Francigena

7 (2021)

Strumenti e criteri per la lemmatizzazione del franco-italiano: verso la costruzione di un *corpus* lemmatizzato della *Geste Francor*

Sira Rodeghiero
(Università degli Studi di Padova)

Università degli Studi di Padova

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

ISSN 2724-0975

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

info@francigena-unipd.com

INDICE

https://doi.org/10.25430/2420-9767/V7-008
DOI: 10.25430/2420-9767/V7-008

In ricordo di Simon Gaunt

Strumenti e criteri per la lemmatizzazione del franco-italiano: verso la costruzione di un *corpus* lemmatizzato della *Geste Francor*

Sira Rodeghiero
sira.rodeghiero@unipd.it

(Università degli Studi di Padova)

Abstract:
L'articolo discute strumenti e criteri per la costruzione di un *corpus* lemmatizzato di franco-italiano, presentando i risultati di una lemmatizzazione sperimentale delle *Enfances Bovo* della *Geste Francor.*

This article discusses tools and criteria for the construction of a lemmatized corpus of franco-italian texts by presenting the results of an experimental lemmatization of the *Enfances Bovo* of the *Geste Francor*.

Keywords:
Lemmatization – franco-italian – lemmatized corpus - *Geste Francor*
Lemmatizzazione – franco-italiano – *corpus* lemmatizzato – *Geste Francor*

1. *Introduzione*

La costruzione di un dizionario specificamente dedicato al franco-italiano rappresenta da oltre mezzo secolo una delle principali ambizioni degli studiosi del settore[1]. Il lessico costituisce, infatti, uno dei campi di indagine più complessi di questa singolare lingua letteraria e poterne acquisire una conoscenza quanto più possibile ampia e approfondita è fondamentale per la corretta interpretazione dei testi e per la loro descrizione linguistica. Molti sono però gli ostacoli che si frappongono alla realizzazione di una simile opera. La prima difficoltà è di ordine ontologico: esiste una 'lingua' franco-italiana? La questione sulla natura sostanziale o accidentale del franco-italiano[2] grava enormemente sulla progettazione di un suo dizionario e rende estremamente complessa la selezione del *corpus* così come la definizione dell'interazione dei sistemi linguistici coinvolti[3]. D'altro canto, abbondano anche difficoltà specifiche: la cospicua dose di *hapax*, le forme rare prodotte dall'interferenza linguistica, l'ardua, talvolta impossibile, ricostruzione etimologica di molte parole[4] concorrono a delineare un quadro piuttosto scoraggiante.

[1] Cfr. Holtus 1981: 155, Capusso 2007: 173, Morlino 2010: 65.
[2] Per una sintesi sulla questione si veda Barbato 2015.
[3] Cfr. Gambino 2020.
[4] Cfr. Morlino 2010: 65.

Eppure, le strategie per affrontare l'impresa non mancano, purché si accetti di adeguare in parte la propria mentalità alle risorse disponibili. In un mondo sempre più proteso alla digitalizzazione, queste possono essere individuate innanzitutto nei mezzi informatici, soprattutto se intesi non solo come strumento di produzione, ma anche come mezzo di pubblicazione e fruizione dei risultati. Con ciò non si intende che il digitale consenta di superare i problemi e le insormontabili difficoltà inerenti al franco-italiano. Esso può però perlomeno alleviarne alcune criticità, favorendo l'adozione di una logica di lavoro incrementale e aperta al dibattito *in fieri*, particolarmente vantaggiosa per questo tipo di lavoro. Il supporto digitale, infatti, consente di ridurre i tempi di pubblicazione dei risultati e di renderli immediatamente fruibili in linea, offrendoli al dibattito della comunità scientifica gradualmente, di pari passo con la crescita del lavoro. Ne consegue una rapida ed efficace integrazione di commenti, segnalazioni, migliorie e modifiche a tutto vantaggio della qualità del prodotto e senza gli oneri della pubblicazione cartacea. Del resto, la possibilità di automatizzare alcune fasi di lavoro e la conseguente necessità di criteri di analisi oggettivi e uniformi stimolano a scelte e classificazioni che aiutano a rendere più governabili situazioni apparentemente inestricabili.

Insomma, il supporto digitale sembra poter costituire un punto di forza nella creazione di un dizionario di franco-italiano, opera che coinvolge un oggetto di studio tanto complesso e aggrovigliato da richiedere gradualità nel lavoro, apertura ad un'ampia discussione scientifica, definizione di criteri di analisi e che, più che mai, necessita di continua revisione e modifica.

Questa è precisamente la linea su cui si muove il *DiFrI* (*Dizionario del Franco-italiano)*, un progetto diretto da Francesca Gambino, e nato in seno al *RIALFrI*, che si propone di creare un dizionario completo del franco-italiano a partire da un repertorio digitalizzato e interrogabile di testi della letteratura franco-italiana[5]. Oggetto di indagine sono in particolare le opere originali scritte in francese da autori italiani, un insieme di testi linguisticamente eterogeneo, che copre un arco cronologico che va dal secondo quarto del Duecento fino alla fine del Quattrocento[6]. Attraverso la creazione di un dizionario del franco-italiano, il progetto *DiFrI* mira ad offrire uno strumento in grado di evidenziare non solo la differenziazione lessicale e semantica dei testi studiati rispetto al francese, ma anche la loro componente comune, così da contribuire alla descrizione di questa particolare lingua letteraria, ricostruendone il processo di elaborazione e le sue dinamiche variazionali[7]. Una simile opera presuppone naturalmente un ampio lavoro di lemmatizzazione dei testi, che richiede tempo e implementazione, ma che ha già conosciuto i primi risultati nella lemmatizzazione di parte di due testi esemplari

[5] Cfr. Gambino 2020.
[6] *Ibid*.
[7] *Ibid*.

della letteratura franco-italiana: l'*Entrée d'Espagne* (a cura di Floriana Ceresato) e le *Enfances Bovo* della *Geste Francor*.

Il presente articolo si concentrerà proprio sulle *Enfances Bovo*, illustrando gli strumenti e i criteri impiegati per la sua lemmatizzazione, al fine di testarne l'efficacia e di contribuire a tracciare alcune linee guida valide per la lemmatizzazione degli altri testi inclusi nel *DiFrI*.

2. *Un* corpus *digitale lemmatizzato del franco-italiano*

La lemmatizzazione consiste nel ricondurre tutte le forme flesse e le diverse realizzazioni grafiche di una parola ad un'unica forma canonica, individuata come entrata di dizionario. Così, ricorrendo ad un esempio già impiegato da Roberto Busa[8], la forma flessiva *uomini* e la variante grafica *omo* vengono attribuiti ad uno stesso esponente lessicale *uomo.* Quest'ultima forma base è detta *lemma* e le diverse varianti (flessive e grafiche) sono invece le *forme* del lemma.

Tale operazione di raggruppamento di diverse forme sotto un unico lemma può essere svolta in modo semi-automatico grazie all'impiego di elaboratori elettronici. Il risultato è la creazione di un *corpus* digitale lemmatizzato con relativi vantaggi in termini di velocizzazione del processo, di quantità di dati trattati e immagazzinati, nonché di ottimizzazione della loro esplorazione.

Alla base della realizzazione di un simile prodotto vi è il concetto di *corpus* come collezione di dati selezionati e organizzati secondo criteri funzionali al tipo di analisi e alle informazioni che da tali dati si intende ricavare. Pertanto, la costruzione di un *corpus* elettronico prevede alcuni passaggi preliminari fondamentali, ovvero:

1) l'individuazione degli elementi costitutivi della raccolta;
2) la definizione degli scopi del *corpus*;
3) la scelta degli strumenti[9].

La creazione di un *corpus* digitale, infatti, dipende innanzitutto dalla precisa definizione dei dati che si intendono destinare al trattamento della macchina e dalla corretta e completa individuazione delle relazioni e dei vincoli che essi intrattengono. Ciò consente di elaborare un primo progetto astratto del *corpus*, in grado di determinarne le caratteristiche e l'articolazione logica prima ancora della sua concreta realizzazione e indipendentemente dalla macchina. La consapevolezza degli obiettivi del *corpus*, vale a dire delle informazioni che si intendono

[8] Busa 1987: 113.
[9] Cfr. Tomasi 2008: 84.

estrarre dalla raccolta, è d'altro canto imprescindibile per stabilire i criteri di rappresentazione informatica dei dati e le modalità di interrogazione[10]. A ciò si aggiunge, infine, la necessità di decidere, in coerenza con il modello elaborato, quali linguaggi e quali programmi impiegare per la gestione del processo di creazione, manipolazione, pubblicazione e fruizione del *corpus*[11]. Quanto più consapevole e precisa è la riflessione su ciascuno dei punti menzionati tanto maggiori sono le possibilità di successo del *corpus.*

Applicato alla creazione di un *corpus* lemmatizzato, quanto detto sopra evidenzia la necessità di considerare accuratamente le forme presenti nei testi inclusi nella collezione e i problemi che esse possono presentare, e suggerisce di individuare in anticipo un elenco di lemmi a cui ricondurre tali forme, prevedendo eventuali difficoltà di associazione. Anche l'inserimento di eventuali altre informazioni (ad esempio annotazioni di natura morfo-sintattica) va progettato fin dal principio. Lo scopo del *corpus*, che è quello di associare forme a lemmi e ad altre eventuali informazioni, chiarisce anche la struttura logica di organizzazione dei dati, che è di tipo relazionale, ovvero rappresentabile con una struttura a tabella che assegni valori a determinati attributi (o campi). Quanto alle modalità di interrogazione, le finalità del *corpus* definiscono perlomeno due possibilità, vale a dire una ricerca per forme e una ricerca per lemmi, agevolate dalla scelta di un'adeguata interfaccia utente.

2.1 *Progettare un* corpus *lemmatizzato del franco-italiano*: *problemi e obiettivi*

Il franco-italiano è rappresentato da un insieme di testi redatti tra il XIII e il XV secolo in una forma linguistica che combina variamente tratti francesi e italiani. Tale mescolanza si manifesta in un'interferenza tra sistemi linguistici che ha effetti pervasivi a vari livelli: fonetico, morfologico e semantico. Peculiarità del franco-italiano è, dunque, la vasta produzione di ibridismi. In qualche occasione, però, la combinazione tra il sistema del francese e dell'italiano può assumere anche caratteristiche analoghe al *code-mixing*[12] attraverso un «uso alternato delle due lingue per porzioni sintattiche»[13]. In ogni caso, l'uso dei termini 'francese' e 'italiano' non deve far dimenticare che né l'uno né l'altro rappresentano realtà omogenee. Ciò va considerato soprattutto in riferimento alla componente linguistica peninsulare. Benché esistano elementi panitaliani, gli italianismi inglobano per lo più tratti locali e regionali propri dei volgari diffusi in alcune aree (specialmente settentrionali) nella penisola italiana. Si delinea così un

[10] Cfr. Tomasi 2008: 84.
[11] *Ibid.*
[12] A tal proposito si veda, ad esempio, lo studio proposto da Gambino 2016 per il *Bovo udinese*.
[13] Renzi 1970: 85.

panorama estremamente variegato di cui è difficile fornire una descrizione linguistica. Alcune caratteristiche comuni possono certamente emergere da opportune classificazioni tassonomiche, come da quella proposta da Barbato[14] che, in base a parametri genetici e linguistici, individua testi di origine francese e testi di origine italiana, a propria volta suddivisibili in produzioni a fondo linguistico francese o italiano. Tuttavia, va tenuto presente che il franco-italiano costituisce «un *continuum* che va dal francese all'italiano, in cui ogni testo è un caso a sé»[15], quale manifestazione linguistica unica, dotata di caratteristiche proprie[16].

La realtà rappresentata dal franco-italiano è, dunque, piuttosto complessa e rende la progettazione di un *corpus* lemmatizzato particolarmente delicata. Le problematiche connesse ad una lingua medievale così mescidata, infatti, non sono poche né di facile soluzione. Non si tratta di individuare solamente i testi da includere nella selezione e di affidarne la trattazione automatica ad un programma già esistente che traduca in rappresentazione informatica il modello di *corpus* desiderato. Occorre piuttosto avere consapevolezza della natura dell'oggetto analizzato e prevedere le difficoltà che possono manifestarsi in corso d'opera.

In questa prospettiva, un primo problema è già individuabile nella natura dei dati. La presenza di forme ibride pone la questione della scelta del tipo di lemmi da includere nel *corpus*: si considerano lemmi solo francesi? Devono essere inclusi anche lemmi italiani[17]? Va previsto l'inserimento anche di lemmi propri di altre varietà dialettali antiche? La difficoltà va però ben oltre e implica anche la relazione tra i dati: una forma ibrida va associata ad un solo lemma (es. francese) o a più lemmi (es. francese e italiano)? Inoltre, se si desidera corredare la lemmatizzazione con informazioni di natura morfo-sintattica, il problema si espande ulteriormente: come gestire i dati nel caso di forme grammaticali anomale?

