Francigena

7 (2021)

L'analisi lessicale *dell'Entrée d'Espagne*: bilancio di una prima sperimentazione

Floriana Ceresato
(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 2420-9767

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

info@francigena-unipd.com

INDICE

https://doi.org/10.25430/2420-9767/V7-009
DOI: 10.25430/2420-9767/V7-009

In ricordo di Simon Gaunt

L'analisi lessicale dell'*Entrée d'Espagne*: bilancio di una prima sperimentazione

Floriana Ceresato
floriana.ceresato@gmail.com

(Università degli Studi di Padova)

Abstract:
L'*Entrée d'Espagne*, capolavoro della letteratura franco-italiana, è stata scelta come banco di prova per testare l'efficacia di un'analisi lessicale semi-automatica, realizzata mediante il *tagger Pie* e l'interfaccia di post-correzione *Pyrrha*. Il presente contributo ripercorre le varie tappe del lavoro svolto e illustra i risultati ottenuti.

The *Entrée d'Espagne*, masterpiece of Franco-Italian literature, was chosen to test the efficacy of a semi-automatic lexical analysis program realized by using the Pie tagger and the Pyrrha post-correction interface. This paper retraces the different stages of the experimentation and illustrates the final results.

Keywords:
Franco-italiano – lemmatizzazione – annotazione morfosintattica – *Pyrrha* – *Entrée d'Espagne.*
Franco-italian – lemmatization – morpho-syntactic annotation – *Pyrrha* – *Entrée d'Espagne.*

1. *Introduzione*

L'idea di applicare l'analisi lessicale semi-automatica all'opera manifesto della letteratura medievale franco-italiana, l'*Entrée d'Espagne*, nasce da un duplice proposito: comprendere in quale misura gli strumenti a nostra disposizione per l'antico francese si adattino allo studio di una lingua storica mescidata e non normata; condividere dati utili ad approfondire la conoscenza del patrimonio letterario franco-italiano, rendendolo maggiormente fruibile.

Per questa prima sperimentazione che, dopo la costituzione di un ampio repertorio informatizzato di testi franco-italiani, apre la seconda fase del progetto *RIALFrI*[1], ci si è serviti di *Pie*[2] e di *Pyrrha*[3], rispettivamente un *tagger*[4] e un'inter-

[1] Il *Repertorio Informatizzato dell'Antica Letteratura Franco-Italiana* è un progetto diretto da Francesca Gambino presso il Dipartimento di Studi Linguistici e Letterari dell'Università degli Studi di Padova.
[2] Cfr. Manjavacas – Kádár – Kestemont 2019; Manjavacas – Clérice – Kestemont 2021.
[3] Cfr. Camps – Clérice – Pinche 2020; Clérice – Pilla – Camps 2019.
[4] Un *tagger* (o 'disambiguatore morfosintattico') è un programma che esplicita la categoria grammaticale di ogni forma presente in un testo, basandosi sull'analisi dei contesti sintattici. Per ulteriori approfondimenti si rimanda a Abeillé 2003.

faccia di post-correzione. Da essi siamo partiti per verificare l'efficacia di un determinato tipo di analisi lessicale, documentandone pro e contro. Il modello e gli strumenti sui quali ci basiamo attualmente funzionano anche per il franco-italiano? E, in caso di risposta negativa, sarebbe sufficiente rimaneggiarli o diventerebbe necessario sviluppare un nuovo paradigma? A prescindere dalle risposte che immaginavamo a queste domande e dai risultati effettivi che poi abbiamo conseguito, ci è sembrato importante proporre un esempio completo di analisi lessicale di un testo franco-italiano non solo come obiettivo compiuto in sé, ma anche come obiettivo intermedio da cui ripartire per ulteriori ricerche. Il materiale raccolto durante il nostro studio, infatti, può costituire un interessante bagaglio di informazioni strutturate per chi volesse intraprendere approfondimenti linguistici, lessicografici, stilistici.

Nella trattazione che seguirà, per indicare il lavoro complessivo svolto sul testo dell'*Entrée d'Espagne*, si è preferito ricorrere all'espressione generica 'analisi lessicale' perché onnicomprensiva di tutte le singole operazioni effettuate, che qui citiamo[5]:

1. Tokenizzazione: la segmentazione del testo in *tokens* (o lessemi), le unità di base del testo digitale sulle quali si basano i successivi livelli di elaborazione[6].
2. Lemmatizzazione: il ricondurre il *token* al rispettivo esponente lessicale.
3. *POS* (*Part Of Speech*) *tagging*: l'annotazione della categoria lessicale del *token*.
4. *MSD* (*Morphosyntactic Description*) *tagging*: l'annotazione morfosintattica del *token*.
5. Post-correzione: la revisione dei risultati ottenuti nelle operazioni precedenti.

Segnaliamo che le prime quattro fasi dell'analisi lessicale sono state demandate all'azione di *Pie* e, quindi, sono state realizzate simultaneamente e in modo automatico. L'intervento manuale, analitico e graduale, è subentrato solo nella fase di post-correzione e revisione finale eseguita in *Pyrrha*.

1.1. *Gli strumenti:* Pie *e* Pyrrha

Pyrrha è un'interfaccia per la visualizzazione e la post-correzione manuale dell'analisi lessicale automatica svolta da *Pie*, un *tagger* costruito su algoritmi di

[5] Si aggiunga un'ulteriore precisazione terminologica. Definiamo 'forme' o 'vocaboli' le unità lessicali presenti nel testo nella loro forma flessa, coniugata o invariabile; 'occorrenze' le ripetizioni di una forma nel testo; 'lemmi' le parole ricondotte al rispettivo esponente lessicale del dizionario.
[6] La definizione è ripresa da Lenci – Montemagni – Pirrelli 2020: 102, volume al quale si rimanda per ulteriori approfondimenti riguardo il processo di tokenizzazione e l'interpretazione di *token* come unità atomica dell'analisi linguistica.

Machine Learning[7]. *Pie* si basa sull'apprendimento automatico di tipo induttivo ed è svincolato dai problemi riscontrati generalmente nei *taggers* che si basano su un dizionario o su un *set* di regole predefinite[8]. Di conseguenza, *Pie* è indipendente dalla lingua trattata e da tutte le varianti che essa può presentare. *Pie* apprende la lingua comparando i risultati che ottiene su un *corpus* di prova con quelli ricavati da un *corpus* di addestramento precedentemente annotato a mano. Estrae inoltre il modello organizzativo dei dati, che potrà essere in seguito utilizzato per altri *corpora* simili. Più aumentano la quantità e la varietà dei testi, maggiori diventano l'affidabilità e la precisione di *Pie*.

