus*, inteso come *gold standard* di riferimento (cfr. *supra*). I primi dati sull'accuratezza sono stati abbastanza incoraggianti: i modelli di lemmatizzazione e di POS-tagging hanno raggiunto in media rispettivamente l'86 e l'87%.

Tuttavia, analizzando più nel dettaglio i dati sulle prestazioni dei singoli testi, si nota che le divergenze notevoli sono dovute sia a peculiarità linguistiche, sia alle caratteristiche testuali delle opere selezionate. In merito a quest'ultimo punto, è stato inserito nel *test corpus* un documento, considerato come testo singolo, contenente la raccolta delle iscrizioni presenti nel *corpus* complessivo, al fine di sperimentare l'analisi automatica di tipologie testuali brevi e frammentarie[20]. In questo caso, la precisione della lemmatizzazione automatica è stata ben al di sotto

[16] Cfr. *ibid.*

[17] Al momento, non sono stati effettuati *test* utilizzando tecniche basate sulle *AI* e sui *LLM* (*Large Language Models*), in virtù del fatto che tecniche di questo genere hanno bisogno di una notevole mole di dati ben distribuiti per poter essere efficaci, e dunque di *corpora* molto più estesi e diversificati rispetto al *corpus DiFrI*.

[18] La dimensione attuale del *corpus* nella sua interezza, allo stato attuale, è di circa 4000000 di *token*; il campione di *test*, dunque, è stato allargato nelle successive fasi del progetto, al fine di migliorare la rappresentatività, in rapporto alla popolazione testuale complessiva (cfr. par. 6).

[19] Cfr. Gambino (*et alii*) 2024: 303-306.

[20] Nonostante si tratti di un insieme di testi, la dimensione totale di questa sotto-raccolta è molto esigua: i *token* sono poco più di 250.

della media: soltanto il 73% dei lemmi sono stati riconosciuti correttamente, a differenza invece del dato sul POS-tagging, che rientra invece perfettamente nel valore intermedio (86%).

Ciò che però si nota maggiormente è il fatto che le principali discrepanze nei dati sono prevalentemente dovute non tanto alla tipologia testuale di appartenenza, quanto più al livello di ibridismo dei testi. Infatti, i versi francesi del *Dittamondo* di Fazio degli Uberti, dunque per forza di cose più vicini alle varietà di riferimento di *Pyrrha*, addestrato su testi di francese antico, raggiungono un livello di accuratezza molto elevato, precisamente del 97% per la lemmatizzazione e del 96% per il riconoscimento delle categorie grammaticali. Al contrario, un testo maggiormente mescidato, colmo di italianismi, ossia il *Roland* del manoscritto V4 (cfr. *infra*), presenta un calo notevole nell'accuratezza: 75% per il riconoscimento dei lemmi, 79% per le categorie grammaticali.

A tal proposito, un altro elemento da considerare concerne gli indici relativi a una sottocategoria del POS-tagging, facente parte sempre dell'annotazione morfologica, ovverosia le etichette riguardanti il campo che in *Pyrrha* viene definito MORPH, comprendente le proprietà morfologiche associate alla specifica categoria grammaticale (il modo, il tempo, la persona, il numero ecc.). In questo caso, la media degli errori è molto più elevata rispetto al POS-tagging (l'accuratezza raggiunge in media solo il 74%), sia perché il numero di etichette disponibili è molto più ampio, e di conseguenza è più alta anche la probabilità che il modello compia errori, sia perché nel campo MORPH sono presenti proprio i *tag* relativi all'annotazione degli elementi ibridi.