Cruciale è stabilire preventivamente le finalità del *corpus*: si vuole esaminare solo lo scarto lessicale e semantico del franco-italiano rispetto al francese o interessano anche i punti di convergenza? Ci si vuole limitare allo studio del lessico? Si contempla una futura espansione allo studio di altri aspetti linguistici? Fino a che punto interessa estrarre informazioni relative alla mescolanza linguistica, ovvero, premessa l'enormità della questione, con quale grado di approssimazione possono essere rappresentate tali informazioni?

Anche la scelta degli strumenti richiede particolare cura. Esistono, infatti, diversi programmi di lemmatizzazione automatica, ma non tutti efficaci per il trattamento di lingue a forte variazione grafica e poco regolari nell'applicazione

[14] Barbato 2015: 46-47.
[15] Ivi: 36.
[16] Si vedano in proposito Rajna 1998: 200-201, Viscardi 1941: 44-47, Vidossi 1956: LXIX, Wunderli 2003: 1.
[17] Si precisa che nel presente articolo si intende l'italiano antico secondo i termini stabiliti da Salvi – Renzi 2010.

delle norme morfo-sintattiche. Inoltre, per quanto performante sia il programma, è chiaro che l'*output* della lemmatizzazione automatica di una lingua ricca di forme aberranti come il franco-italiano presenterà sempre un margine di errore piuttosto ampio. Poter disporre di un'interfaccia che renda agevole la correzione manuale dei risultati diventa pertanto una reale necessità. Per nulla secondario, infine, è lo strumento di interrogazione e rappresentazione dei dati, che deve tenere conto di tutte le difficoltà già menzionate ed essere funzionale agli scopi del *corpus*.

Come si nota, i problemi sollevati dalla costruzione di un *corpus* lemmatizzato di franco-italiano sono dunque molteplici ed estremamente condizionanti. Prendere decisioni preliminari in linea con i propri obiettivi e progettare con lungimiranza il lavoro costituisce una necessità fondamentale. D'altro canto, però, la singolarità di ogni testo spesso impedisce di prevedere e proporre soluzioni per tutti i problemi che si potrebbero presentare. Alcuni aspetti emergeranno inevitabilmente in corso d'opera, imponendo una revisione e un raffinamento tanto dell'organizzazione e delle modalità di lavoro quanto dei mezzi impiegati. In tale prospettiva, si comprende perciò come la costruzione di un *corpus* lemmatizzato del franco-italiano non possa prescindere da una prima fase di sperimentazione. Di qui deriva allora il carattere, per così dire, pionieristico della lemmatizzazione dei primi testi del *DiFrI*, che mira proprio a collaudare gli strumenti di lavoro e a elaborare un sistema di gestione e trattamento dei dati quanto più possibile funzionale allo scopo del *corpus*: descrivere il processo di elaborazione del franco-italiano nella sua differenziazione e convergenza rispetto al francese, tenendo conto delle diverse dinamiche di interazione con i volgari italiani non solo dal punto di vista semantico, ma anche a livello morfologico e grafico-fonetico[18].

Con questo spirito, i prossimi paragrafi presenteranno dunque le principali fasi di lavoro e illustreranno le risorse e gli strumenti adottati per la costruzione e rappresentazione del *corpus* (cfr. § 3), onde descrivere, sulla base di tali premesse, l'esperienza di lemmatizzazione delle *Enfances Bovo* (cfr. § 4) e l'apporto che essa può offrire nella definizione dei criteri di annotazione e nell'implementazione degli strumenti impiegati (cfr. § 5).

3. *La lemmatizzazione nell'ambito del progetto* DiFrI

L'approccio del *DiFrI* alla lemmatizzazione del franco-italiano tiene conto dell'estrema complessità di tale operazione e si muove all'insegna della gradualità e della prudenza, cercando di sfruttare al meglio le risorse disponibili. La consapevolezza di quanto l'oggetto di indagine sia difficile da governare induce ad

[18] Cfr. Gambino 2020.

accettare un certo grado di approssimazione iniziale. Non vi è presunzione di offrire immediatamente un prodotto perfetto. La filosofia di base è piuttosto quella di un progressivo avvicinamento al problema che, sgrezzando e raffinando via via strumenti e metodi di lavoro, renda affrontabili questioni in prima battuta apparentemente inestricabili.

Poche scelte hanno dunque guidato l'avvio dei lavori: ricondurre le forme quanto più possibile al lemma francese, ricorrendo ad uno strumento di lemmatizzazione automatica; fornire un'indicazione inizialmente solo sommaria dell'interferenza linguistica per le forme ibride; pubblicare i risultati sul sito del *RIALFrI*, offrendo uno strumento di interrogazione per forme e per lemmi.

Nel complesso, la lemmatizzazione digitale dei testi considerati per il *DiFrI* prevede quindi cinque fasi di lavoro:

1. La digitalizzazione dei testi
2. La lemmatizzazione automatica
3. La correzione manuale dei risultati della lemmatizzazione automatica
4. L'esportazione dei risultati corretti
5. La pubblicazione dei risultati perché possano essere interrogati.

3.1 *La prima fase: la digitalizzazione del testo*

La prima fase di lavoro è preparatoria e consiste nel predisporre il testo affinché esso possa essere analizzato dalla macchina. Occorre dunque innanzitutto scegliere un'edizione di riferimento e trasformarne il testo in formato digitale. Nel caso del *DiFrI* tale operazione è semplificata in quanto può avvalersi dei testi già selezionati e digitalizzati in *RIALFrI*. Quindi, il testo digitalizzato deve essere manipolato per permettere alla macchina di riconoscere le unità di analisi e di eseguire le operazioni previste. L'attenzione, durante questa fase, deve essere duplice: non perdere informazioni testuali e andare incontro alle esigenze di elaborazione automatica. Le prime operazioni da eseguire in tal senso sono affidate per il *DiFrI* a Luigi Tessarolo, che assiste il progetto per gli aspetti informatici. Esse prevedono l'eliminazione di tutto ciò che può interferire con il linguaggio di programmazione, come, ad esempio, la presenza di parentesi quadre, e alcuni interventi dettati dalla necessità di mantenere informazioni essenziali. L'aggiunta del simbolo di paragrafo (§), ad esempio, è indispensabile per conservare l'indicazione di fine verso.

Un aspetto cruciale in questa fase è il concetto di parola, ovvero l'unità di analisi, che per l'elaboratore elettronico non può essere altro che una stringa di caratteri compresa tra spazi bianchi. Ciò implica un rimaneggiamento della punteggiatura e dell'elisione al fine di separare le parole da segni di interpunzione e apostrofi. Accanto a questi interventi meramente tecnici, ve ne sono però altri che richiedono necessariamente il coinvolgimento dello studioso. Tra questi vi è quello di considerare la possibilità di spezzare alcune parole altrimenti non analizzabili dalla macchina. Il caso tipico è rappresentato dalle preposizioni articolate e da alcuni fenomeni di contrazione (specialmente pronominale), ma vi

sono anche altre forme composte che richiedono una considerazione più scrupolosa (cfr. § 5.1). Alcune di queste scelte dipendono dal programma che si intende utilizzare. Pertanto occorre dedicare qualche parola al *software* scelto per la lemmatizzazione del franco-italiano: *Pyrrha*.

3.2 *Le fasi centrali: l'annotazione automatica con* Pyrrha *e la correzione manuale dei risultati*

Pyrrha è un'interfaccia di annotazione e post-correzione sviluppata dall'É*cole nationale des Chartes*[19]. Essa si appoggia ad un programma di annotazione Pie[20] che consente di associare ogni forma ad un lemma, ad una categoria del discorso e ad una descrizione morfo-sintattica, basandosi su algoritmi che permettono alla macchina di perfezionare progressivamente l'annotazione, confrontando i risultati prodotti automaticamente con quelli corretti manualmente dallo studioso[21]. Non c'è dunque bisogno di una grammatica predefinita in base alla quale stabilire le regole algoritmiche di annotazione. Si tratta piuttosto di un programma in grado di 'apprendere' la lingua alla quale si applica (*machine learning*)[22]. Ciò risulta particolarmente efficace per il trattamento di lingue poco omogenee nella realizzazione grafica e nell'applicazione di regole morfosintattiche[23], quale appunto il franco-italiano.

Per la lemmatizzazione del franco-italiano, il *DiFrI* si appoggia al modello originariamente elaborato da *Pyrrha* per il francese antico[24], che si basa sul dizionario *Tobler-Lommatzsch* e sul protocollo di annotazione *Cattex09*[25] sviluppato dall'università di Lione.

Nel complesso, *Pyrrha* permette, dunque, di lemmatizzare un testo arricchendolo anche di una descrizione morfo-sintattica, di verificare e correggere i risultati dell'annotazione automatica e di esportarli (in formato CSV o XML-TEI) perché possano essere conservati e interrogati.

Pyrrha è liberamente accessibile *online*, previa registrazione. Accedendo con il proprio *account* all'interfaccia è possibile creare un proprio *corpus* e importare il testo che si intende lemmatizzare, dopo averlo opportunamente predisposto (cfr. § 3.1). Dalla pagina di creazione del *corpus* si seleziona un modello di riferimento (es. francese antico) e la lista di controllo inclusiva di tutti i valori utilizzabili per l'annotazione. In questa fase, è possibile definire anche altre opzioni, come la quantità di contesto che si intende visualizzare attorno ad ogni forma e il *layout*

[19] Clérice – Pilla – Camps 2019.
[20] Manjavacas – Kestemont – Clérice 2018.
[21] Cfr. Pinche 2019: 50.
[22] *Ibid.*
[23] *Ibid*.
[24] Clérice – Camps 2021.
[25] Guillot – Prévost – Lavrentiev 2013a e 2013b.

del *corpus*, decidendo, ad esempio, se si desidera la rappresentazione completa di tutti i risultati (lemmi, parte del discorso e descrizione morfologica) o solo di alcuni di essi. Quindi, inviando le proprie scelte attraverso il bottone *Submit*, viene prodotta la lemmatizzazione automatica. I risultati, distribuiti su più pagine, appaiono all'interno di una struttura a tabella (cfr. §2). Ogni forma occupa una riga e viene associata ad un valore per ciascuno dei campi inclusi nella rappresentazione, come nell'esempio seguente:

Id	**Form**	**Lemma**	**POS**	**Morph**	**Context**	**Similar**	**Save**	**+**
34	*les*	*le*	DETdef	NOMB.=p|GENRE=m|CAS=r	Sor tot **les** autres fu de major renon	30	Save	+

L'utente, a questo punto, può procedere alla correzione manuale modificando i valori errati nei campi dedicati all'analisi (*Lemma*, *POS*, *Morph*). La scelta deve essere confermata dal salvataggio (funzione *Save)*. Se il valore immesso non è incluso nella lista di controllo selezionata al momento della creazione del *corpus*, non sarà possibile salvare la modifica.

3.3 *Valori ammessi e principi generali di annotazione con Pyrrha*

Come già ribadito, i valori che possono essere impiegati per l'annotazione costituiscono un *set* prestabilito e fissato in una lista di controllo. Per poter meglio comprendere i binari entro cui si muove la lemmatizzazione delle *Enfances Bovo* e il contributo che essa può offrire per la messa a punto degli strumenti di lavoro, non sarà dunque inutile presentare brevemente i valori a disposizione per l'annotazione del *corpus* di franco-italiano e i principi che ne guidano l'utilizzo.

Quanto ai lemmi, si è già detto che il *DiFrI*, appoggiandosi al modello creato da *Pyrrha* per il francese antico, può disporre di un elenco di lemmi esclusivamente francesi definito sulla base del dizionario *Tobler-Lommatzsch*[26].

Per quanto concerne la descrizione morfo-sintattica, *Pyrrha* adotta invece il sistema di annotazione del *Cattex09*, che prevede un *set* di etichette per l'analisi delle parti del discorso (*Part of Speech*, d'ora in poi *POS*) e un insieme di etichette per le informazioni morfologiche da inserire nel campo *Morph* di *Pyrrha*.