Concepita e sviluppata presso l'*École Nationale des Chartes* dall'*équipe Humanités Numériques*, l'interfaccia *Pyrrha* è liberamente accessibile in rete, previa registrazione e creazione di un *account* personale, in due modalità. La prima mediante il sito https://dev.chartes.psl.eu/pyrrha/, che ospita il cosiddetto *development environment*, ovvero una versione di prova per esercitarsi, nella quale i materiali caricati e il lavoro svolto non vengono salvati sul lungo termine; la seconda tramite il sito https://dh.chartes.psl.eu/pyrrha/, che invece ospita il cosiddetto *production environment* e salva i dati in modo perenne[9]. Una volta effettuato l'accesso, si procede cliccando su *New Corpus* nella barra di comando in alto e si eseguono alcune operazioni preliminari, che consistono nel:

- nominare il proprio *corpus*[10] e stabilire l'ampiezza del passaggio testuale da visualizzare accanto al vocabolo da analizzare (fig. 1);
- impostare l'assetto della pagina di lavoro (fig. 2);
- importare il proprio *corpus*[11] e scegliere il modello di analisi lessicale[12] (fig. 3);
- scegliere la lista di controllo da utilizzare[13] (fig. 4).

[7] Cfr. Pinche 2019: 50.

[8] Come accade, ad esempio, per il *tagger LGeRM* dell'*ATILF*.

[9] È possibile richiedere la cancellazione dei dati in qualsiasi momento, contattando i gestori del sito. I contenuti che ogni utente crea sono collegati esclusivamente al suo *account*, quindi non risultano accessibili ai gestori o ad altri utilizzatori, a meno che l'utente stesso non lo consenta attraverso un invito a collaborare ad uno o più *corpora* in qualità di *user* o di *administrator* (in tal caso il collaboratore potrà intervenire anche sulla *control list*).

[10] Con il termine 'corpus' in *Pyrrha* si indica in modo neutro il materiale testuale importato a prescindere dalla sua lunghezza e dalla sua composizione; a titolo esemplificativo, il testo di una *chanson de geste* e un compendio di poesie, tratte da canzonieri diversi, per *Pyrrha* risultano equivalenti perché contenuti nel singolo *file* che si sceglie di caricare.

[11] Con un semplice copia-incolla del testo puro. Se il testo caricato non è già tokenizzato, *Pyrrha* consente di eseguire anche la tokenizzazione.

[12] Attualmente *Pyrrha* dispone di cinque modelli: *Ancient Greek*, *Français Classique (Modernisé)*, *Français Classique (Non Modernisé)*, *Old French*, *Latin* (modello elaborato dal *LASLA*, *Laboratoire d'Analyse Statistique des Langues Anciennes* dell'Università di Liegi).

[13] La lista di controllo, così come il modello di analisi lessicale, si riferisce alla lingua del *corpus* analizzato. Costituisce una sorta di 'grammatica' del paradigma che viene applicato al *corpus* e si

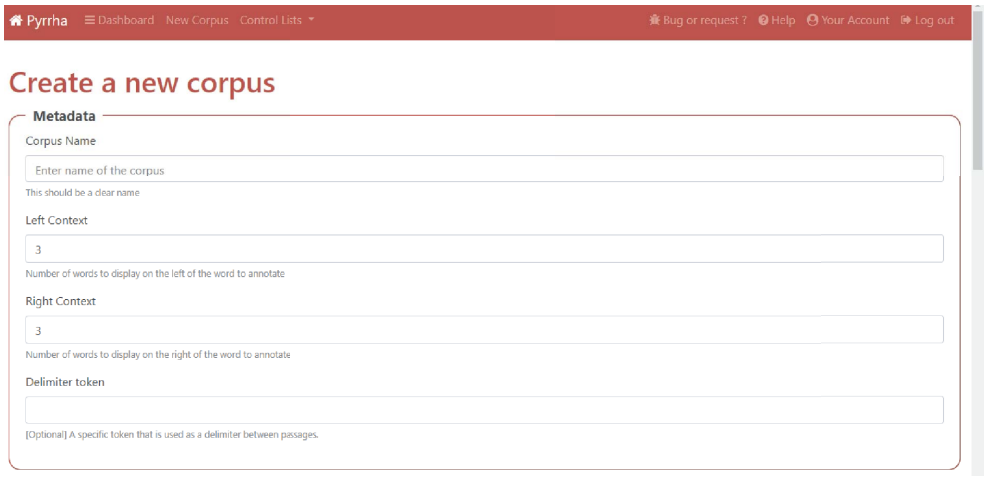

Fig. 1. Denominazione del *corpus* e delimitazione del contesto da visualizzare.

Fig. 2. Impostazione dell'assetto della pagina di lavoro.

Fig. 3. Importazione del *corpus* e scelta del modello di analisi lessicale.

compone a sua volta di tre liste specifiche: la lista dei lemmi (o dizionario di riferimento), la lista delle etichette *POS*, la lista delle annotazioni *Morph*. Se non si trova una *control list* adatta tra quelle proposte (*Ancien Français – École des Chartes*, *Ancien Occitan*, *Classical Latin (LASLA-Derived)*, *Français modernisé*, *Français non modernisé*), *Pyrrha* prevede anche l'opzione *Write your own*. In quest'ultimo caso la lista di controllo sarà privata, quindi accessibile solo al creatore del *corpus* e, se previsti, agli utenti associati come collaboratori. Il responsabile della lista possiede comunque la facoltà di proporla ai gestori di *Pyrrha* al fine di renderla pubblica come le altre cinque già presenti ed utilizzabili da tutti. Anche eventuali modifiche o integrazioni alle liste di controllo pubbliche devono essere sottoposte alla moderazione dei gestori.

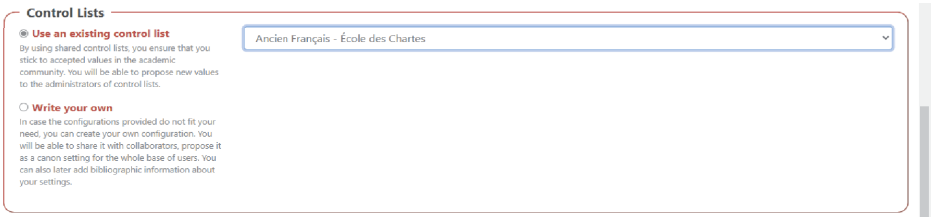

Fig. 4. Scelta della lista di controllo.

Una volta impostati tutti i criteri, si procede ad avviare l'analisi lessicale automatica, il cui esito compare in una nuova finestra dedicata alla post-correzione manuale.

La schermata che si apre all'utente (fig. 5), e che corrisponde alla funzione *Correct tokens* del menù visibile sulla sinistra dello schermo, è costituita da una tabella composta da nove colonne, ognuna corrispondente ad una categoria:

1. *Id*: il numero d'identificazione attribuito a ciascun *token*;
2. *Form*: il *token* come appare nel testo[14];
3. *Lemma*: il lemma associato al *token*;
4. *POS* (*Part Of Speech*): l'etichetta riferita alla categoria lessicale;
5. *Morph*: l'etichetta morfosintattica;
6. *Context*: il contesto testuale nel quale compare il *token*;
7. *Similar*: il numero di *token* presenti nel *corpus* equivalenti al *token* che si sta analizzando[15];
8. *Save*: il salvataggio delle modifiche apportate all'etichettatura;
9. + : le opzioni di modifica: correzione di una forma errata (*Edit the form*), cancellazione (*Delete the row*) o aggiunta di un *token* (*Add a token after this one*)[16].