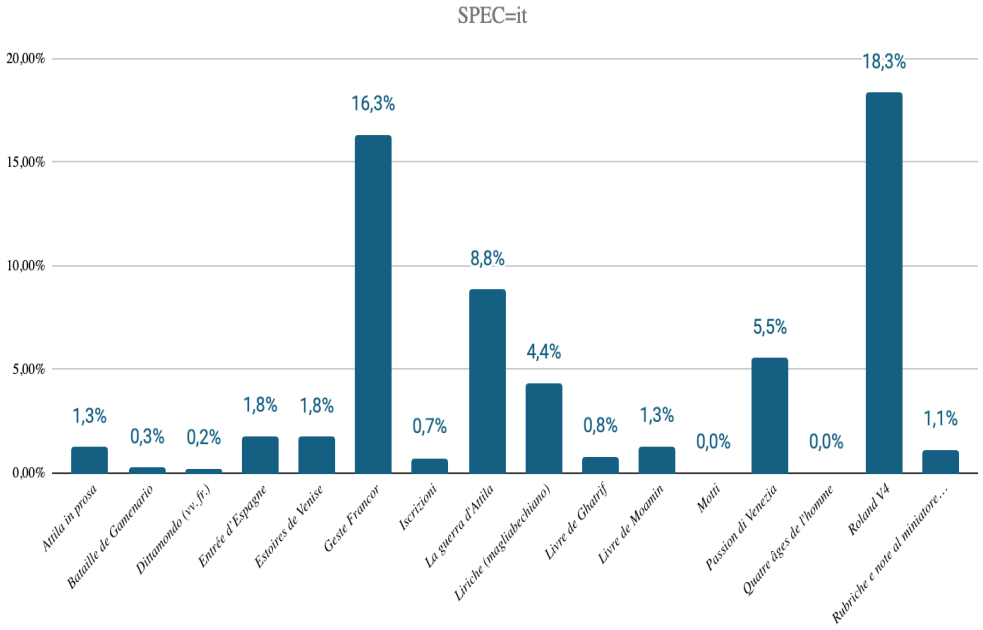
Infatti, il *test corpus* non si prefigura soltanto come punto di riferimento per la creazione di nuovi strumenti di analisi computazionale, ma anche come primo *corpus* interrogabile, su cui operare delle analisi preliminari sui fenomeni di contatto. Per questo motivo, la seconda fase del lavoro ha previsto l'estrazione e l'analisi quantitativa degli elementi ibridi, in particolar modo di tutti quei tratti riconoscibili in varia misura come 'italianismi'. Per raggiungere tale obiettivo, in fase di annotazione del *test corpus* sono state introdotte *ad hoc* alcune etichette che sono relative a specificità linguistiche e testuali; dunque, ancora non presenti nel novero dei *tag* previsti attualmente da *Pyrrha*. Si tratta delle etichette identificate come SPEC (che sta appunto per 'specificità'), che al momento riguardano le forme adattate per la rima, segnate come «SPEC=rim», mentre i latinismi sono etichettati «SPEC=lat». Inoltre, è stata scelta l'etichetta «SPEC=probl» che isola occorrenze linguisticamente o filologicamente problematiche[21].

Quella che interessa gli italianismi è invece l'etichetta «SPEC=it». Nell'ottica del progetto, l'obiettivo principale della taggatura 'italianismo' è quello di rilevare tratti devianti rispetto ad altri testi oitanici. Alcuni di questi tratti saranno sicura-

[21] Cfr. Gambino (*et alii*) 2024: 296-298.

mente di origine italiana, mentre altri potranno essere spiegati solo parzialmente con il contatto con l'italiano. Inoltre, dal punto di vista linguistico, oltre ai fatti grafici e grafico-fonetici, che rappresentano il piano linguistico attualmente meglio noto, l'annotazione permetterà, in fase di analisi sistematica, di cogliere tratti morfologici e lessicali eventualmente caratteristici.

Il grafico sottostante rappresenta la percentuale degli elementi etichettati come «SPEC=it» nei singoli testi del campione, in rapporto al numero totale dei *token*.