Le etichette *POS* comprendono le sigle (in maiuscolo) delle nove parti del discorso: VER (verbo); NOM (nome); ADJ (aggettivo); PRO (pronome); DET (determinante); ADV (avverbio); PRE (preposizione); CON (congiunzione); INJ (interiezione). Tali sigle sono ulteriormente specificate, dove pertinente, dall'ag-

[26] I lemmi del dizionario sono adottati con alcuni adattamenti. Si veda la documentazione riportata in Camps – Albarran – Cochet – Ing 2019: https://github.com/Jean-Baptiste-Camps/Geste.

giunta di informazioni (in minuscolo) sulla particolare natura ("tipo") di ogni categoria. L'elenco completo delle etichette è riportato di seguito in due colonne:

VERcjg	*verbe conjugué*	DETdef	*déterminant défini*
VERinf	*verbe infinitif*	DETndf	*déterminant non défini*
VERppe	*participe passé*	DETdem	*déterminant démonstratif*
VERppa	*participe présent*	DETpos	*déterminant possessif*
NOMcom	*nom commun*	DETind	*déterminant indéfini*
NOMpro	*nom propre*	DETcar	*déterminant cardinal*
ADJqua	*adjectif qualificatif*	DETrel	*déterminant relatif*
ADJind	*adjectif indéfini*	DETint	*déterminant interrogatif*
ADJpos	*adjectif possessif*	DETcom	*déterminant défini composé*
ADJcar	*adjectif cardinal*	ADVgen	*adverbe général*
ADJord	*adjectif ordinal*	ADVneg	*adverbe de négation*
PROper	*pronom personnel*	ADVint	*adverbe interrogatif*
PROimp	*pronom impersonnel*	ADVing	*adverbe interrogatif négatif*
PROadv	*pronom adverbial*	ADVsub	*adverbe «subordonnant»*
PROpos	*pronom possessif*	PRE	*préposition*
PROdem	*pronom démonstratif*	CONcoo	*conjonction de coordination*
PROind	*pronom indéfini*	CONsub	*conjonction de subordination*
PROcar	*pronom cardinal*	INJ	*interjection*
PROord	*pronom ordinal*		
PROrel	*pronom relatif*		
PROint	*pronom interrogatif*		

A tale lista si aggiungono inoltre le etichette: PON (*ponctuation)*; ETR (*mot étranger)*; ABR (*abréviation)* RED (*mot redondant*), meno rilevanti agli scopi del presente articolo, e l'etichetta OUT (*catégorie temporaire*), utilizzata nella rara eventualità che due forme grafiche debbano essere associate ad un'unica etichetta, come nel caso del pronome relativo analitico *le quel*[27]. In questo caso, per non violare i principi di annotazione (vedi *infra*), l'etichetta di pronome relativo (PROrel) viene attribuita solo a *quel*, mentre il determinante *le* riceve il valore OUT[28].

È importante sottolineare, inoltre, che il sistema di annotazione *Cattex09* include anche un gruppo di etichette per la rappresentazione di forme complesse risultanti da fenomeni di enclisi o di proclisi. Ciò permette, in linea di principio, di alleggerire la fase preparatoria di manipolazione del testo, evitando di dover spezzare tali forme per garantirne il riconoscimento da parte della macchina (cfr. § 3.1), un aspetto sul quale, tuttavia, sarà necessario tornare più avanti (cfr. § 5.1). Le combinazioni previste da *Cattex09* disponibili in *Pyrrha* sono le seguenti:

[27] Guillot – Prévost – Lavrentiev 2013a: 4.
[28] Guillot – Prévost – Lavrentiev 2013b: 16.

ADVgen.PROadv	ADVgen.PROper	ADVneg.PROper
CONcoo.DETdef	CONcoo.PROper	CONsub.ADVgen
CONsub.DETdef	CONsub.PROper	PRE.DETcom
PRE.DETdef	PRE.DETrel	PRE.PROper
PRE.PROrel	PROper.PROper	PROrel.PROadv
PROrel.PROper		

Quanto all'annotazione morfologica, i principi adottati permettono di indicare, laddove pertinente, le categorie di modo, tempo, persona, numero, genere, caso e grado. Le etichette previste sono riassunte nella tavola seguente:

CATEGORIA MORFOLOGICA	ETICHETTE(*CATTEX09*)	LEGENDA(*CATTEX09*)
MODO	ind	*indicatif*
	imp	*impératif*
	con	*conditionnel*
	sub	*subjonctif*
TEMPO	pst	*présent*
	ipf	*imparfait*
	fut	*futur*
	psp	*passé simple*
PERSONA	0	*impersonel*
	1	*1ère personne*
	2	*2ème personne*
	3	*3ème personne*
NUMERO	s	*singulier*
	p	*pluriel*
GENERE	m	*masculin*
	f	*féminin*
	n	*neutre*
CASO	n	*nominatif*
	r	*régime*
	i	*régime indirect*
GRADO	p	*positif*
	c	*comparatif*
	s	*superlatif*

Ogni forma chiaramente necessita di essere associata a diverse informazioni morfologiche. Pertanto la *control list* di *Pyrrha* prevede tutte le possibilità con cui tali valori possono combinarsi nel campo *Morph*. Così, ad esempio un aggettivo femminile plurale di caso nominativo e grado superlativo viene descritto dalla stringa: NOMB.=p|GENRE=f|CAS=n|DEGRE=s.

Oltre alle etichette sopra elencate, la lista di controllo di *Pyrrha* include anche il valore NOMB.=x|GENRE=x|CAS=x per situazioni in cui l'attribuzione di genere numero e caso non sia applicabile, come nel caso del gerundio. Qualora, la forma analizzata non presenti flessione (forma invariabile), *Pyrrha* prevede, infine, l'etichetta '*empty*', per indicare l'assenza di informazioni morfologiche.

Il principio fondamentale che guida l'annotazione è che ogni forma deve essere associata ad una e una sola etichetta per ciascun campo. In caso di ambiguità, l'annotatore è comunque obbligato a operare una scelta univoca, aiutandosi con l'osservazione delle forme precedenti e seguenti, oppure prendendo nota dei propri dubbi altrove[29].

A tale proposito, è importante sottolineare che l'annotazione morfo-sintattica avviene su base contestuale, secondo criteri perlopiù morfologici e in qualche occasione distribuzionali[30]. Una stessa forma dimostrativa, ad esempio, può svolgere sia la funzione di pronome che quella di determinante, ma l'attribuzione dell'uno o dell'altro valore dipende esclusivamente dal contesto, come si vede nel caso seguente, dove la medesima forma *cele* assume il valore pronominale quando ricorre da sola (1b) e il valore di determinante quando è seguita da un sostantivo (1a).

(1) a. *Et la damoisele torne* ***cele*** *part si tost come il son pres* (*Graal*) [*cele* = DETdem]
b. *Et* ***cele*** *dit que onques deseritee n'en fu* (*Graal*) [*cele* = PROdem][31]

Tale principio di annotazione si applica in maniera evidente in quei casi in cui una forma assuma, in un particolare contesto morfo-sintattico, un valore non corrispondente a quello registrato nel lessico. La forma *dame*, ad esempio, è associata nel lessico a proprietà esclusivamente nominali. Tuttavia, nel contesto seguente, riportato nella documentazione del *Cattex*, *dame* assume una funzione assimilabile a quella di un aggettivo a causa della presenza del modificatore avverbiale *plus*.

(2) *Et cele qui estoit la plus* ***dame*** *le menoit par la main et ploroit mout tendrement* (*Graal*)[32]

Si comprende così l'importanza della doppia etichettatura, morfo-sintattica (*POS*) e morfologica (*Morph*), che consente di restituire ad ogni forma una duplice annotazione, descrivendone sia la funzione svolta nel particolare contesto

[29] Guillot – Prévost – Lavrentiev 2013a: 2.
[30] *Ibid.*
[31] *Ibid*.
[32] L'esempio è tratto da ivi: 3.

sintattico, sia la forma morfologica quale essa realmente appare. Ciò permette in questo modo di analizzare anche quelle forme in cui la funzione sintattica non coincide con la morfologia, proprio come nel caso di *dame* in (2) che, in base ai criteri di annotazione esposti, viene trattata come aggettivo (ADJqua) sul piano dell'analisi sintattica, ma dal punto di vista morfologico è descritta come un sostantivo, in coerenza con le sue proprietà lessicali.

3.4 *Vantaggi e limiti di* Pyrrha *e primi adattamenti per il* DiFrI

Alla luce di quanto esposto sopra, i vantaggi offerti da *Pyrrha* per la costruzione di un *corpus* lemmatizzato di franco-italiano sono molteplici. Tra questi, un aspetto particolarmente favorevole è rappresentato dall'interfaccia, che risulta immediatamente intuitiva e agevole per l'utente. Ciò, come si è detto, costituisce un particolare per nulla secondario, dal momento che il risultato dell'annotazione automatica dei testi maggiormente italianizzati presenta un margine di errore piuttosto ampio, che rende necessario un cospicuo intervento di correzione manuale.

D'altra parte, se è vero che l'automatizzazione produce per alcuni testi una abbondante quantità di errori, il vantaggio di poter disporre di uno strumento che consente di analizzare efficacemente e automaticamente almeno ciò che pertiene al sistema linguistico del francese antico è comunque notevole, in assenza di un mezzo appositamente progettato per il franco-italiano.

Quanto alla fase di correzione manuale, la presenza di liste di controllo, il fatto che il programma suggerisca il completamento dell'annotazione durante la compilazione dei campi e che impedisca il salvataggio di stringhe non corrette sono ottime tutele rispetto al rischio di errore umano. Particolarmente vantaggiose, infine, sono la possibilità di correggere simultaneamente più forme identiche (cliccando sul numero che compare nel campo *Similar*) e la funzione *Search tokens*, che permette di eseguire ricerche per forme, lemmi, etichette *POS* o *Morph.* Tutto ciò, infatti, non solo riduce ulteriormente il pericolo di inserire errori, ma consente anche di verificare di aver adottato in ogni situazione un'annotazione coerente, offrendo così un un'ottima garanzia rispetto al requisito fondamentale di ogni *corpus*: l'uniformità.

A fronte dei molti vantaggi, va detto però che *Pyrrha* è un programma recente che, per alcuni aspetti, è ancora in via di implementazione. Inoltre, per gli scopi del progetto, le stesse funzionalità che garantiscono l'uniformità del *corpus* si rivelano talora rigide e limitanti rispetto alle esigenze di annotazione del franco-italiano. L'impossibilità, ad esempio, di inserire lemmi non previsti dal dizionario di riferimento, se non richiedendo ai tecnici di *Pyrrha* un'integrazione ufficiale della *control list*, costituisce un ostacolo notevole all'avanzamento del lavoro di lemmatizzazione.

Per questo, una delle prime esigenze manifestate dai collaboratori del *DiFrI* è stata quella di creare una lista di controllo appositamente dedicata al franco-italiano, da realizzare in collaborazione con gli informatici dell'*École nationale des*

Chartes. Il primo intervento in questo senso è rappresentato dalla creazione di un'etichetta specificamente dedicata alla descrizione delle forme ibride, tipiche del franco-italiano: SPEC=it. L'etichetta al momento trova spazio nel campo *Morph* e può combinarsi con tutti i valori previsti per l'annotazione morfologica. Tale collocazione non è del tutto felice dal momento che l'ibridismo linguistico non coinvolge necessariamente solo la morfologia, ma può manifestarsi anche a livello fonologico o semantico. Si tratta, dunque, di una soluzione ancora temporanea. Essa costituisce però un primo passo che consente perlomeno di raccogliere tutte le forme ibride in vista di poterle meglio specificare in futuro.

3.5 *Le ultime fasi: esportazione e pubblicazione dei risultati*

Una volta completata la correzione manuale, i risultati possono essere estratti da *Pyrrha* per essere conservati e pubblicati. La procedura elaborata per il *DriFrI* prevede che il prodotto dell'annotazione sia esportato nel formato TSV (*tab-separated values*), un semplice *file* di testo che conserva la struttura tabellare dei dati. Quindi, il *file* viene affidato all'assistente informatico del progetto, che vi applica una serie di procedure *Java*, lo elabora e lo trasforma infine in formato XML (*Extensible Markup Language*), che traduce il documento in una sintassi leggibile sia dall'uomo che dalla macchina. Questo *file* costituisce la base per ogni elaborazione successiva, e dunque anche per eventuali interventi a modifica dell'annotazione.

La sede scelta per la pubblicazione dei risultati è il sito del *RIALFrI*, dove vengono caricati i file XML dei testi lemmatizzati. Qui gli utenti possono consultare i dati della lemmatizzazione attraverso due modalità. La prima prevede la consultazione diretta del testo di interesse: con un *click* del *mouse* è possibile attivare per ogni parola l'apertura di un *pop-up* che contiene l'analisi completa della forma, riportandone sia il lemma, sia la descrizione morfo-sintattica. Il secondo metodo, invece, è la funzione di *Ricerca avanzata* proposta dal sito, che offre una maschera di interrogazione attraverso la quale è possibile eseguire ricerche per forme e per lemmi, naturalmente interrogando, nel secondo caso, solamente la porzione di *corpus* lemmatizzata.

4. *La lemmatizzazione della* Geste Francor*: le* Enfances Bovo.

Le prime prove di lemmatizzazione per il *DiFrI* sono state affidate simultaneamente a due opere esemplari della letteratura franco-italiana: l'*Entrée d'Espagne* e la *Geste Francor*. Della prima sono state lemmatizzate a cura di Floriana Ceresato le porzioni di testo dell'edizione di Thomas riviste e tradotte da Infurna[33]. Della

[33] La lemmatizzazione dell'*Entrée d'Espagne*, pubblicata sul sito del *RIALFrI*, ha utilizzato i

seconda, che costituisce l'oggetto del presente articolo, è stata lemmatizzata la prima storia, le *Enfances Bovo*. La scelta dei testi di partenza non è stata casuale. Si tratta, infatti, in entrambi i casi di opere di origine italiana che tuttavia presentano caratteristiche molto diverse, essendo l'una a fondo linguistico francese e l'altra a fondo italiano[34]. Ciò ha permesso dunque di testare gli strumenti informatici e i criteri di annotazione impiegati rispetto ad un'ampia gamma di situazioni ritenute rappresentative delle problematiche più ricorrenti nella lemmatizzazione del franco-italiano.