[14] Ovvero il tradizionale 'forma' o 'vocabolo', come specificato nella nota 5.

[15] Ovvero le tradizionali 'occorrenze', come specificato nella nota 5. Questa funzione facilita e velocizza la correzione perché rimanda all'elenco completo dei *tokens* che presentano la medesima annotazione e si riferiscono allo stesso lemma: l'utente può stabilire se applicare la correzione effettuata a tutti i *tokens* individuati oppure ad una selezione di *tokens* all'interno della lista visualizzata a schermo.

[16] Tali opzioni di modifica consentono di correggere il testo qualora si riscontrassero, ad esempio, dei refusi dovuti al riconoscimento ottico dei caratteri (cfr. nota 27) utilizzato preliminarmente per digitalizzare il documento.

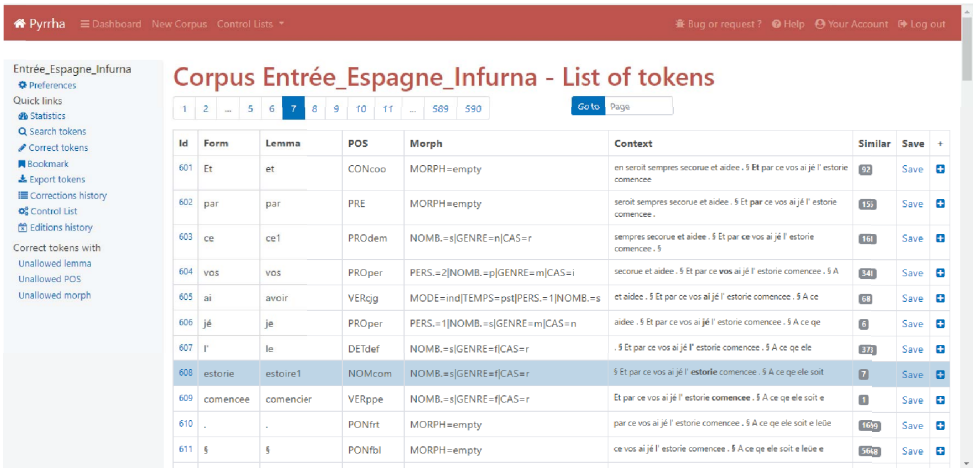

Fig. 5. Schermata di lavoro di *Pyrrha*.

L'utente scorre la tabella con l'elenco dei *tokens* e delle rispettive annotazioni (100 per ogni pagina) e, in caso di errore, può intervenire direttamente sul *token* stesso (colonna 9), sul lemma (colonna 3), sulla categoria lessicale (colonna 4), sull'etichetta morfosintattica (colonna 5). È sufficiente iniziare a digitare le prime lettere di un'etichetta o di un lemma e *Pyrrha* suggerisce in automatico le soluzioni disponibili. Ogni modifica deve essere salvata cliccando su *Save*. Le righe che si colorano di azzurro indicano che l'intervento manuale è stato salvato correttamente; se invece la correzione inserita non è contemplata nella lista di controllo del *corpus*, la casella si colora di rosso e non si può procedere al salvataggio dei dati[17].

Pyrrha presenta anche una finestra dedicata all'esplorazione del *corpus* (*Search tokens*), nella quale è possibile effettuare ricerche per *token*, per lemma, per *POS*, per morfologia o per più criteri combinati tra loro[18]. Inoltre, mediante le funzioni *Corrections history* e *Editions history* si accede, rispettivamente, allo storico delle modifiche apportate nell'etichettatura e allo storico delle correzioni effettuate sui *tokens* servendosi delle operazioni offerte dalla colonna 9.

[17] Due i possibili scenari. Uno formale: il lemma, il valore *POS* o il valore *Morph* inserito non è digitato nella forma prevista dalla lista di controllo oppure non è stato selezionato correttamente il completamento automatico della parola o dell'etichetta suggerito da *Pyrrha*. L'altro sostanziale: nella colonna *Lemma*, *POS* o *Morph* si è scritto un valore non contemplato nella lista di controllo, ignorando il suggerimento automatico, oppure la procedura di validazione per l'aggiunta di un nuovo lemma o etichetta nella lista di controllo non è andata a buon fine e deve quindi essere ripetuta.

[18] Per perfezionare la ricerca, alle stringhe si possono aggiungere anche i caratteri jolly contemplati da *Pyrrha*, che qui riportiamo: «* can be used to match partial words, eg. ADV*; ! can be used to negate a match, eg. !PRE; | can be used to perform an OR operation, eg. s*t | s*s. To search forms which do not contain 'e': ! *e*».

1.2. *Il modello di analisi lessicale per l'antico francese*

Come accennato nel paragrafo precedente, nella fase preparatoria di scelta dei criteri di analisi, *Pyrrha* richiede di specificare quale modello e quale lista di controllo si vogliono utilizzare per il trattamento del *corpus* importato. Ciò serve ad impostare il *tagger Pie* sul paradigma di riferimento che si ritiene corrispondente o, quanto meno, vicino alla lingua del testo sottoposto all'analisi lessicale. Nel nostro caso, il modello più affine all'*Entrée d'Espagne* era senza dubbio *Old French*, frutto dell'addestramento di *Pie* su *corpora* in antico francese costituiti in gran parte presso l'*ENC*[19]. Su tale modello opera la lista di controllo *Ancien Français*, che a sua volta si compone di tre liste specifiche:

1. la lista dei lemmi;
2. la lista delle *Part of Speech* (*POS*);
3. la lista delle annotazioni morfosintattiche (*Morph*).

I lemmi contenuti nella prima lista sono stati ricavati dall'edizione elettronica del dizionario TL, mentre le etichette delle altre due liste provengono dal sistema *Cattex09*, sviluppato presso la *Base de Français Médiéval* a Lione[20].

In *Cattex09* l'annotazione *POS* e *MSD* si distribuisce su due sezioni: *POS*, che indica la categoria lessicale del vocabolo, e *Morph*, che ne descrive la flessione. Le etichette *POS*, corrispondenti alle tradizionali nove parti del discorso, possono essere composte da uno o due campi. Il primo campo rappresenta la categoria e il suo valore è espresso da tre lettere maiuscole:

VER (verbo), **NOM** (nome), **ADJ** (aggettivo), **PRO** (pronome), **DET** (determinante), **ADV** (avverbio), **PRE** (preposizione), **CON** (congiunzione), **INJ** (interiezione)[21].

Il secondo campo, se presente, indica la tipologia e il suo valore è espresso da tre lettere minuscole:

VER: VERcjg (verbo coniugato), VERinf (verbo all'infinito), VERppe (participio passato), VERppa (participio presente).