Come si osserva dal grafico, per la maggior parte dei campioni testuali analizzati l'uso dell'elemento italo-romanzo o presunto tale, almeno a livello meramente quantitativo, è piuttosto limitato. Vediamo però come per alcuni dei testi più rappresentativi della letteratura franco-italiana del XIV secolo la presenza diventa maggiormente significativa, e in particolare c'è un picco nella *Geste Francor* e nella *Chanson de Roland* del manoscritto identificato come V4. Notoriamente, il manoscritto V13 testimone della *Geste* presenta una *scripta* altamente influenzata da elementi italiani settentrionali[22]; anche il *Roland* V4 presenta molti tratti caratteristici del trevisano-bellunese. Tuttavia, oltre a darci numericamente conferme sulle ipotesi analitiche degli studi specifici, questa preliminare analisi quantitativa rivela un primo dato potenzialmente interessante. Sappiamo infatti che V4 è frutto di due redazioni distinte tratte da due diversi codici della *Chanson de Roland*, uno assonanzato e l'altro rimato. La sezione iniziale, assonanzata e più arcaica, con-

[22] Cfr. Mascitelli 2023.

tiene un grado superiore di italianismi rispetto alla sezione finale, rimata[23]. Altrettanto noto è che a fare da intermezzo c'è l'episodio della *Prise de Narbonne*, tratto da un terzo modello estraneo alla tradizione del poema. Questa cerniera è quella annotata all'interno del *test corpus*, grazie alla quale si potrà indagare il rapporto tra le sezioni a seconda anche del diverso grado di italianizzazione, che dipende in varia misura dalle vicende di trasmissione testuale.

Questo primo esperimento apre numerose prospettive, innanzitutto nel tentativo di fornire risposte sia sui tratti di 'sicura italianità', sia su quelli che per il momento sono considerati italianismi, come per esempio le forme dell'infinito del tipo *aver*, tratto caratteristico del franco-italiano, ma che compaiono pure nei testi del francese d'Oltremare[24].

Un fenomeno come quello appena menzionato, dunque, può trattarsi sia di un italianismo effettivo, sia di tratto poligenetico; soltanto a partire dall'analisi sistematica dei singoli fenomeni riscontrabili all'interno del *corpus* complessivo sarà possibile dare indicazioni su questa e su altre possibili domande di ricerca.

6. *Sviluppi in corso e conclusioni*

Le fasi successive del progetto *FringE* hanno previsto e prevedono, innanzitutto, l'allargamento del *test corpus*, sia nell'aggiunta di parti annotate dei testi già presenti, sia nell'inserimento di nuovi testi, mantenendo gli stessi criteri di bilanciamento e di rappresentatività esposti nel paragrafo precedente.

L'utilizzo di nuovi dati potrà permettere, come si è più volte menzionato all'interno del contributo, il riaddestramento dei modelli secondo la prospettiva di un adattamento alle specificità delle varietà prese in esame, sia specializzando i modelli sul francese antico di *Pyrrha*, sia riferendosi ad altri strumenti, in particolare quelli riguardanti la galassia delle *Universal Dependencies* (*UD*), attualmente lo *standard* di riferimento per l'annotazione linguistica[25]. Le ormai numerose banche dati annotate di *UD* raccolgono primariamente testi *standard* di lingue contemporanee, ma sono rappresentate anche alcune varietà non *standard* e varietà di lingue antiche (incluso l'antico francese), in particolare il *corpus PROFITEROLE*[26], contenente testi annotati di francese antico dal IX al XV secolo. A partire da questo *corpus* e dai modelli di trattamento automatico sviluppati su di esso, sono in corso degli esperimenti di riaddestramento dei modelli di lemmatizzazione e di POS-tagging, utilizzando gli strumenti cosiddetti '*UD*-

[23] Cfr. Beretta 2023.
[24] Cfr. Gambino (*et alii*) 2024: 297.
[25] Cfr. De Marneffe (*et alii*) 2021.
[26] Cfr. Prévost (*et alii*) 2024.

compliant' come *Stanza*[27], capace di raggiungere, mediante il processo di *training* basato su moduli a reti neurali, delle elevate percentuali di accuratezza. A tal proposito, i primi risultati sul riaddestramento dei modelli di lemmatizzazione (sono in corso quelli per il POS-tagging), sono stati piuttosto elevati: l'accuratezza media sui testi del campione ha raggiunto il 93,5%.