In questa prospettiva, dopo un breve cenno sulla lingua della *Geste Francor*, i paragrafi seguenti illustreranno la lemmatizzazione delle *Enfances Bovo*, sottolineando le criticità sollevate da un testo pesantemente italianizzato ed esponendo i criteri conseguentemente adottati per la sua annotazione.

4.1 *La lingua della* Geste Francor

Nel già variegato panorama della letteratura franco-italiana, la lingua della *Geste Francor*, ospitata dal codice marciano Francese Z 13 (256), spesso citato come V13, rappresenta un caso singolare. La sua natura sfuggente trova senz'altro ragione nella storia della tradizione, come esito di una stratificazione di elementi linguistici prodotta da successive fasi di rielaborazione testuali[35], a cui concorrono, da un lato, lo sforzo di imitare i prestigiosi modelli francesi, dall'altro, un livellamento in direzione dell'italiano atto a soddisfare le esigenze comunicative di un nuovo pubblico[36]. Come osserva Maria Grazia Capusso, infatti, «anche le parti della *Geste Francor* provviste di riconoscibili ed illustri modelli si caratterizzano per un accentuato interventismo compositivo»[37], attraverso il quale l'anonimo compositore mette in atto al massimo grado le «potenzialità compromissorie dei due interloquenti sistemi linguistici»[38] per venire incontro «alle attese ideologiche ed estetiche di un pubblico ormai ben lontano dall'originaria aristocrazia feudale»[39]. Un tale ibridismo si manifesta nella mescidanza continua di francese e italiano e nella presenza di una cospicua quantità di vocaboli di incerto etimo e significato con effetti «surreali e insieme sapidamente espressivi»[40]. La lingua della *Geste Francor* può essere così qualificata, secondo le parole di Mascitelli, come

medesimi strumenti impiegati per la *Geste Francor* (*Pyrrha*, sistema di annotazione *Cattex09*). In questa fase sperimentale, però, i criteri di annotazione specificamente adottati per il franco-italiano hanno seguito per i due testi soluzioni non sempre comuni.

[34] Cfr. Barbato 2015: 47.

[35] Cfr. Mascitelli 2020: 257.

[36] Ivi: 261.

[37] Capusso 2007: 179

[38] *Ibid.*

[39] *Ibid.*

[40] *Ibid.*

> un idioletto, dotato di una marcata connotazione poetica e […] sensibilmente proteso verso l'italiano, nel quale gli apporti del francese 'letterario', del francese di uso pratico e della lingua materna dell'autore e dei copisti si sovrappongono fino a dissolversi, realizzando una sintesi in cui caratteri e funzioni dei due sistemi linguistici in interazione finiscono […] per fondersi e confondersi[41].

4.2 *Note generali sull'annotazione digitale delle* Enfances Bovo

Per la lemmatizzazione delle *Enfances Bovo* si è utilizzato il testo digitalizzato in *RIALFrI*, che assume come riferimento l'edizione della *Geste Francor* curata da Leslie Zarker Morgan. L'attribuzione dei lemmi e la descrizione morfologica delle forme presenti nel testo seguono, dunque, in linea di principio, l'interpretazione proposta dall'editrice e ricostruibile dalle note e dal glossario, offerti a corredo dell'opera. L'adozione di questo criterio operativo convive con la consapevolezza che diverse interpretazioni testuali e linguistiche avanzate dall'edizione di Zarker Morgan si prestano ad un'ampia discussione[42]. Tuttavia, il dovere di oggettività a cui l'annotatore si impegna è tale da rendere tale *modus operandi* una scelta obbligata, al fine di escludere, o comunque limitare al massimo, ogni soluzione arbitraria e garantire una fedele adesione al testo di riferimento. Pertanto, laddove l'applicazione di tale criterio sia apparsa impossibile, costringendo a discostarsi dalle scelte dell'edizione, si è tenuto conto delle divergenze annotandole in un'apposita documentazione.

Nell'insieme, l'annotazione delle *Enfances Bovo* si è dimostrata assai complessa. L'assenza di una traduzione di supporto e la presenza di passi lacunosi hanno reso talvolta difficile l'interpretazione di alcune porzioni testuali e la descrizione delle forme, costringendo a convivere con il dubbio delle scelte operate.

Inoltre, la concomitanza dell'attività di annotazione con alcuni lavori di implementazione della *control list* dedicata al franco-italiano ha reso indisponibile per un lungo periodo la funzione di controllo sulla correttezza dei lemmi inseriti, esponendo il lavoro ad un certo margine di errore umano. Occorre dunque avvertire che, nonostante l'annotazione sia stata eseguita con il massimo scrupolo, non si esclude per il futuro la necessità di eventuali interventi di correzione.

4.3 *Criteri per la lemmatizzazione*

Nella fase di lemmatizzazione, la problematicità dell'annotazione del franco-italiano si manifesta in tutta la sua complessità. Come già detto (cfr. § 2.1), il trattamento di testi linguisticamente misti mette in dubbio la possibilità di impiegare lemmi pertinenti esclusivamente ad uno dei sistemi linguistici coinvolti. Ciò, nel caso delle *Enfances Bovo,* è apparso immediatamente evidente. La massic-

[41] Mascitelli 2020: 262.
[42] A tal proposito, si vedano, tra gli altri, Beretta 2011 e Giannini 2012.

cia presenza di forme schiettamente italiane ha reso palese, infatti, l'insufficienza dell'elenco di lemmi francesi disponibili in *Pyrrha* e ha dimostrato l'urgenza di una lista di lemmi multipla, comprendente sia lemmi francesi che lemmi italiani, anche eventualmente specifici di alcuni volgari settentrionali. Una tale necessità ha avuto tuttavia implicazioni di natura sia tecnica che teorica. Dal punto di vista tecnico, è pesata in particolare l'impossibilità di integrare autonomamente la *control list* di *Pyrrha* (cfr. §§ 3.3, 3.4). Del resto, sporadiche richieste di intervento ai tecnici dell'*École des Chartes* per l'aggiunta di singoli lemmi sarebbe stata troppo dispendiosa in termini di tempo, data la significativa quantità di italianismi. Si calcola, infatti, che al momento le forme da ricondurre ad un lemma italiano costituiscano circa il 10% delle forme totali presenti nel testo. Pertanto, in attesa di una lista di controllo per lemmi appositamente creata per il franco-italiano (cfr. § 5.2), l'unica soluzione applicabile per le *Enfances Bovo* è stata la compilazione manuale del file TSV estratto da *Pyrrha*.

Sul piano teorico, invece, la possibilità di ricorrere a lemmi appartenenti a sistemi linguistici diversi ha posto il problema dei criteri in base ai quali lemmatizzare le singole forme. Si è trattato di una scelta complessa che ha dovuto tenere conto della presenza tanto di forme francesi e italiane quanto di una messe inestricabile di forme ibride.

Considerata la volontà degli autori di scrivere in francese[43], la scelta di *default* è stata quella di ricondurre le forme al lemma francese tutte le volte che fosse possibile, riservando il lemma italiano (o dialettale) esclusivamente alle forme schiettamente italiane (o dialettali), secondo i criteri seguenti[44]:

a) Privilegiare sempre l'attribuzione del lemma francese.
b) Ricondurre al francese le forme ibride che condividano l'etimologia e la semantica francese.
c) Ricondurre al lemma italiano (o dialettale) le forme (anche ibride) non riconducibili ad alcun lemma francese.
d) Ricondurre all'italiano le forme italiane perfettamente prodotte anche quando condividono l'etimologia e la semantica del francese.

Così, ad esempio, una forma ibrida come *espea*, in cui il morfema desinenziale *-a* denuncia chiaramente l'interferenza con l'italiano, è stata ricondotta al lemma francese *espee* piuttosto che all'italiano *spada*, in quanto la semantica e l'etimologia

[43] A tal proposito, si vedano, tra gli altri, Viscardi 1941: 46-47 e Capusso 2007: 166-169, 172-173.

[44] Questi criteri sono stati presentati per la prima volta da Francesca Gambino e Sira Rodeghiero in occasione del convegno «GraVO», tenutosi all'Università di Padova, dal 5 al 6 dicembre 2019. Titolo dell'intervento: *Dizionario del franco-italiano (DiFrI): la lemmatizzazione dei testi, le prime voci*.

della parola sono le stesse condivise anche dal francese. Diversamente, la forma *così*, inesistente in francese, è stata lemmatizzata come italiana, similmente ad una forma come *colpo*, che, pur condividendo la semantica e l'etimologia del francese *coup*, risulta perfettamente formata in italiano.

Quanto ai lemmi impiegati, se per il francese antico si è mantenuto il riferimento al dizionario *Tobler-Lommatzsch*, per l'italiano si è scelto di appoggiarsi al *TLIO* e al *GDLI* (Battaglia), ricorrendo invece al dizionario di Boerio per il veneto. Il rinvio ai dizionari menzionati è avvenuto secondo una norma a scalare: se un lemma non era presente nel *TLIO*, si è rinviato al *GDLI*; se era assente anche qui e si trattava di una forma veneta, allora si è ricondotto al vocabolario di Boerio. A tal proposito, occorre però specificare che la distinzione tra generici settentrionalismi e venetismi può risultare ardua e non sempre possibile. Pertanto, nel corso della lemmatizzazione delle *Enfances Bovo*, si è ritenuto di integrare i criteri c) e d) sopracitati con un'ulteriore norma prudenziale:

e) Ricondurre di preferenza al lemma italiano tutte le forme chiaramente ascrivibili ai sistemi linguistici peninsulari, dialetti compresi.

Di conseguenza, forme come *amigo*, *fiolo*, *mejo*, *guera* sono state rinviate rispettivamente ai lemmi italiani *amico*, *figliuolo, meglio, guerra*. Continuando con gli esempi, lo stesso trattamento è stato riservato ad *aradegé* nel contesto seguente:

(3) *E m'en partì cun l'ovra fu finé,* | *Por venir enver la mersalé.* |*Ma in le boscho e fu* ***aradegé****,* | *Por altra via eo fu açaminé.* (11.856-859)

La forma, come riportato dalla nota dell'edizione al verso 858, è molto probabilmente una forma settentrionale dal significato di 'vagare, errare, perdersi' (*erraticare). Nel dizionario di Boerio esiste un lemma *radegar*, tuttavia, in base al principio appena enunciato, si è preferito rinviare la forma al lemma italiano *eradegar* del *TLIO*.

L'adozione dei criteri menzionati, benché utile a dirimere un'abbondante quantità di situazioni, non è tuttavia risolutiva di ogni problema. La lemmatizzazione delle *Enfances Bovo* ha messo in luce diversi casi di ambiguità. Alcune forme, infatti, possono apparire a pari diritto francesi o italiane. Si pensi, ad esempio, a *fu*, che può corrispondere tanto alla terza persona singolare del *passé simple* francese quanto a quella del passato remoto italiano, oppure a *dona*, che può essere interpretato sia come la terza persona singolare del *passé simple*, sia come la terza persona singolare del presente italiano o del perfetto semplice settentrionale.

Alla luce di tali ambiguità, si è deciso di generalizzare l'attribuzione del lemma francese a tutte le forme appartenenti a determinate categorie, quali determinanti definiti (cioè articoli determinativi), determinanti, aggettivi e pronomi possessivi e forme verbali, aggiungendo ai criteri la seguente norma:

f) Ricondurre di *default* al lemma francese tutti i determinanti definiti, tutti i determinanti, gli aggettivi e i pronomi possessivi e tutte le forme verbali.

Nel caso di forme individuabili senza ambiguità come italianismi schietti o forme ibride, l'indicazione di italianismo/italianizzazione è stata demandata unicamente all'annotazione morfologica tramite l'applicazione dell'etichetta SPEC=it, come nell'esempio seguente:

(4) *Coment Bovo fo in* ***lo*** *castel Siginbaldo* (Rubrica 4)

Forma	Lemma	MORPH
lo	*le*	SPEC=it|NOMB.=s|GENRE=m|CAS=r

Ciò non è valso però per alcune forme verbali (tra cui *aradegé* dell'esempio 3), che, prive di un corrispondente lemma francese, sono state ricondotte al lemma italiano in deroga al criterio f). Si vedano a tal proposito i lemmi verbali nella lista di forme ricondotte all'italiano elencate in Appendice.