NOM: NOMcom (nome comune), NOMpro (nome proprio).

[19] Cfr. Pinche 2019: 51.

[20] Cfr. Guillot – Lavrentiev – Prévost 2013a; Guillot – Lavrentiev – Prévost 2013b.

[21] *Cattex09* prevede inoltre cinque categorie supplementari, funzionali all'annotazione: PON (punteggiatura), ETR (parola straniera), ABR (abbreviazione), RED (parola ridondante), OUT (parola che si vuole escludere dall'analisi).

ADJ: ADJqua (aggettivo qualificativo), ADJind (aggettivo indefinito), ADJcar (aggettivo cardinale), ADJord (aggettivo ordinale), ADJpos (aggettivo possessivo).

PRO: PROper (pronome personale), PROimp (pronome impersonale), PROadv (pronome avverbiale), PROpos (pronome possessivo), PROdem (pronome dimostrativo), PROind (pronome indefinito), PROcar (pronome cardinale), PROord (pronome ordinale), PROrel (pronome relativo), PROint (pronome interrogativo), PROcom (pronome composto).

DET: DETdef (determinante definito), DETndf (determinante non definito), DETdem (determinante dimostrativo), DETpos (determinante possessivo), DETind (determinante indefinito), DETcar (determinante cardinale), DETrel (determinante relativo), DETint (determinante interrogativo), DETcom (determinante composto).

ADV: ADVgen (avverbio generale), ADVneg (avverbio negativo), ADVint (avverbio interrogativo), ADVsub (avverbio subordinante).

CON: CONcoo (congiunzione coordinante), CONsub (congiunzione subordinante).

A queste *POS* si deve inoltre aggiungere un gruppo di etichette definite 'complesse', che rendono conto dei fenomeni di enclisi e proclisi nei quali due unità linguistiche si fondono in un'unità grafica. Tali etichette complesse sono formate a loro volta da due etichette semplici separate da un punto:

PROper.PROper (pronome personale + pronome personale)
PROrel.PROper (pronome relativo + pronome personale)
PROrel.PROadv (pronome relativo + pronome avverbiale)
PROrel.ADVneg (pronome relativo + avverbio negativo)
ADVgen.PROper (avverbio generale + pronome personale)
ADVneg.PROper (avverbio negativo + pronome personale)
PRE.DETdef (preposizione + determinante definito)
PRE.DETcom (preposizione + determinante composto)
PRE.DETrel (preposizione + determinante relativo)
PRE.PROper (preposizione + pronome personale)
PRE.PROrel (preposizione + pronome relativo)
PRE.PROint (preposizione + pronome interrogativo)
CONsub.PROper (congiunzione subordinante + pronome personale)

Le etichette *Morph* esprimono le caratteristiche morfosintattiche del vocabolo. Per la flessione verbale:

MODE (modo): *ind* (indicativo), *imp* (imperativo), *con* (condizionale), *sub* (congiuntivo).

TEMPS (tempo): *pst* (presente), *ipf* (imperfetto), *fut* (futuro), *psp* (passato remoto).

PERS (persona): *0* (impersonale), *1* (prima persona), *2* (seconda persona), *3* (terza persona).

Per la flessione nominale, aggettivale e pronominale:

NOMB (numero): *s* (singolare), *p* (plurale).

GENRE (genere): *m* (maschile), *f* (femminile), *n* (neutro).

CAS (caso): *n* (*cas sujet*), *r* (*cas régime*), *i* (*cas régime indirect*).

DEGRE[22] (grado): *p* (positivo), *c* (comparativo), *s* (superlativo).

PERS[23] (persona): *0* (impersonale), *1* (prima persona), *2* (seconda persona), *3* (terza persona).

1.3. *La preparazione del testo*

Sono state sottoposte ad una prima prova di analisi lessicale le porzioni testuali dell'edizione Thomas dell'*Entrée d'Espagne* riviste da Marco Infurna[24]: 1-7 (fino al v. 169); 10 (fino al v. 279); 386-393; 423-426; 429-435 (fino al v. 10042); 437 (dal v. 10080 al v. 10089); 438-439 (fino al v. 10136); 444-452; 458 (dal v. 10552 al v. 10563); 459 (fino al v. 10600); 477-492 (fino al v. 11362); 497-564 (fino al v. 13253); 566-595; 607-609; 614-615; 619-621 (fino al v. 14492); 628 (dal v. 14633 al v. 14647); 629-681.

Salvate in un unico file *txt* come testo puro, sono state manipolate con uno specifico sistema di marcatura al fine di impedire la perdita di informazioni testuali (suddivisione in lasse e confini dei singoli versi) nella fase finale di esportazione dei dati[25]. Si è inoltre stabilito di conservare la punteggiatura inserita

[22] Solo per gli aggettivi.

[23] Solo per i possessivi e per i pronomi personali e impersonali.

[24] Cfr. *L'Entrée d'Espagne* (ed. Thomas 1913; ed. Infurna 2011). È stata rispettata la segmentazione in unità grafiche del testo critico. Generalmente un'unità grafica coincide con un'unità linguistica. Per l'esame dettagliato delle possibili eccezioni si rimanda a Guillot – Lavrentiev – Prévost 2013a: 4-6.

[25] L'elaborazione del sistema di marcatura si deve a Luigi Tessarolo, responsabile tecnico del *RIALFrI*, che ringraziamo.

dall'editore critico, in modo tale da poter ricostituire facilmente la *mise en page* dell'edizione per una successiva pubblicazione in rete del testo annotato e lemmatizzato. *Pie* è programmato per distinguere tra punteggiatura forte, che delimita le frasi (punto, punto interrogativo, punto esclamativo, puntini di sospensione), e debole, che scandisce internamente la frase (virgola, punto e virgola, due punti, trattino)[26].

A posteriori, una volta caricato il testo in *Pyrrha*, abbiamo riscontrato un numero esiguo di refusi (72 su un totale di 58990 *tokens*) dovuti alla fase di riconoscimento ottico dei caratteri, che siamo riusciti a correggere utilizzando le opzioni di modifica previste nell'interfaccia di post-correzione alla colonna 9 (*Edit the form*, *Delete the row*, *Add a token after this one*)[27].

2. *La sperimentazione sull'Entrée d'Espagne*

È stato ampiamente dimostrato che l'automatizzazione dell'analisi lessicale implica un notevole risparmio di tempo e di lavoro, riducendo l'intervento manuale ad una fase secondaria di revisione. Per l'antico francese il tasso di correttezza di *Pie* si aggira attorno al 93%, nonostante solo un 30% del *corpus* di apprendimento sia annotato anche morfosintatticamente[28]. I risultati ottenuti su testi riconducibili ad un modello di analisi già presente in *Pyrrha* sono quindi rilevanti. Com'era prevedibile, invece, la riuscita su testi linguisticamente eterodossi rispetto ai paradigmi disponibili è ridotta: le percentuali di rettifica manuale sull'annotazione e sulla lemmatizzazione automatiche risultano elevate e la fase di post-correzione si rivela alquanto dispendiosa, dato che comporta un'attenta e puntuale rilettura *token* per *token*. Le statistiche rivelano che nel nostro caso specifico siamo dovuti intervenire sull'84% dei lemmi, sul 69% delle morfologie e sul 90% delle *POS*. L'alto valore del tasso di correzioni, tuttavia, non è da

[26] Nel sistema *Cattex09* i due tipi di punteggiatura sono identificati rispettivamente dalle etichette PONfrt e PONfbl.