Si tratta, tuttavia, dei primi dati a disposizione. L'obiettivo finale sarà quello di applicare una sperimentazione dei modelli ottenuti sul *corpus* complessivo, al fine di ottenere una risorsa interrogabile da parte degli studiosi e, nel caso del *corpus* convertito secondo gli *standard UD*, interoperabile e riutilizzabile per ulteriori esperimenti su altre varietà.

Bibliografia

I. Manoscritti

V4	Venezia	Biblioteca Nazionale Marciana	francese	4
V13	Venezia	Biblioteca Nazionale Marciana	francese	13

II. Opere

Geste Francor

La *Geste Francor*, edition of the *Chansons de geste* of MS. Marc. Fr. XIII (=256), with glossary, introduction and notes by Leslie Zarker Morgan, Tempe, ACMRS, 2009 («Medieval & Renaissance texts & studies», 348).

Roland (V4)

The Franco-Italian Roland (V4), edited by Geoffrey Robertson-Mellor, Salford, University of Salford Reprographic Unit, 1980.

III. Studi e strumenti

Antonelli 2003

Giuseppe Antonelli, *Tipologia linguistica del genere epistolare nel primo Ottocento. Sondaggi sulle lettere familiari di mittenti cólti*, Roma, Edizioni dell'Ateneo, 2003.

[27] Cfr. Qi (*et alii*) 2020.

Basile – Sangati 2016
Angelo Basile, Federico Sangati, *D(h)ante: A new set of tools for XIII century Italian*, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, Portorož (Slovenia), European Language Resources Association, 2016, pp. 2825-2828.

Beretta 2023
Carlo Beretta, *La* Chanson de Roland *del ms. V4*, in Gambino – Beretta, pp. 3-30.

Bollmann 2019
Marcel Bollmann, *A large-scale comparison of historical text normalization systems*, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, Association for Computational Linguistics, 2019, pp. 3885-3898.

De Felice (*et alii*) 2018
Irene De Felice, Felice Dell'Orletta, Giulia Venturi, Alessandro Lenci, Simonetta Montemagni, *Italian in the Trenches: Linguistic Annotation and Analysis of Text of the Great War*, in *Proceedings of 5th Italian Conference on Computational Linguistics (CLiC-It)*, edited by Elena Cabrio, Alessandro Mazzei, Fabio Tamburini, Torino, Accademia University Press, 2018, pp. 160-164.

De Marneffe (*et alii*) 2021
Marie-Catherine De Marneffe, Christopher D. Manning, Joakim Nivre, Daniel Zeman, *Universal Dependencies*, in «Computational Linguistics», 47/2 (2021), pp. 255-308.

Favaro (*et alii*) 2022
Manuel Favaro, Elisa Guadagnini, Eva Sassolini, Marco Biffi, Simonetta Montemagni, *Towards the Creation of a Diachronic Corpus for Italian: A Case Study on the GDLI Quotations*, in *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA 2022)*, edited by Rachele Sprugnoli, Marco Passarotti, Marseille, European Language Resources Association (ELRA), 2022, pp. 94-100, https://aclanthology.org/2022.lt4hala-1.0 [cons. 7. VII. 2024].

Favaro (*et alii*) 2023
Manuel Favaro, Marco Biffi, Simonetta Montemagni, *POS Tagging and Lemmatization of Historical Varieties of Languages. The Challenge of Old Italian*, in «Italian Journal of Computational Linguistics», 9/2 (2023), http://journals.openedition.org/ijcol/1325 [cons. 31. I. 2025].