Normalmente il lemma attribuito condivide la categoria sintattica della forma analizzata. In alcuni casi, tuttavia, l'adesione alle scelte del dizionario *Tobler-Lommatzsch* ha comportato alcune divergenze. È questo il caso di *conbatant*, che, pur descritto come sostantivo, è stato rinviato al lemma verbale *combatre*. Similmente è avvenuto per molti avverbi in *-ment/-mant*. Per il futuro si progetta di integrare la *control list* di *Pyrrha* in modo da poter rinviare le forme a lemmi avverbiali. Per ora, però, tali forme sono state rimandate generalmente agli aggettivi corrispondenti, secondo le scelte del *Tobler-Lommatzsch*:

Forma	**Lemma**
alegramant	*haliegre*
altament	*haut*
altrament	*autre*
be(l)lemant	*bel1*
çelcemant	*celer1*
comunelmant	*comunal*
cortesement	*cortois*
cortoismant	*cortois*
dolçemant	*douz*
estoitament	*estroit*
isnelamant/isnelemant	*isnel*
lialmant	*lëal*
lojalmant/lojalment	*lëal*
longemant	*lonc*
lungament	*lonc*
malament/malemant	*mal1*

primemant	*prim*
primeremant	*premier*
richament	*riche*
saçamant	*sage*
saviamant	*sage*
seguramant/segurament	*sëur2*
stroitamente	*estroit*
tenderemant/tenderamant	*tendre2*
vigorosamant	*vigoros*

In qualche occasione, è capitato che il *Tobler-Lommatzsch* non riuscisse a soddisfare le esigenze di lemmatizzazione, mancando di corrispondenze rispetto ad alcune forme riconducibili al francese. In queste circostanze, si è deciso allora di ricorrere in seconda istanza al *DMF*. Così, ad esempio, la forma *anbasea* in (5), che il glossario dell'edizione rimanda a *ambasse/-ee*, non era presente nel *TL* ed è stata dunque rinviata al lemma *ambassie* del *DMF*.

(5) *S'**anbasea** le dient con homes pros e ber* (10.672)

L'annotazione delle *Enfances Bovo* ha dimostrato che talvolta la difficoltà ad individuare criteri per la lemmatizzazione (come per la descrizione morfologica) del franco-italiano è complicata, oltre che dalle forme mescidate, anche dalla presenza di errori e di forme di ambigua interpretazione nel contesto. In casi simili, si è fatto valere il principio generale enunciato al § 4.2: il dato è stato analizzato unicamente per quale esso appariva nel testo edito, evitando ogni interpretazione personale e seguendo ciecamente eventuali suggerimenti del glossario o delle note. Si veda a tal proposito l'esempio successivo, dove la forma *persant* è di problematica interpretazione. Il significato atteso è quello di 'ordine, comando', ma, come riportato in nota, questo significato non è solitamente associato con il lemma *persant* ('Persiano'), tanto da fare pensare ad una deformazione della parola in contesto di rima. Poiché il problema è apparentemente insolubile, lungi dal prendere posizioni arbitrarie, la lemmatizzazione rispetta la scelta editoriale e riconduce la forma al lemma francese *persant*.

(6) *Poco v'amò, qi vos dè li **persant***, (16.1141)

Vale la pena, infine, menzionare un ultimo aspetto problematico, ovvero la lemmatizzazione dei nomi propri. In questo caso, infatti, il dizionario di riferimento non offre alcun supporto e la scelta di una forma base a cui ricondurre nomi di persona o di luogo si presenta alquanto difficile di fronte all'ampia quantità di varianti con cui essi si presentano. Nel caso della *Geste Francor*, l'interferenza con il sistema linguistico dell'italiano è poi particolarmente pervasiva in tale contesto. Per evitare decisioni arbitrarie, nella lemmatizzazione delle *Enfances Bovo* si è deciso dunque di riferirsi, nell'ordine, ai repertori di Moisan e di Flutre, assu-

mendo come lemma per i nomi la forma di citazione proposta nel repertorio. In questo modo, si è sempre cercato di privilegiare il francese come forma base, anche se in alcuni casi è stato necessario assumere come lemma la forma di citazione proposta nel volume di Moisan dedicato ai nomi propri citati nelle opere straniere. L'intervento di annotazione è stato anche in questo caso, come per i lemmi italiani, del tutto manuale ed è avvenuto sul file TSV estratto da *Pyrrha*. L'elenco completo dei nomi propri citati nelle *Enfances Bovo* è riportato in Appendice (App. II).

4.4 *Criteri per l'annotazione morfo-sintattica (POS)*

L'attribuzione delle forme alle diverse parti del discorso è avvenuta seguendo i principi generali del *Cattex09*, senza sostanziali modifiche. Vale la pena, tuttavia, precisare pochi aspetti per i quali l'annotazione del *corpus* ha richiesto minime integrazioni o adattamenti rispetto alle linee guida.

A tal proposito, un ambito nel quale il manuale di riferimento risulta meno efficace è il trattamento delle forme di quantificazione. Nel *Cattex09*, infatti, i quantificatori non costituiscono una categoria morfo-sintattica a parte, dotata di un'etichetta *POS* propria, ma sono distribuiti tra le categorie dei determinanti, degli aggettivi e dei pronomi indefiniti, secondo le consuetudini della grammatica tradizionale. Un uso avverbiale della forma *tout/tous* è tuttavia riconosciuto quando essa precede un aggettivo qualificativo (es. *tute blanche*[45]). È opportuno tuttavia sottolineare che la stessa funzione avverbiale può presentarsi anche quando la forma modifica un participio passato come in (7). In questi contesti il potenziale limite della scelta di non dedicare ai quantificatori una propria etichetta *POS* si manifesta nell'incongruenza tra l'annotazione morfo-sintattica e quella morfologica. A dispetto degli altri avverbi, infatti, il quantificatore *tout* con valore avverbiale riceve necessariamente assegnazione di genere, numero e caso.

(7) *Donde la cort fu* ***tota*** *resvigoré* (11.869)

Form	*POS*	*Morph*
tota	ADVgen	SPEC=it|NOMB.=s|GENRE=f|CAS=n

Inoltre, continuando a proposito dei quantificatori, va riconosciuta la possibilità che essi possano svolgere funzione aggettivale anche se posposti rispetto al nome a cui si riferiscono, come in (8). Ciò non sembra previsto dal *Cattex09* che definisce l'aggettivo indefinito come «une forme indéfinie précédée d'un déterminant et suivie d'un nom»[46], menzionando come eccezione solamente l'uso posposto di *même*, annoverato tra gli indefiniti invece che tra i dimostrativi.

[45] Guillot – Prévost – Lavrentiev 2013b: 13.
[46] Guillot – Prévost – Lavrentiev 2013b: 9.

(8) *Se Bovo enver de vos porta lialtà* ***tant****;* (13.981)

Form	*POS*	*Morph*
tant	ADJind	NOMB.=s|GENRE=f|CAS=r|DEGRE=p

Nell'ambito dei pronomi, una precisazione merita invece il trattamento del clitico *se*, che, nelle *Enfances Bovo*, può svolgere, oltre alla funzione di pronome riflessivo, anche le funzioni impersonale e passivante proprie del *si* dell'italiano. In tali casi, la forma, ricondotta al lemma *si* del *GDLI*, è stata descritta come pronome impersonale privo di flessione morfologica.

(9) *Ela sa qe il est li milor çivaler* | *Qe* ***se*** *poust en tot li mondo trover.* (10.657-658)

Form	*Lemma*	*POS*	*Morph*
se	*si*	PROimp	SPEC=it|MORPH=empty

Infine, un'ultima nota marginale è richiesta per il fenomeno della sostantivizzazione. A tale proposito, il principio generale del *Cattex09* prevede che la forma venga trattata come sostantivo dal punto di vista morfo-sintattico, ricevendo però, sul piano morfologico, la descrizione propria della sua categoria originaria. Così, ad esempio, un aggettivo sostantivato riceve l'etichetta NOMcom per il campo *POS*, ma è annotato morfologicamente come un aggettivo. Rispetto a tali criteri, l'annotazione delle *Enfances Bovo* diverge per il trattamento dell'infinito sostantivato, che, anziché ricevere l'annotazione morfologica dell'infinito (nel *corpus* di franco-italiano, NOMB.=x|GENRE=x|CAS=x), riceve la descrizione morfologica propria dei nomi[47], con attribuzione di *default* di genere maschile e numero singolare, come in (10):

(10) *De ço q'è fato, preso sui de* ***l'amender****.* (12.941)

Form	*POS*	*Morph*
amender	NOMcom	NOMB.=s|GENRE=m|CAS=r

4.5 *Criteri per l'annotazione morfologica*

Il riempimento del campo *Morph* in *Pyrrha* impone di affrontare due ordini di problemi: stabilire l'eventuale italianità/italianizzazione delle forme e fornirne una corretta descrizione morfologica.

La prima informazione è affidata all'etichetta SPEC=it (cfr. § 3.4), che, come si è detto, pur essendo collocata nel campo *Morph*, ha una portata ampia che eccede talvolta i confini di osservazioni esclusivamente morfologiche. Nella *Geste*

[47] Tale criterio è assunto per uniformità con l'annotazione degli altri testi lemmatizzati inclusi nel progetto.

Francor, data l'estesissima pervasività dell'elemento italiano (cfr. § 4.1), l'aggiunta di tale informazione potrebbe accompagnarsi alla gran parte delle forme incluse nel testo. Pertanto, per evitare di rendere inefficace l'impiego di tale etichetta con uso troppo intensivo, nell'annotazione delle *Enfances Bovo*, si è scelto di riservare temporaneamente il valore SPEC=it solamente ad alcuni contesti. Tra questi sono state incluse in generale:

a) Forme ricondotte a un lemma italiano.
b) Forme che manifestano determinati fenomeni grafico-fonetici[48], quali:
   - uso del digramma *ch-* per indicare la pronuncia velare dell'italiano (es. *chi*, *poche*, *richa*);
   - uso di *x* per designare la sibilante sonora (es. *Druxiana*);
   - fenomeni metafonetici (es. *quisti*);
   - fenomeni di degeminazione (es. *spala*);
   - casi di sonorizzazione delle occlusive sorde in posizione intervocalica (es. *digo*, *perigolo*);
   - aggiunta di *a-* prostetica in contesto verbale (es. *aseré*, *asaçer*);
   - aferesi di *e-* (es. *spea*, *scanpé*, *scremie*).

Dal punto di vista propriamente morfologico, poiché la *Geste Francor*, salvo pochissime eccezioni, fa un uso quasi completamente indifferenziato di *cas sujet* e *cas régime*, si è scelto di ignorare il mancato rispetto della distinzione bicasuale del francese antico. L'attributo SPEC=it è stato invece assegnato nei casi indicati di seguito[49].

c) Articoli:
   - indeterminativi: *una*;
   - determinativi: *lo*, *i (masch. plur).*

d) Sostantivi e aggettivi: forme con suffissazione italiana, dunque:
   -maschili singolari in *-o* (es. *çevo*, *flanco*; *sano*, *legro*);
   -maschili plurali in *-i* (es. *brandi*, *colpi*; *richi*, *culverti*);
   -femminili singolari in *-a* (es. *çanbra*, *lança*; *richa*, *droita*);
   -femminili plurali in *-e* (es. *vile*, *polçele*; *poche*).

[48] L'individuazione dei fenomeni grafico-fonetici derivanti dall'interazione dei sistemi linguistici del francese e dell'italiano richiede ulteriore attenzione. L'elenco fornito è parziale e il *corpus* necessita nel prossimo futuro di interventi di integrazione e correzione. Tra questi, si inserisce, ad esempio, la necessità di una definizione dei contesti in cui l'uso della grafia *ç* costituisce una spia dell'interferenza linguistica, così come l'inclusione dei fenomeni di apertura di *-e-* in *-a-* davanti a nasale (es. *çant*).

[49] L'individuazione dei contesti morfologici a cui attribuire l'etichetta SPEC=it si è avvalsa notevolmente dello studio linguistico della *Geste Francor* offerto da Mascitelli 2020: 286-337.

e) Sostantivi: metaplasmi di genere. Si considerino a tal proposito le forme: *amor*, che, dal genere femminile tipico del francese, passa talvolta al maschile, come in italiano: *vostro amor* (v. 10.685); *braçe*, che dal neutro plurale passa al femminile, come in: *entro ses braçe* (v. 8.557).

f) Dimostrativi: *questo*, *quisti*, *questa*, *cesto*, *cesta*, *sta*, *ste*, *quel*, *quelo*, *qelo*, *quela*, *quella*, *quele*, *cela*, *cella*, *qui*, *cestu*, *colu*.

g) Possessivi: *me*, *moja*, *to*, *so*, *soa*, *nostro*, *nostra*, *nostri*, *vostro*, *vestra*.

h) Pronomi personali:
- in funzione di soggetto: *e*, *eo*, *elo*, *ello*, *el*, *ela*, *ella*, *ila*, *la*, *nu*, *vu*;
- in funzione di complemento (forme toniche): *lui* (preceduto dalla preposizione *a* per esprimere valore dativale[50]), *loro*;
- in funzione di complemento (forme atone): *lo*, *ne*, *ve*, *vi*, *li*.
- Pronome riflessivo: *si*.

i) Relativi:
- *qe* con funzione di soggetto;
- pronome relativo analitico: *la quale*, *la qual*;
- le forme: *quanto*, *donde.*

j) Particella pronominali: *ge*, *li*.

k) Indefiniti: *uno*, *altro*, *altri*, *molto*, *qualche*, *tanti*, *tanta*, *tuto*, *tuta*, *tuti*, *tute*, *tota*, *toti*, *cotanto*.

l) Numerali: *do*, *doa*, *sé*, *sete*, *octo*, *quaranta*, *cento*, *mile*, *mili*, *anbi*.

m) Verbi:

- Forme con radicale francese che presentano desinenze flessive italiane o viceversa (es. *alirò*, *vene*, *spaventé*).
- Errori, ipercorrettismi o forme ipercaratterizzate in direzione del francese. Si veda, ad esempio, l'uso disinvolto dell'affisso *-oi-* in luogo di *-ai-* nelle forme *foit*, *soit*, *voit*, *oit*[51].
- Forme verbali italiane *tout court* (es. *guardò*, *consegue*, *dato*).

n) Avverbi:
- forme in *-a* tipiche dell'italiano settentrionale (es. *adoncha*).