[27] Cfr. nota 16. I refusi osservati sono tutti riconducibili a due tipologie di errore: errata lettura della sequenza grafica o errata segmentazione della sequenza grafica. Il primo tipo di errore si può manifestare come confusione tra singole lettere (ad es. *voit* per *voil*), confusione tra più lettere (ad es. *en* per *ert*), omissione di una lettera nel corpo della parola (ad es. *repont* per *respont*) o nella terminazione (ad es. *Espaign* per *Espaigne*), omissione di una lettera che corrisponde ad un'unità grafica (ad es. *quant fil* per *quant a fil*), aggiunta di una lettera finale (ad es. *or* per *o*). Il secondo tipo di errore si riscontra quando un'unità grafica viene scomposta in due componenti (ad es. *a prés* per *aprés*) oppure quando due unità grafiche vengono unite in un unico elemento (ad es. *dite* per *dit e*). Si potrebbe aggiungere anche una terza tipologia di errore, che combina i due sopra illustrati e che si verifica in presenza dell'apostrofo. Si tratta di un'errata lettura che implica un'errata segmentazione (ad es. *qil* per *q'il*).

[28] Cfr. Pinche 2019: 51.

imputare totalmente alla *performance* del *tagger* e all'inadeguatezza, per determinati aspetti, del modello d'analisi utilizzato (concepito, lo ricordiamo, per l'antico francese e non per il franco-italiano). Esso, infatti, dipende in parte anche dalla configurazione stessa degli strumenti impiegati: il dizionario TL e il sistema d'annotazione *Cattex09*.

Nei paragrafi seguenti passeremo in rassegna le problematiche riscontrate durante la fase di post-correzione ed illustreremo brevemente le soluzioni messe a punto per risolvere, o almeno aggirare in tempi contenuti, quelle difficoltà. A tal proposito, è utile chiarire che per praticità abbiamo ricondotto l'eterogeneità di forme linguistiche presenti nell'*Entrée d'Espagne* a quattro categorie:

1. forme francesi;
2. forme ibride nella fonetica e/o nella morfologia che condividono etimologia e significato con il corrispettivo lemma francese;
3. forme italiane o italianizzate che condividono etimologia e significato con il corrispettivo lemma francese;
4. forme italiane che non trovano riscontro etimologico in un lemma francese.

2.1. *La post-correzione della lemmatizzazione e il dizionario Tobler-Lommatzsch*

La revisione della lemmatizzazione dell'*Entrée d'Espagne* si è basata su un unico criterio operativo: ricondurre tutte le forme possibili al corrispettivo esponente lessicale francese registrato nel TL. Per ovvi motivi, tale principio non si è potuto rispettare per le forme appartenenti alla quarta categoria sopra elencata. La lemmatizzazione di un testo franco-italiano richiederebbe la compresenza, nel modello di analisi, di due dizionari di riferimento, uno per ogni 'polo linguistico' coinvolto. Per l'italiano si sta infatti valutando la possibilità di predisporre un secondo elenco alfabetico di lemmi, ricavato dal *Tesoro della Lingua Italiana delle Origini*.

Per ovviare nell'immediato a tale mancanza, per l'*Entrée d'Espagne* si è deciso di adottare una soluzione di compromesso, consistente nell'aggiunta manuale dell'esiguo numero di lemmi italiani alla lista di entrate francesi già presente in *Pyrrha*. L'integrazione si è dimostrata preziosa, ma ha posto un'ulteriore questione: come mantenere traccia del fatto che le nuove entrate fossero degli italianismi? I lemmi inseriti, infatti, andavano semplicemente ad aggiungersi al lungo elenco alfabetico delle entrate già presenti nella lista di controllo e l'informazione 'italianismo' andava persa anche nell'interfaccia di post-correzione, poiché non segnalabile né nella colonna 2 (*Form*) né nella colonna 3 (*Lemma*). L'annotazione morfosintattica costituiva il solo spazio di intervento rimasto a disposizione. Si è stabilito dunque di esprimere l'informazione 'italianismo' mediante un nuovo attributo, SPEC=it, che è stato anteposto alle etichette di una seconda lista *Morph*. Qualche esempio:

MODE=ind|TEMPS=pst|PERS.=1|NOMB.=s > SPEC=it|MODE=ind|TEMPS=pst|PERS.=1|NOMB.=s

PERS.=1|NOMB.=s|GENRE=m|CAS=n > SPEC=it|PERS.=1|NOMB.=s|GENRE=m|CAS=n

Il riesame minuzioso della lemmatizzazione ha mostrato inoltre i limiti intrinseci allo stesso TL, alcuni dei quali già osservati da Jean-Baptiste Camps nel corso dell'analisi lessicale del corpus *Geste*[29].

Innanzitutto, nel TL risulta difficoltoso raccapezzarsi tra le numerose entrate omografe, con le quali occorre familiarizzarsi prima di intraprendere la post-correzione in *Pyrrha*. I casi di omografia da disambiguare possono essere di due tipi[30]:

- lemmi che derivano da basi latine diverse, come ad esempio

Lemma	**Base latina**	**Categoria lessicale**
que1	*quam*	ADVgen/CONsub (*que* comparativo)
que2	*qui, quem, quam, quod*	PROrel
que3	*quid*	PROint
que4	*quia*	CONsub

- lemmi che derivano dalla stessa base latina, ma che appartengono a categorie lessicali differenti, come ad esempio

Lemma	**Base latina**	**Categoria lessicale**
bien1	*bene*	ADVgen
bien2	*bene*	NOMcom

In secondo luogo, il TL digitalizzato presenta alcune discrepanze nei rinvii interni; talvolta, infatti, nelle pagine di consultazione l'ordine delle entrate è

[29] Cfr. Camps 2016, sezione *Wiki*, *Erreurs et cas fréquents* (https://github.com/Jean-Baptiste-Camps/Geste/wiki/Erreurs-et-cas-fr%C3%A9quents).

[30] Nel trattare le entrate omografe il TL provvede a numerare i lemmi per differenziarli. Lo stesso espediente è stato adottato nella nostra sperimentazione al fine di integrare eventuali nuovi lemmi omografi rispetto a quelli già registrati nella lista di controllo di *Pyrrha*.

invertito rispetto a quello dell'elenco alfabetico dei lemmi. Ad esempio, all'interno della sezione *moi-mol* si trovano rispettivamente *moillier s.f.* e *moillier1 v.* ma, se si clicca sui vocaboli per visualizzarne la definizione, si incontra prima il verbo *moillier* < *MOLLIARE e poi il sostantivo *moillier* < MULIER.