Gambino (*et alii*) 2024
Francesca Gambino, Andrea Beretta, Maura Sonia Barillari, Floriana Ceresato, Giacomo Costa, Rachele Fassanelli, Manuel Favaro, Jacopo Fois, Elisa Guadagnini, Federico Guariglia, Matteo Parodi, Carlo Rettore, *Il 'francese d'Italia' e il progetto "FrIngE". Panoramica generale e casi di studio*, in «Francigena», 10 (2024), pp. 285-340.

Gambino – Beretta 2023
Antologia del francese d'Italia. XIII-XV secolo, a cura di Francesca Gambino e Andrea Beretta, Bologna, Pàtron, 2023 («Storia e testi», 4).

GDLI
Grande dizionario della lingua italiana, di Salvatore Battaglia (poi diretto da Giorgio Bárberi Squarotti), 21 voll., Torino, UTET, 1961-2002; 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, https://www.gdli.it/ [cons. 3. III. 2025].

Lenci (*et alii*) 2020
Alessandro Lenci, Simonetta Montemagni, Federico Boschetti, Irene De Felice, Stefano Dei Rossi, Felice Dell'Orletta, Michele Di Giorgio, *Voices of the Great War: A Richly Annotated Corpus of Italian Texts on the First World War,* in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, European Language Resources Association, 2020, pp. 911-918.

Mascitelli 2023
Cesare Mascitelli, *Anonimo,* Geste Francor, in Gambino – Beretta, pp. 53-68.

Marazzini – Maconi 2018
Claudio Marazzini, Ludovica Maconi, *Il 'Vocabolario dinamico dell'italiano moderno' rispetto ai linguaggi settoriali. Proposta di voce lessicografica per il redigendo VoDIM*, in «Italiano digitale», 7 (2018), pp. 101-120.

Mengaldo 1987
Pier Vincenzo Mengaldo, *L'epistolario di Nievo. Un'analisi linguistica*, Bologna, Il Mulino, 1987.

Palmero Aprosio (*et alii*) 2022
Alessio Palmero Aprosio, Stefano Menini, Sara Tonelli, *BERToldo, the historical BERT for Italian,* in *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, Marseille, European Language Resources Association, 2022, pp. 68-72.

Pennacchiotti – Zanzotto 2008
Marco Pennacchiotti, Fabio M. Zanzotto, *Natural language processing across time: An empirical investigation on Italian*, in *Proceedings of GoTAL – 6th International Conference on Natural Language Processing*, edited by Bengt Nordström, Aarne Ranta, Berlin-Heidelberg, Springer-Verlag, 2008, pp. 371-382.

Prévost (*et alii*) 2024
Sophie Prévost, Loïc Grobol, Mathieu Dehouck, Alexei Lavrentiev et Serge Heiden, *Profiterole: un corpus morpho-syntaxique et syntaxique de français médiéval*, in «Corpus», 25 (2024), https://doi.org/10.4000/corpus.8538 [cons. 3. III. 2025].

Pyrrha
Pyrrha, A language independent post correction app for POS and lemmatization, https://pyrrha.huma-num.fr/ [cons. 3. III. 2025].

Qi (*et alii*) 2020
Peng Qi, Zhang Yuhao, Zhang Yuhui, Jason Bolton, Christopher D. Manning, *Stanza: a Python Natural Language Processing toolkit for many human languages*, in *ACL2020 System Demonstration*, Online, 2020, https://aclanthology.org/2020.acl-demos.14/ [cons. 31. I. 2025].

RIALFrI
Repertorio Informatizzato Antica Letteratura Franco-Italiana (*RIALFrI*), diretto da Francesca Gambino, Università degli Studi di Padova, Dipartimento di Studi Linguistici e Letterari, versione 2.0, 2022, www.rialfri.eu [cons. 3. III. 2025].

Yang – Eisenstein 2016
Yi Yang, Jacob Eisenstein, *Part-of-Speech tagging for historical English*, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kevin Knight, Ani Nenkova, and Owen Rambow, San Diego, Association for Computational Linguistics, 2016, pp. 1318-1328.