[50] Cfr. Mascitelli 2020: 293.
[51] Cfr. Mascitelli 2020: 303.

o) Congiunzioni: *ma*, *che*, *quando*, *como*.

p) Preposizioni: *con*, *cum*, *cun*, *da*, *in*, *apreso*, *avanti*, *contra*, *entro*, *enverso*, *sença*, *defora*, *deverso*, *sovra*.

Quanto alla descrizione morfologica vera e propria, molte difficoltà di annotazione dipendono dalle frequenti irregolarità grammaticali del franco-italiano, di fronte alle quali occorre assumere dei criteri omogenei. A tale proposito, un primo aspetto problematico nell'annotazione delle *Enfances Bovo* è rappresentato dal mancato rispetto delle norme di accordo morfologico. In tali circostanze, per poter agire in modo uniforme, si è scelto di attribuire a determinanti e modificatori il genere, il numero e il caso del sostantivo a cui si accompagnano, come nell'esempio (11), dove il determinante *la*, pur presentando morfologia femminile, è descritto come maschile in base al genere del sostantivo italiano *boscho* a cui si riferisce.

(11) *Qe pitete fu de* ***la boscho*** *sevré,* (11.853).

Form	*POS*	*Morph*
la	DETdef	NOMB.=s|GENRE=m|CAS=r

Un simile criterio non rappresenta certamente una soluzione ottimale, ma è perlomeno coerente con il principio generale del *Cattex* già applicato al campo *POS*, secondo il quale l'annotazione deve avvenire in base al contesto sintattico. Inoltre, tale scelta consente in qualche misura di conservare memoria dell'irregolarità: ipotizzando, ad esempio, una ricerca per forme, la presenza di una forma femminile *la* descritta come maschile segnala la presenza di un errore di qualche natura.

Un ambito di annotazione particolarmente complesso è poi rappresentato dalla morfologia verbale. In tale contesto, infatti, l'interferenza tra i sistemi linguistici del francese e dell'italiano genera non poche ambiguità. Sovente capita, ad esempio, di incontrare forme che potrebbero appartenere all'una o all'altra lingua, come nel caso del già citato *dona* o di *monta* che potrebbero rappresentare tanto la terza persona del *passé simple* francese, quanto quella del presente italiano, oltre che del perfetto semplice veneto. In caso di simili ambiguità, se il contesto non è dirimente, la scelta di annotazione dell'*Enfances Bovo* prevede di attribuire la descrizione morfologica corrispondente alla forma francese.

Anche l'attribuzione corretta del tempo verbale è spesso problematica e rappresenta un aspetto dell'annotazione su cui intervenire in prossimi aggiornamenti del *corpus*. Emblematico in questo senso è il caso delle forme *oit*, *voit*, *soit*, *poit*[52], per il momento descritte come terza persona del tempo presente, con l'eccezione di *poit* più frequentemente interpretata come imperfetto.

[52] Sui problemi presentati da tali forme si vedano Capusso 1980: 23 e Mascitelli 2020: 303.

Quanto all'attribuzione della persona, occorre sottolineare che il criterio di concordanza sintattica assunto per l'attribuzione del numero nella morfologia nominale non vige in contesto verbale per la terza persona. Le forme di terza persona singolare usate secondo l'uso tipico dell'italiano settentrionale anche con riferimento a soggetti plurali sono infatti descritte come tali:

(12) *Bovo va pur davanti, li soldaer* ***va*** *dre;* (4.154)

Form	*POS*	*Morph*
va	VERcjg	MODE=ind|TEMPS=pst|PERS.=3|NOMB.=s

Nei tempi composti con ausiliare *avere*, il participio può concordare o meno per genere e numero con l'oggetto:

(13) *Monta a çival, si oit* ***trata*** *la spe,* (4.169)
(14) *Bovo oit* ***pris*** *una lança feré,* (4.212)

Qualora tuttavia la morfologia del participio presentasse ambiguità, la forma è stata descritta di *default* come maschile singolare.

Nel campo della morfologia nominale, richiede attenzione il caso dei sostantivi che ammettono il doppio genere, maschile e femminile. Per la descrizione di tali forme, nell'annotazione delle *Enfances Bovo*, ci si è avvalsi per quanto possibile delle informazioni fornite dal contesto. La presenza di determinanti o modificatori maschili o femminili, infatti, può costituire, in questi casi, un segnale dirimente per l'assegnazione del genere. Tuttavia, in assenza di chiari indizi contestuali, la descrizione morfologica ha cercato di operare scelte generalizzate. Così, ad esempio, i sostantivi *dolor* e *onor* sono stati trattati come femminili in quanto questa sembra l'opzione preferita dal dizionario *Tobler-Lommatzsch*, mentre le forme *oste* e *ost* sono state considerate maschili eccetto che in presenza di determinanti e modificatori femminili. Diverso è il caso di forme in cui l'alternanza di genere è determinata dall'interferenza con l'italiano, come il già citato caso di *amor*. In tal caso, nell'impossibilità di ricavare informazioni dal contesto, si opta per il genere della parola francese.

Accanto ai problemi intrinseci del franco-italiano, l'annotazione morfologica ha posto qualche questione anche in merito all'impiego delle etichette disponibili in *Pyrrha*. Un problema significativo, in tal senso, è presentato dai determinanti e dagli aggettivi possessivi per i quali l'annotazione prevista da *Phyrra* appare imprecisa e insufficiente. Data la disponibilità limitata di tre etichette per la persona (PERS.=1; PERS.=2; PERS.=3) e di un'unica etichetta per il numero (NOMB.=s; NOMB.=p), è impossibile fornire una descrizione completa che specifichi la persona (singolare o plurale) e l'accordo morfologico (singolare o plurale) della forma. In questi contesti, in attesa di potere adottare una soluzione più efficace, si è scelto di impiegare l'annotazione NOMB.=s/NOMB.=p esclusivamente per descrivere il numero della persona[53]:

(15) *Qi est quel, qi mena tel ferté,* |***Ses*** *armes a bicor pituré?* (4. 191-192)

Form	*POS*	*Morph*
ses	DETpos	PERS.=3|NOMB.=s|GENRE=f|CAS=n

5. *Contributi per l'implementazione di metodi e strumenti*

Accanto alla definizione di alcuni criteri di annotazione, il lavoro sulle *Enfances Bovo* ha permesso anche di testare i metodi e gli strumenti adottati per l'allestimento e la pubblicazione del *corpus*, contribuendo a suggerire e ad apportare diverse migliorie. I paragrafi seguenti illustrano brevemente gli interventi proposti rispetto ad ogni fase del lavoro.

5.1 *Interventi per la manipolazione del testo*

Come si è detto (cfr. § 3.1), la fase preparatoria nell'allestimento di un *corpus* consiste nel manipolare il testo al fine di renderlo analizzabile dalla macchina. A tale proposito, rispetto alle operazioni tecniche originariamente previste, l'annotazione delle *Enfances Bovo* ha fatto emergere l'esigenza di un cambiamento importante rispetto al trattamento di unità grafiche complesse in cui più forme siano tra loro combinate in un'unica parola grafica.

In un primo momento, infatti, tali forme sono state mantenute senza alcuna modifica, ritenendo di poterle descrivere grazie allo specifico *set* di etichette disponibili in *Pyrrha* (cfr. § 3.3). Tuttavia, in fase di lemmatizzazione, è apparso evidente che le possibilità di annotazione offerte dal programma non sono sufficienti a coprire l'intera gamma di combinazioni presenti nel testo della *Geste Francor*. Infatti, per soddisfare tale necessità mancano non solo appropriate etichette *POS*, ma anche lemmi adeguati alla rappresentazione di forme tra loro combinate. Inoltre, anche la rappresentazione morfologica appare poco soddisfacente, in quanto si applica non all'insieme delle forme, ma solo ad una di esse, secondo una modalità funzionale a descrivere la combinazione di una forma invariabile e di una forma flessa, ma non quella di due forme dotate di morfologia flessiva, come si vede rispettivamente in (16) e in (17), per la preposizione articolata e per la combinazione di forme pronominali.

(16) *Qi me donast tot li or* ***de·l*** *mon,* (1.34)

Form	*POS*	*Morph*
de·l	PRE.DETdef	NOMB.=s|GENRE=m|CAS=r

[53] Si tratta di una soluzione temporanea adottata per uniformità con gli altri testi lemmatizzati inclusi nel progetto. Una proposta alternativa, per il futuro, è offerta in seguito al § 5.2.

(17) *"Bovo," dist Druxiana, "e no* ***ve·l*** *vojo çeler;* […] (12.879)

Form	*POS*	*Morph*
ve·l	PROper.PROper	?

A fronte di tali difficoltà, per l'annotazione delle *Enfances Bovo* e di tutti i prossimi testi, si è deciso di operare una scelta unitaria e di spezzare tutte le forme complesse, comprese quelle correttamente descrivibili in *Pyrrha*, così da consentire la descrizione individuale di ciascuna componente. Poiché tale operazione comporta l'alterazione del testo proposto dall'edizione, al fine di conservare ogni informazione testuale, l'intervento viene segnalato dall'aggiunta di uno speciale marcatore corrispondente al simbolo 'cancelletto': #[54]. Così, ad esempio, forme come *avantiqe* o *tor·la* possono essere trasformate in *avanti# #qe* e *tor·# #la*, permettendo la descrizione di ciascuna unità:

(18) ***Avantiqe*** *qui de l'oste montese en auferant,* (3.134)

Form	*Lemma*	*POS*	*Morph*
Avanti#	*avanti*	ADVgen	SPEC=it|MORPH=empty
#qe	*que4*	CONsub	MORPH=empty

(19) *Q'el vegna a le a* ***tor·la*** *por muler.* (10.662)

Form	*Lemma*	*POS*	*Morph*
tor·#	*tolir*	VERinf	SPEC=it|NOMB.=x|GENRE=x|CAS=x
#la	*il*	PROper	PERS.=3|NOMB.=s|GENRE=f|CAS=r

Trattandosi ancora di una fase sperimentale, nel caso delle *Enfances Bovo,* l'introduzione della modifica appena descritta è stata piuttosto dispendiosa in termini di tempo, perché ha richiesto l'intervento manuale per la trasformazione di ogni singola forma tramite la funzione *Edit the form* di *Pyrrha.* Nel futuro, però, tale operazione potrà essere semplificata se integrata tra gli interventi preliminari da eseguire nella fase di manipolazione del testo prima del suo caricamento in *Pyrrha*. In tal modo la macchina potrà riconoscere le singole forme e proporre un'annotazione automatica da correggere poi manualmente.

[54] Un aspetto potenzialmente problematico rispetto alla soluzione proposta è rappresentato dalle forme contratte tipiche del francese, quali *au, du* etc., assenti nelle *Enfances Bovo*, ma ampiamente attestate in altri testi franco-italiani. Si ritiene tuttavia che il trattamento descritto sopra possa applicarsi anche in simili contesti scegliendo una suddivisione convenzionale delle forme: *a# #u*, *d# #u* da rinviarsi rispettivamente alle preposizioni *à* e *de* e al determinante *le*. Si tratta di una soluzione certamente discutibile che però ha il vantaggio di garantire un sistema di annotazione uniforme per tutti i casi di forme combinate, evitando di dover moltiplicare la creazione di etichette *ad hoc* per ogni specifica situazione e consentendone la corretta descrizione morfo-sintattica. Si noti peraltro che nel sito del *RIALFrI* il testo appare inalterato e l'analisi disgiunta delle singole componenti è offerta esclusivamente nella finestra di *pop-up* destinata alle informazioni lessicali e linguistiche (cfr. § 3.5).

5.2 *Proposte e contributi per l'annotazione del testo con* Pyrrha

L'uso di *Pyrrha* per la costruzione di un *corpus* lemmatizzato è vantaggioso e l'impiego delle *control list* di francese antico si è dimostrato in generale un valido supporto per l'annotazione automatica della *Geste Francor.* Alcune rigidità insite nell'uso delle liste di controllo manifestano però la necessità di integrazioni e modifiche tali da risolvere alcune criticità tanto nella descrizione delle forme in generale, quanto nel trattamento specifico dei testi franco-italiani.