Ricordiamo, infine, la scarsità di avverbi in *-ment* registrati come veri e propri lemmi nel TL (gli unici che abbiamo trovato per la nostra lemmatizzazione sono *ensement* ed *escordement1*), che preferisce rimandare direttamente all'aggettivo dal quale la forma avverbiale deriva. Svolgendo il lavoro di post-correzione ci siamo accorti che avere a disposizione due esponenti lessicali differenti e indipendenti, uno per l'aggettivo e uno per l'avverbio, facilitava la lemmatizzazione. Abbiamo così provveduto a creare delle entrate *ad hoc* per gli avverbi in *-ment* e le abbiamo integrate alla lista di lemmi in *Pyrrha*, prestando particolare attenzione al fatto che in alcuni casi esistevano già nel TL dei vocaboli omografi in *-ment* appartenenti ad un'altra categoria lessicale (ad esempio *forment s.m.* < FRUMENTUM)[31].

Un ultimo aspetto collegato collateralmente al dizionario di riferimento è rappresentato dal nutrito gruppo di toponimi e nomi propri presenti nel testo dell'*Entrée d'Espagne*, ma privi di riscontro nel TL. Anche in tal caso i nuovi lemmi sono stati aggiunti manualmente, riconducendoli, ove possibile, alla forma registrata nel *DMF*.

2.2. *La post-correzione dell'annotazione e il sistema Cattex09*

La revisione dell'annotazione morfosintattica ha richiesto talvolta la manipolazione delle etichette *Cattex09*, al fine di dar conto delle caratteristiche di ogni forma con la maggior precisione possibile. Abbiamo, ad esempio, precisato che la nostra annotazione rappresenta la funzione del vocabolo (*cas sujet*, *cas régime*, *cas régime indirect*) a prescindere dal rispetto formale delle marche di declinazione bicasuale, criterio che si è reso obbligatorio dato il profilo linguistico dell'*Entrée d'Espagne*. Della stessa natura è il trattamento che si è deciso di riservare ai participi passati che formano una voce verbale composta con ausiliare 'avere', per i quali l'accordo appare molto instabile. Se il participio passato è parola-rima, nel campo *Morph* compare l'etichetta con valore nullo NOMB.=x|GENRE=x|CAS=x, poiché la rima condiziona sensibilmente la desinenza della parola; ma se il participio passato a fine verso è declinato al femminile, allora in *Morph* si segnala l'accordo. Se l'oggetto della frase precede il verbo o si trova tra ausiliare e participio, si segnala l'accordo, anche se il participio è in rima; in tutti gli altri casi, si ricorre all'etichetta NOMB.=x|GENRE=x|CAS=x per indicare la mancanza di accordo esplicito.

[31] Abbiamo inoltre conservato le forme avverbiali in *-ment* già presenti in *Pyrrha*, derivanti da analisi lessicali precedenti, come quella svolta sul *corpus Geste*.

In un solo caso abbiamo stabilito di omettere un elemento in un'etichetta, allo scopo di non appesantire troppo l'annotazione: si tratta dell'attributo DEGRE=p destinato agli aggettivi. Ne risulta che il grado degli aggettivi viene indicato solo se si tratta di comparativi o di superlativi, mentre in tutti gli altri casi, si lascia sottointesa l'informazione 'grado positivo'.

Interventi che vanno nella direzione dell'esplicitazione o dell'estensione dei parametri *Cattex09* sono il fatto di considerare *cas régime* il pronome riflessivo che assume la funzione di oggetto nei verbi pronominali, oltre che l'attribuzione del tratto 'neutro' (per i pronomi) all'impersonale *il*, al dimostrativo *ce*/*c'* con funzione presentativa, all'indefinito *on*, ai relativi (talvolta interrogativi) *coi2*, *dont*, *où* e *que2*, e agli indefiniti *aucant*, *autre*, *autretel*, *cant2*, *itant*, *nïent*, *petit*, *plus*, *poi*, *rien*, *tant*, *tel*, *tot*, *trop*, quando non si riferiscono ad un antecedente maschile o femminile concreto ed esplicitato, bensì ad un concetto, un pensiero, un'intera frase. Inoltre, l'etichetta *POS* PROimp e gli attributi PERS.=0, GENRE=n, riservati originariamente al pronome impersonale *il*, sono stati estesi anche al pronome riflessivo *se* (lemma *soi1*), quando si configura come un italianismo che svolge la medesima funzione del *si* impersonale in italiano.

Ad un adattamento dell'uso delle etichette sono invece da ascrivere le innovazioni apportate nei campi *Morph* (e talora *POS*), quando si verificano cambiamenti di categoria lessicale o quando occorre distinguere più funzioni di un medesimo vocabolo. Nel primo caso si tratta del passaggio di una forma da una categoria lessicale non flessa ad una flessa (il caso più frequente è quello degli infiniti sostantivati: VERinf > NOMcom): nel campo *Morph* si sostituisce all'etichetta di valore nullo NOMB.=x|GENRE=x|CAS=x l'annotazione morfologica completa. Nel secondo caso si tratta della discriminazione tra le due funzioni che il participio presente può assumere (participio o gerundio): poiché il *POS* rimane invariato (VERppa), abbiamo differenziato l'annotazione morfologica, rendendo conto della flessione per la funzione participio e segnalando il valore nullo per la funzione gerundio. All'intersezione tra quest'ultima casistica e il principio di estensione di un parametro, si colloca la scelta di etichettare *POS* = VERinf e *Morph* = MODE=imp|PERS.=2|NOMB.=s quegli imperativi negativi alla seconda persona singolare costruiti con negazione + infinito. In tal modo, nel campo *POS* si restituisce la categoria del vocabolo (infinito) e nel campo *Morph* la funzione (imperativo) e si evita di perdere parte dell'informazione.

Con l'attributo NOMB nell'annotazione dei possessivi ci siamo trovati di fronte ad un caso di ambiguità connaturata all'etichetta: non è chiaro, infatti, se esso richiami il numero della persona o il numero del sostantivo al quale il possessivo si riferisce. Poiché *Cattex09* non prevede di assegnare una doppia etichetta e poiché non si voleva appesantire troppo l'annotazione con un'aggiunta nel campo *Morph*, in questo primo tentativo di annotazione abbiamo preferito riferirci al numero della persona del possessivo[32].