Sul piano generale, ad esempio, sarebbe utile prevedere l'uso di un'etichetta da dedicare appositamente a quei passi lacunosi per i quali risulti impossibile fornire la corretta analisi della forma. La presenza di molte lacune nelle prime lasse delle *Enfances Bovo* ha manifestato con urgenza tale necessità. Un'opzione sarebbe certamente quella di escludere completamente dall'annotazione le forme dubbie cancellandole dal *corpus* (funzione *Delete the row)*. Tuttavia, in molti casi questa operazione comporterebbe una spiacevole perdita di informazioni: se non è possibile individuare il lemma o fornire una corretta descrizione grammaticale di una forma non significa, ad esempio, che non sia ricostruibile la sua categoria morfo-sintattica. L'introduzione di un'annotazione specifica per segnalare i contesti dubbi sarebbe dunque in tal senso risolutiva. Per ora, nelle *Enfances Bovo*, tale difficoltà è stata risolta in maniera artigianale sul file TSV tramite l'inserimento manuale di un punto di domanda (?), per indicare l'impossibilità ad indicare un lemma, e l'introduzione dell'informazione 'LACUNA' all'interno del campo *POS*, qualora non fosse determinabile neppure la parte del discorso. Il campo *Morph* è stato riempito invece in *Pyrrha* con l'etichetta MORPH=empty:

(20) *Do de Magançe fu retorné arer |E lasa* ***ato****… er* (2.41-42)

Form	*Lemma*	*POS*	*Morph*
ato	?	LACUNA	MORPH=empty

Un altro aspetto che necessita di modifica riguarda il caso già citato della descrizione dei possessivi. Come si è visto (cfr. § 4.5), le etichette morfologiche offerte da *Pyrrha* lasciano nell'ambiguità la definizione della persona e del numero. Al fine di risolvere tale criticità, si propone per il futuro di modificare l'annotazione morfologica dedicata alla persona integrandola con tre etichette aggiuntive: PERS.=4; PERS.=5; PERS.=6, corrispondenti rispettivamente alla prima, alla seconda e alla terza persona plurale[55]. In tal modo, sarebbe possibile sciogliere l'ambiguità delle etichette originarie rappresentando i possessivi come nell'esempio seguente.

[55] La scelta di una tale soluzione implica necessariamente, quale ulteriore modifica, l'esclusione del valore NOMB (numero) dalla stringa di descrizione morfologica dei pronomi personali prevista da *Pyrrha*. Di conseguenza, la forma *nos* in *Sor un de nos i sont plus de çant* (v.98) sarebbe descritta nel campo *Morph* come PERS.=4|GENRE=m|CAS=r.

(21) *Qe* ***nostri*** *çivaler ont reduti aré* (4.178)

Form	*POS*	*Morph*
nostri	DETpos	SPEC=it|PERS.=4|NOMB.=p|GENRE=m|CAS=r

La questione più urgente per l'annotazione del franco-italiano riguarda però la necessità di ricondurre alcune forme a lemmi italiani. Come si è visto (cfr. § 4.3), fino ad ora la lemmatizzazione delle forme italiane è potuta avvenire soltanto manualmente sul *file* estratto da *Pyrrha*. Ciò costituisce però un'operazione dispendiosa in termini di tempo e rischiosa per l'inserimento di errori umani. La ricerca di una soluzione al problema è dunque della massima rilevanza, specialmente nel caso di testi pesantemente italianizzati come la *Geste Francor*. Pertanto l'annotazione delle *Enfances Bovo* è stata l'occasione per mettere in atto un'importante implementazione della *control list* del franco-italiano. Cercando la collaborazione dei tecnici dell'*École des Chartes*, si è richiesta, infatti, la creazione di una *control list* di lemmi specifica per il franco-italiano, in grado di includere, oltre ai lemmi del francese antico, anche l'elenco delle entrate lessicali pubblicate nel lemmario generale del *TLIO* e gentilmente concesso per l'utilizzo in *Pyrrha*. Si apre così un'incoraggiante prospettiva per la lemmatizzazione del franco-italiano che, se tutto procede bene, potrà presto usufruire dei vantaggi di un'annotazione automatica e di liste di controllo per la correzione manuale più adeguate alle sue esigenze.

5.3 *La presentazione dei risultati: alcuni interventi*

La pubblicazione dei risultati della lemmatizzazione delle *Enfances Bovo* ha offerto l'opportunità di migliorare anche alcuni aspetti relativi alla rappresentazione dei risultati sul sito del *RIALFrI.* Gli interventi si sono concentrati in particolare sulla visualizzazione dei risultati nella finestra di *pop-up* che si apre cliccando sulle singole forme del testo[56]. La modifica ha riguardato specificamente l'annotazione *POS*. Considerato, infatti, che l'assegnazione delle forme alle diverse categorie morfo-sintattiche avviene su base contestuale e che dunque l'appartenenza di una forma a questa o quella parte del discorso non è soltanto una proprietà intrinseca del lemma, ma corrisponde alla funzione specifica che esso assume nel particolare contesto sintattico, è parso opportuno, rispetto ad una prima versione, spostare tale informazione dal campo *Lemma* al campo *Analisi*, fornendo una rappresentazione coerente con il seguente schema:

[56] Il testo lemmatizzato delle *Enfances Bovo* è disponibile sul sito del *RIALFrI* all'indirizzo http://www.rialfri.eu/rialfriPHP/public/testo/testo/codice/rialfri%7CgesteFrancor_1%7C001.

Forma-
Lemma
Analisi *POS; MORPH*

Pertanto, la rappresentazione dell'annotazione di una forma come *lion* al v. 3 delle *Enfances Bovo* è ora raffigurata come di seguito:

(22) *Ver lui s'en voit cosi fer cun* ***lion****,*

lion
Lemma LÏON **Analisi** Nome comune, singolare, maschile, nominativo

Tale schema rappresentativo, oltre ad essere formalmente più corretto, si dimostra funzionale anche alla compilazione delle voci del dizionario. Attraverso una semplice ricerca, si potrà infatti agevolmente risalire alle diverse funzioni svolte dai singoli lemmi nei vari contesti, consentendo così una descrizione più ampia e precisa delle proprietà morfo-sintattiche generalmente possedute da ciascun lemma.

Infine, accanto all'intervento appena descritto, per permettere agli utenti una più agevole fruizione dei risultati, si è deciso di sciogliere le etichette di annotazione utilizzate in *Pyrrha*, proponendone una traduzione italiana estesa, secondo le corrispondenze riportate poco sotto. In tal senso, la prova pratica di annotazione delle *Enfances Bovo* si è dimostrata utile per rendere più appropriato lo scioglimento di alcune sigle e per valutare l'opportunità di rappresentare o meno determinate informazioni[57].

[57] I campi lasciati in bianco nelle tabelle corrispondono alle informazioni escluse dalla rappresentazione nel *RIALFrI*.

POS tags	**Corrispondenza**
ADJcar	aggettivo numerale cardinale
ADJind	aggettivo indefinito
ADJord	aggettivo numerale ordinale
ADJpos	aggettivo possessivo
ADJqua	aggettivo qualificativo
ADVgen	avverbio
ADVing	avverbio interrogativo negativo
ADVint	avverbio interrogativo/esclamativo
ADVneg	avverbio di negazione
ADVsub	avverbio subordinante
CONcoo	congiunzione coordinante
CONsub	congiunzione subordinante
DETcar	determinante numerale cardinale
DETcom	determinante definito composto
DETdef	determinante definito
DETdem	determinante dimostrativo
DETind	determinante indefinito
DETint	determinante interrogativo/esclamativo
DETndf	determinante non definito
DETpos	determinante possessivo
DETrel	determinante relativo
INJ	interiezione
LACUNA	
NOMcom	nome comune
NOMpro	nome proprio
NUM	numero
OUT	
PRE	preposizione
PROadv	pronome avverbiale
PROcar	pronome numerale cardinale
PROdem	pronome dimostrativo
PROimp	pronome impersonale
PROind	pronome indefinito
PROint	pronome interrogativo
PROord	pronome numerale ordinale
PROper	pronome personale
PROpos	pronome possessivo
PROrel	pronome relativo
RED	
VERcjg	verbo
VERinf	verbo infinito
VERppa	verbo participio presente
VERppe	verbo participio passato

MORPH tags	**Corrispondenza**
CAS=i	*régime indirect*[58]
CAS=n	nominativo
CAS=r	*régime*
CAS=x	
DEGRE=-	
DEGRE=c	comparativo
DEGRE=p	
DEGRE=s	superlativo
GENRE=f	femminile
GENRE=m	maschile
GENRE=n	neutro
GENRE=x	
MODE=con	condizionale
MODE=imp	imperativo
MODE=ind	indicativo
MODE=sub	congiuntivo
MORPH=empty	
NOMB.=p	plurale
NOMB.=s	singolare
NOMB.=x	
PERS.=0	
PERS.=1	1ª persona
PERS.=2	2ª persona
PERS.=3	3ª persona
SPEC=it	italianismo
SPEC=lat	latino
TEMPS=fut	futuro
TEMPS=ipf	imperfetto
TEMPS=psp	perfetto
TEMPS=pst	presente

6. *Conclusioni*

L'annotazione delle *Enfances Bovo* ha permesso di testare la validità degli strumenti di lemmatizzazione scelti dal *DiFrI* su un testo che presenta un altissimo grado di mescidanza linguistica. L'esperienza ha dimostrato che l'impiego di uno

[58] Per le sigle CAS=i e CAS=r si è ritenuto di riportare eccezionalmente la corrispondente traduzione francese: *régime indirect* e *régime*.

strumento di lemmatizzazione automatica progettato, come *Pyrrha*, per il francese antico offre un supporto molto valido che tuttavia richiede di essere aggiustato assecondando le esigenze specifiche del franco-italiano. In tal senso, la creazione di una *control list ad hoc* inclusiva di lemmi italiani e dialettali e l'introduzione di etichette di annotazione dedicate costituiscono una necessità di primaria importanza.

D'altro canto, la pesante infiltrazione dell'ingrediente italiano della *Geste Francor* ha reso evidente l'enorme complessità rappresentata dalla lemmatizzazione del franco-italiano, mostrando l'urgenza di definire principi di lemmatizzazione e annotazione quanto più possibile uniformi. In tale prospettiva, i criteri elaborati per le *Enfances Bovo* contribuiscono a delineare alcune linee guida valide per la lemmatizzazione dei prossimi testi.

L'implementazione di strumenti e metodi di lavoro richiede ancora molto impegno. Tuttavia, la pubblicazione sul *RIALFrI* del primo testo lemmatizzato della *Geste Francor* costituisce, pur nella provvisorietà dei suoi risultati, un passo importante, che permette di aprire il confronto con la comunità scientifica, rendendo sempre più concreta la possibilità di realizzare un *corpus* lemmatizzato completo quale risorsa preziosa per lo studio lessicale e linguistico della letteratura franco-italiana.

7. *Appendice*

I. *Lemmi italiani*

Si riporta di seguito l'elenco delle forme ricondotte ad un lemma italiano. Dove non diversamente specificato, i lemmi sono attribuiti sulla base del *TLIO*.

FORME RICONDOTTE AL LEMMA ITALIANO	
FORME	**LEMMA**
adoncha	*adunque*
albergo	*albergo*
altro, *altra*, *altri*	*altro*
amigo, *amisi*	*amico*
ancoi	*ancoi*
ancora	*ancora*
ani	*anno*
apreso	*appresso*
aradegé	*eradegar*
arpa	*arpa*
asaçer	*assapere*

asaçon	*assaggiare*
aspeté, *aspeter*	*aspettare*
atenti	*attento*
avanti	*avanti*
bagno	*bagno*
bagordi	*bagordo*
banco	*banco*
boscho	*bosco*
brando	*brando*
burgi	*borgo*
canpo	*campo*
cason	*cagione*
cento	*cento*
certo	*certo*
colpo, *colpi*	*colpo*
colu	*colui*
como	*come*
con, *cun*, *cum* (preposizione)	*con*
conpagni	*compagno*
contra	*contro*
corno	*corno*
corona	*corona*
corpo	*corpo*
corte	*corte*
cosi	*così*
cotanto, *cotant,*	*cotanto*
da	*da*
dama	*dama*
dardo	*dardo*
dato	*dare*
davanti	*davanti*
dè	*dare*
defora	*difuori*
deo	*dio*
desovra	*disopra*

desghiré	*dischierare*
doncha	*dunque*
donde	*donde*
ecote	*ecco*
elmo	*elmo*
enparer	*imparare*
entarder	*intardare*
entro	*entro*
erba	*erba*
fanti	*fante*
fantin	*fantino*
femena	*femmina*
fiolo	*figliuolo*
fonte	*fonte*
fora	*fuori*
forte	*forte*
ge/g'	*ci*
groso	*grosso*
guera	*guerra*
homo	*uomo*
ideo	*dio*
in	*in*
inparè, inparò	*imparare*
intarder	*intardare*
inverno	*inverno*
leto, leito	*letto*
li	*lido*
li (pronome avv.)	*gli*
lì	*lì*
luna	*luna*
lungo	*lungo*
ma	*ma*
martirio	*martirio*
meço, me	*mezzo*
mejo	*meglio*

meno	*meno*
mesi	*messo*
mille	*mille*
miracolo	*miracolo*
mo	*modo*
molto	*molto*
mondo	*mondo*
monti	*monte*
morte	*morte*
mura	*muro*
nean	*neanche*
nian	*neanche*
nome	*nome*
novella, *novela*	*novella*
octo	*otto*
ora	*ora*
oste	*oste*
parola	*parola*
pena	*pena*
perigolo	*pericolo*
perigoloso	*pericoloso*
pié	*pigliare*
plaga	*piaga*
poco	*poco*
poma	*poma*
ponte	*ponte*
porta	*porta*
porti	*porto*
pur	*pure*
qelo	*quello*
qual	*quale*
qualche	*qualche*
quando	*quando*
quanto	*quanto*
quel (dimostrativo), *quello*, *quella*, *quela*	*quello*

questo, *questa*	*questo*
qui (avverbio)	*qui*
qui	*quello*
quisti	*questo*
quilois	*quiloga*
regno	*regno*
rico	*ricco*
roba	*roba*
sala	*sala*
salvamento	*salvamento*
sangue	*sangue*
sano	*sano*
santo	*santo*
schané	*scannare*
se (impersonale)	*si* (*GDLI*)
segno	*segno*
sella	*sella*
selva	*selva*
sença	*senza*
seno	*senno*
senpre, *sempre*	*sempre*
servisio	*servizio*
sete	*sette*
soldo	*soldo*
soto	*sotto*
sovente	*sovente*
sovra	*sopra*
spala	*spalla*
suso	*suso*
tanto, *tanta*	*tanto*
tempo, *tenpo*	*tempo*
tera	*terra*
tore	*torre*
tosto	*tosto*
tradimento	*tradimento*

tropo	*troppo*
tutti	*tutto*
unde	*onde*
unguento, *unguenti*	*unguento*
verde	*verde*
verità	*verità*
versi	*verso*
verso (preposizione)	*verso*
via (avverbio)	*via*
via (nome)	*via*
vita	*vita*
vivo	*vivo*
viso	*viso*
voluntà	*volontà*

II. *Nomi propri*

Di seguito è riportato l'elenco dei nomi propri contenuti nelle *Enfances Bovo* con la relativa lemmatizzazione. Se non diversamente specificato, i nomi propri sono sempre ricondotti alla forma di citazione proposta dal repertorio di Moisan (1986). Le forme con asterisco (*) assumono come lemma la forma indicata nel volume dedicato ai nomi propri citati nelle opere straniere. Alcune forme rimangono tuttavia incerte (?).

Nomi propri di persona	
Forme	**Lemma**
Apolin	*Apolin*
Armenion	*Hermin*
Blondoja, *Blondoie*, *Blionda*	*Blondoia**
Bovo, *Bovon*, *Boves*	*Bueves*
Braidamont, *Bradamont*	*Braidamont**
Clarença	*Clarence*
Dodo, *Does*, *Do*	*Doon*
Drixiana, *Druxiana*, *Druxiane*, *Drusiane*	*Drusiana**
Garner, *Guarner*	*Garnier* (?)
Gujon, *Gui*	*Gui*
Jesu	*Jesus*

Latro	*Latro*
Luchafer	*Luchafer*
Machabrun	*Macabrun*
Macon, *Macometo*	*Mahon*
Marie	*Marie*
Oliver	*Oliver* (?)
Oria	*Orie2**
Pipin	*Pepin*
Pulicant, *Pulican*	*Pulican**
Rondel	*Arondel* (?)
Simon, *Symon*	*Simon**
Synibaldo, *Siginbaldo*, *Sinibaldo*	*Sinibaldo**
Teris	*Thiery*
Uberto	*Uberto* (?)

Nomi di luogo	
Forme	**Lemma**
Antone, *Ntone*, *Antona*	*Hamtone*
Arminie, *Arminia*	*Ermenie*
Baiver	*Baiviere*
Beniant	*Beleant*
França, *Françe*, *France*	*France*
Magança, *Magançe*, *Maganc*	*Maience*
Syndonia, *Syndonie*, *Sydonia*	*Sidoine*

Bibliografia

I. Opere

Geste Francor (ed. Morgan)

La Geste Francor*, edition of the chansons de geste of MS. Marc. Fr. XIII (=256)*, with glossary, introduction, and notes by Leslie Zarker Morgan, 2 voll., Tempe, Arizona Center for Medieval and Renaissance Studies, 2009.

II. Studi e strumenti

Barbato 2015
Marcello Barbato, *Il franco-italiano: storia e teoria*, in «Medioevo Romanzo», 39 (2015), pp. 22-51.

Beretta 2011
Carlo Beretta, Recensione a *Geste Francor* (ed. Morgan), in «Medioevo Romanzo», 35/1 (2011), pp. 196-199.

Boerio 1998 [1856]
Giuseppe Boerio, *Dizionario del dialetto veneziano*, Seconda edizione aumentata e corretta aggiuntovi l'indice italiano veneto già promesso dall'autore nella prima edizione, Venezia, Cecchini, 1856 [ripr. facs. Firenze, Giunti, 1998].

Busa 1987
Roberto Busa, *Fondamenti di informatica linguistica*, Milano, Vita e Pensiero, 1987.

Camps – Albarran – Cochet – Ing 2019
Jean-Baptiste Camps, Elena Albarran, Alice Cochet, Lucence Ing, *Jean-Baptiste-Camps/Geste. Geste: un corpus de chansons de geste, 2016-…*, 5 aprile 2019, https://doi.org/10.5281/zenodo.2630574 [cons. 24.VII.2021].

Capusso 1980
Maria Grazia Capusso, *La lingua del* Divisament dou monde *di Marco Polo*, I. *Morfologia verbale*, Pisa, Pacini, 1980 («Biblioteca degli Studi Mediolatini e Volgari», nuova serie, 5).

Capusso 2007
Maria Grazia Capusso, *La produzione franco-italiana dei secoli XIII e XIV: convergenze letterarie e linguistiche* in *Plurilinguismo letterario*, a cura di Renato Oniga e Sergio Vatteroni, Soveria Mannelli, Rubbettino, 2007, pp. 159-204.

Clérice – Pilla – Camps 2019
Thibault Clérice, Julien Pilla, Jean-Baptiste Camps, *hipster-philology/pyrrha: 2.1.0*, 1 novembre 2019, https://doi.org/10.5281/zenodo.2325427.

Clérice – Camps 2021
Thibault Clérice, Jean-Baptiste Camps, *chartes/deucalion-model-af: 0.4.0Alpha (Version 0.4.0a)*, 4 giugno 2021, http://doi.org/10.5281/zenodo.4898954.

DMF
Dictionnaire du Moyen Français (1330-1500), version 2020 (DMF 2020), ATILF – CNRS – Université de Lorraine, http://www.atilf.fr/dmf/.

Flutre 1962
Louis-Fernand Flutre, *Table des noms propres avec toutes leurs variantes figurant dans les romans du Moyen Âge écrits en français ou en provençal et actuellement publiés ou analysés*, Poitiers, Centre d'Études Supérieures de Civilisation Médiévale, 1962.

Gambino 2016
Francesca Gambino, *Code-mixing nel* Bovo d'Antona *udinese, con una nuova edizione del frammento Udine, Archivio Capitolare, Fondo Nuovi manoscritti 736.28*, in «Francigena», 2 (2016), pp. 35-130.

Gambino 2020
Francesca Gambino, *Il Dizionario del Franco-Italiano (*DiFrI*). La definizione del corpus, le coordinate spazio-temporali, le prime voci*, 31 marzo 2020, https://www.rialfri.eu/rialfriWP/introduzione.

GDLI
Grande dizionario della lingua italiana, iniziato da Salvatore Battaglia, continuato e concluso da Giorgio Bàrberi Squarotti, 21 voll., Torino, UTET, 1961-2002.

Giannini 2012
Gabriele Giannini, Recensione a *Geste Francor* (ed. Morgan), in «Romania», CXXX/519-520 (2012), pp. 505-507.

Guillot – Prévost – Lavrentiev 2013a
Céline Guillot, Sophie Prévost, Alexei Lavrentiev, *Principes d'annotation Cattex09 (Version 2.0)*, Lyon, École normale supérieure de Lyon, 8 aprile 2013, http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf.

Guillot – Prévost – Lavrentiev 2013b
Céline Guillot, Sophie Prévost, Alexei Lavrentiev, *Manuel de référence du jeu Cattex09, (Version 2.0)*, Lyon, École normale supérieure de Lyon, 8 aprile 2013, http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf.

Holtus 1981
Günther Holtus, *Etimologia e lessico franco-italiano*, in *Etimologia e lessico dialettale*. Atti del XII Convegno per gli Studi Dialettali Italiani (Macerata, 10-13 aprile 1979), Pisa, Pacini, 1981, pp. 153-163.

Manjavacas – Kestemont – Clérice 2018
Enrique Manjavacas, Mike Kestemont, Thibault Clérice, *emanjavacas/pie v0.1.0 (Version v0.1.0)*, 28 novembre 2018, http://doi.org/10.5281/zenodo.1637878.

Mascitelli 2020
Cesare Mascitelli, *La* Geste Francor *nel cod. marc. V13. Stile, tradizione, lingua,* Strasbourg, Éditions de linguistique et de philologie, 2020.

Moisan 1986
André Moisan, *Répertoire des noms propres de personnes et de lieux cités dans les Chansons de Geste françaises et les œuvres étrangères dérivées*, Genève, Droz, 1986.

Morlino 2010
Luca Morlino, *Contributi al lessico franco-italiano*, in «Medioevo letterario d'Italia», 7 (2010), pp. 65-85.

Pinche 2019
Ariane Pinche, *Annoter facilement un corpus complexe. L'exemple de Pyrrha, interface de post correction, et Pie, lemmatiseur et tagueur morphosyntaxique, pour l'ancien français*, in *Actes des Rencontres lyonnaises des jeunes chercheurs en linguistique historique*, éditées par Timothée Premat, Ariane Pinche, Lyon, 2019 («Diachronies Contemporaines»), pp. 48-58.

Rajna 1998
Pio Rajna, *La rotta di Roncisvalle nella letteratura cavalleresca italiana*, in *Scritti di filologia e linguistica italiana e romanza*, a cura di Guido Lucchini, premessa di Francesco Mazzoni, introduzione di Cesare Segre, 3 voll., Roma, Salerno Editrice, 1998 («Pubblicazioni del "Centro Pio Rajna"», sez. II/1), vol. I, pp. 190-369.

Renzi 1970
Lorenzo Renzi, *Per la lingua dell'Entrée d'Espagne*, in «Cultura neolatina», 30 (1970), pp. 59-87.

RIALFrI
Repertorio Informatizzato dell'Antica Letteratura Franco-Italiana, diretto da Francesca Gambino, Dipartimento di Studi Linguistici e Letterari, Università degli Studi di Padova, http://www.rialfri.eu/ [cons. 10.VII.2021].

Salvi – Renzi 2010
Grammatica dell'italiano antico, a cura di Gian Paolo Salvi e Lorenzo Renzi, 2 voll., Bologna, il Mulino, 2010.

TLIO

*Tesoro della Lingua Italiana delle Origin*i, fondato da Pietro G. Beltrami e continuato da Lino Leonardi, diretto da Paolo Squillacioti, http://tlio.ovi.cnr.it/TLIO/ [cons. 15.VII.2021].

TL

Altfranzösisches Wörterbuch, Adolf Toblers nachgelassene Materialien bearbeitet und hrsg. von Erhard Lommatzsch, weitergeführt von Hans Helmut Christmann, vollendet von Richard Baum und Willy Hirdt unter Mitwirkung von Brigitte Frey, 11 voll., Berlin – Wiesbaden, Weidmannsche Buchhandlung – Steiner, 1925-2002.

Tomasi 2008

Francesca Tomasi, *Metodologie informatiche e discipline umanistiche*, Roma, Carocci, 2008.

Vidossi 1956

Giuseppe Vidossi, *L'Italia dialettale fino a Dante*, in *Le Origini. Testi latini, italiani, provenzali e franco-italiani*, a cura di Antonio Viscardi, Bruno e Tilde Nardi, Giuseppe Vidossi e Felice Arese, Milano-Napoli, Ricciardi, 1956, pp. XXXIII-LXXI.

Viscardi 1941

Antonio Viscardi, *Letteratura franco-italiana*, Modena, Società Tipografica Modenese, 1941.

Wunderli 2003

Peter Wunderli, *Franko-Italienisch: ein sprach- und literaturgeschichtliches Kuriosum*, in «Vox Romanica», 62 (2003), pp. 1-27.