[32] Al contrario, nel corpus *Geste* ci si riferisce al numero del sostantivo, se il lemma permette di

Un'altra scelta di disambiguazione si è presentata con le etichette *POS* composte, alle quali non è possibile abbinare una doppia annotazione morfologica; di conseguenza, si rende necessario decidere a quale dei due elementi riferire l'annotazione del campo *Morph*. Valutando di volta in volta la composizione delle *POS*, siamo giunti alle seguenti conclusioni[33]:

Composizione	**POS**	**Annotazione morfologica**
elemento non flesso + elemento flesso	PRE.DETdef: *al*, *au*, *as*, *o* (*a3* + *le*); *del*, *della*, *dels*, *des*, *deu*, *do*, *dou* (*de* + *le*); *el*, *es*, *eu*, *ou*, *oul* (*en1* + *le*). ADVgen.PROper: *ses*, *sil* (*si* + *il*). ADVneg.PROper: *nel*, *nes*, *nol* (*ne1* + *il*). CONcoo.DETdef: *el* (*et* + *le*). CONcoo.PROper: *el* (*et* + *il*). CONsub.PROper: *q(u)el* (*que4* + *il*). CONsub.DETdef: *chel* (*que4* + *le*).	descrive il secondo elemento

risalire all'informazione della persona: cfr. Camps 2016, sezione *Wiki*, *Adaptations par rapport à Cattex2009* (https://github.com/Jean-Baptiste-Camps/Geste/wiki/%5BR%C3%89F%5D-flexion:-adaptations-par-rapport-%C3%A0-Cattex2009). La questione si potrebbe risolvere assegnando una scala di valori da 0 a 6, invece che da 0 a 3, alla categoria 'persona'.

[33] Nel corpus *Geste* l'annotazione morfologica in questi casi si riferisce sempre al secondo componente: cfr. Camps 2016, sezione *Wiki*, *Adaptations par rapport à Cattex2009* (https://github.com/Jean-Baptiste-Camps/Geste/wiki/%5BRÉF%5D-flexion:-adaptations-par-rapport-à-Cattex2009).

Composizione	**POS**	**Annotazione morfologica**
elemento flesso + elemento flesso	PROper.PROper: *jel* (*je* + *il*). PROrel.PROper: *qel* (*que2* + *il*); *qil* (*qui* + *il*).	descrive il primo elemento

Composizione	**POS**	**Annotazione morfologica**
elemento non flesso + elemento non flesso	ADVgen.PROadv: *jan* (*ja* + *en2*); *sin* (*si* + *en2*). ADVneg.PROadv: nessuna occorrenza riscontrata nel testo.	MORPH= empty

Creazioni *ex novo* sono la doppia valenza 'm/f' dell'attributo GENRE, pensata per i sostantivi ambigenere (nei casi in cui il contesto non consenta di discriminare tra maschile e femminile) e per i metaplasmi di genere; l'aggiunta dell'attributo SPEC=lat nell'annotazione morfosintattica di un gruppo di 25 vocaboli in latino[34]; l'introduzione di nuove etichette *POS* complesse per descrivere alcune forme non contemplate in *Cattex09*[35]:

chel < *que4* + *le*	CONsub.DETdef
el < *et* + *le*	CONcoo.DETdef
el < *et* + *il*	CONcoo.PROper
qin < *que4* + *en2*	CONsub.ADVgen

[34] A differenza dell'etichetta generica ETR, prevista nel sistema *Cattex09* (Guillot – Lavrentiev – Prévost 2013a: 4), tale soluzione ha il vantaggio di specificare direttamente la lingua che interviene sul testo.

[35] Camps segnalava già la mancanza dell'etichetta CONsub.DETdef per il corpus *Geste*: cfr. Camps 2016, sezione *Wiki*, *Adaptations par rapport à Cattex2009* (https://github.com/Jean-Baptiste-Camps/Geste/wiki/%5BRÉF%5D-flexion:-adaptations-par-rapport-à-Cattex2009).

2.3. *L'attributo SPEC=it*

L'attributo SPEC=it, che abbiamo introdotto per isolare facilmente gli italianismi dal resto dei lemmi francesi, ha risposto efficacemente al bisogno pratico di trovare una modalità ergonomica mediante la quale conservare nell'interfaccia di post-correzione un'informazione linguistica importante. Tuttavia, sappiamo che in un testo franco-italiano l'italianizzazione si può manifestare a diversi livelli e con gradi differenti. Oltre che con forme prettamente italiane che presentano un'etimologia non riconducibile ad un lemma francese (categoria 4 della nostra classificazione), nell'*Entrée d'Espagne* abbiamo a che fare con forme derivanti dalla medesima base etimologica del corrispettivo lemma francese, ma caratterizzate dalla presenza di tratti fonetici e/o morfologici italiani (categorie 2 e 3 della nostra classificazione).

Per consentire di rintracciare agevolmente e rapidamente tutte le forme contraddistinte in un qualche modo dall'italianizzazione, e non solo gli italianismi schietti (presenti nell'*Entrée d'Espagne* in misura minore rispetto alle forme miste), si è stabilito di estendere l'uso dell'attributo SPEC=it, dotandolo di una certa polivalenza[36]. Si tratta di un espediente provvisorio, che rappresenta la scelta meno onerosa che si potesse adottare a questo stadio della sperimentazione (la sua stringatezza, infatti, ben si adatta all'ambiente di lavoro di *Pyrrha*) e che svolge in modo funzionale il compito di raggruppare un insieme preciso di vocaboli.

Ecco l'elenco dei tratti linguistici per i quali è stato assegnato l'attributo SPEC=it nella nostra analisi lessicale, accompagnati da alcuni esempi a titolo indicativo:

1. Particolarità grafico-fonetiche:

- grafia *ç* per l'affricata alveolare o palato-alveolare in posizione iniziale (*çelui*, *çevalce*, *çiere*), centrale (*pieçe*, *remembrançe*, *trabuçer*) o finale (*braç*, *douç*, *pieç*);
- digramma *ch* con valore velare in finale di parola (*avech*, *flanch*, *franch*, *iluech*, *Moroch*) e davanti alle vocali *o* ed *u* in posizione iniziale o centrale (*choumenzier*, *inchuntre*, *schuz*);
- presenza di *h* iniziale in 40 voci del verbo *avoir* su un totale di 933;
- grafia *qe* con valore velare in posizione iniziale o centrale (*donqe*, *qatorçes*, *qevaus*);
- grafia *x* per la sibilante in posizione iniziale (*xeromes*, *xon*), intervocalica (*garixon*, *servixe*, *voixine*) o finale (*citex*, *dux*, *prix*);

[36] Riteniamo che la creazione di due diversi attributi, SPEC=fr-it e SPEC=it, oltre che sovraccaricare ulteriormente il sistema, avrebbe rappresentato un azzardo, poiché la classificazione rigida del tratto 'italiano' o 'franco-italiano' sarebbe risultata altamente discrezionale, data la natura del testo e la sua collocazione cronologica.

- grafia *z* per l'affricata alveolare ad inizio (*zanberlan*, *zantent*, *zastel*) o fine parola (*guainz*, *lignaz*, *soiez*).

2. Fenomeni fonologici:

- aferesi vocalica di *e-* (*sperance*, *spoblïer*, *stroitement*) e di *o-* (*scure*).

3. Fenomeni morfologici:

- participi passati uscenti in *-ei* (*armei*, *ostei*, *visitei*);
- voci verbali uscenti in *-o* alla prima persona singolare del presente indicativo o del *passé simple* (*do*, *fo*, *so*);
- forme italiane o italianizzanti che presentano lo stesso etimo delle forme francesi: articolo *lo*; preposizioni *da*, *in*; pronomi relativi *che*, *chi*; avverbio *tutor*.

4. Costruzioni italianizzanti:

- complemento di termine espresso nella forma analitica *a* + pronome[37] (*E plus doit estre sajes li justixer* | *Qe cil qe vient a lui droit demander*[38]; *Rolant oï del duc la lamentançe*; | *Respond a lui con molt gran piatançe*[39]).

5. Slittamenti di significato e di funzione:

- pronome riflessivo *se* (lemma *soi1*) equivalente al *si* impersonale italiano (*Ceste feit asavoir cum hom se doit pener* | *D'esamplir la loy Deu et as povres aider*[40]; *Car en Espaigne ne se jüe ne rit*[41]).

6. Lemmi italiani:

- preposizione *con*
(*Con sa gient est Rollant departiz et anblez*[42]; *Le piez lui veit baser con lee ciere et pie*[43]);

[37] In *Morph* il pronome è segnalato come italianismo, in modo tale che non si perda l'informazione uniformandola alle forme sintetiche francesi.
[38] *Entrée d'Espagne* (ed. Infurna 2011): 126, vv. 11236-11237.
[39] Ivi: 370, vv. 15554-15555.
[40] Ivi: 44, vv. 22-23.
[41] Ivi: 180, v. 12123.
[42] Ivi: 68, v. 9185.
[43] Ivi: 96, v. 10333.

- particella avverbiale *li*
  (*Li bon sant il meïme, sil prist a menacer | Que, se il no li aloit, il avroit engombrer*[44]; *E quant l'on li li giete o pains o autre sor*[45]);

- avverbio *pur*, talvolta utilizzato in locuzioni congiuntive
  (*Cevauce pur avaint, qe bien te segirai*[46]; *Pur ch'il retorne sa vie a defensaille, | De perdre honor ne prise une meaille*[47]).

3. *Verso un modello di analisi dedicato al franco-italiano*

Dalla sperimentazione condotta sull'*Entrée d'Espagne* ci pare che emerga una considerazione sostanziale: il modello di analisi lessicale concepito per i testi in antico francese può costituire una solida base di partenza da cui prendere spunto per lo sviluppo di un paradigma adatto al franco-italiano. Si tratta del primo passo verso la definizione di un nuovo modello, al quale si potranno in futuro sottoporre altre opere della produzione letteraria franco-italiana per ottenere un'annotazione morfosintattica e una lemmatizzazione sempre più precise e complete. Come dimostra la trattazione delle problematiche affrontate durante questo primo test, gli interventi da effettuare si concentrano sui due principali strumenti che fanno funzionare il modello, ovvero il dizionario di riferimento e il sistema di etichette.

Abbiamo già espresso il bisogno di disporre di un dizionario di riferimento anche per l'italiano, in modo tale da poter distinguere nella lemmatizzazione le forme prettamente italiane da quelle italianizzate ma comunque riconducibili ad un esponente lessicale francese. Ci sarebbe da riflettere anche sulla scelta del dizionario di riferimento per il francese, considerate le problematiche che il TL pone. Una volta trovato il modo di integrare nella lista di controllo un ulteriore elenco di lemmi, ricavato da un secondo dizionario di riferimento, si potrebbe inoltre valutare l'ipotesi di aggiungere progressivamente più vocabolari per la stessa lingua (ad esempio, per il francese, il *DEAF*, la versione digitalizzata del *Gdf*, il *DMF*, ecc.). Ciò consentirebbe di arricchire la lemmatizzazione con rinvii mirati e di intuire a colpo d'occhio se la forma lemmatizzata è francese, ibrida o italiana, in base ai dizionari citati. Concretamente, nell'interfaccia di lavoro *Pyrrha*, nel campo riservato al lemma (colonna 3), si potrebbe registrare ogni singola entrata affiancandola alla sigla indicante il dizionario utilizzato[48].

[44] Ivi: 46, vv. 33-34.
[45] Ivi: 88, v. 10120.
[46] Ivi: 68, v. 9165.
[47] Ivi: 246, vv. 13156-13157.
[48] Jean-Baptiste Camps, che ringraziamo per averci fornito indicazioni preziose durante la stesura di questo contributo, ci segnala che tale soluzione, per quanto interessante, porrebbe dei problemi in termini di 'predizione del lemma', poiché le regole di costruzione non risulterebbero omogenee.

Riteniamo che il perfezionamento del sistema di etichette, al di là delle lievi variazioni che ogni testo franco-italiano richiede per esplicitare in modo adeguato tutte le sue peculiarità, passi attraverso la distinzione dei vari livelli linguistici coinvolti (grafico-fonetico, morfosintattico, semantico) e del grado di mescidanza tra i due codici, il francese e l'italiano. Provvisoriamente, nella nostra sperimentazione abbiamo deciso di impiegare l'attributo SPEC=it che, per la sua concisione e per la sua praticità, ben si adattava all'ambiente di lavoro di *Pyrrha*[49]. L'obiettivo a lungo termine, però, è realizzare uno strumento che riesca ad indagare le dinamiche variazionali e distributive del franco-italiano all'interno di un testo o di un *corpus*. In tale prospettiva si colloca il prototipo che proponiamo e che vorremmo sviluppare. Esso prende in considerazione due serie di fattori, comprendenti tre parametri ciascuna: nella prima serie i parametri riferiti all'ambito linguistico, nella seconda quelli relativi al posizionamento della forma esaminata tra i due poli del francese e dell'italiano.

	Semantic	**Morphology**	**Spelling and phonetics**
French form	FS	FM	FP
Hybrid form (Franco-Italian form)	HS	HM	HP
Italian form	IS	IM	IP

3.1. *Gli scopi dell'analisi lessicale*

La riflessione circa un nuovo modello di analisi lessicale per il franco-italiano, a nostro avviso, non può prescindere da tre ulteriori considerazioni, più o meno pragmatiche: le finalità reali (e il pubblico) che si vogliono raggiungere; il grado di finezza da attribuire all'annotazione morfosintattica e alla lemmatizzazione; le risorse e il tempo a disposizione. Partiamo dall'ultimo punto. Inizialmente si era supposto di limitare la sperimentazione sull'*Entrée d'Espagne* al *POS tagging* e alla lemmatizzazione; dopo aver testato *Pyrrha* ed averne apprezzato la funzionalità, però, abbiamo ampliato l'analisi anche al *MSD tagging*. Ciò ha sicuramente dilatato le tempistiche e lo sforzo di analisi, ma ci ha consentito di migliorare il secondo punto, ovvero lo scendere maggiormente nei dettagli, cercando di inserire quante più informazioni possibili, in maniera particolareggiata, nelle etichette.

La soluzione ideale, ma di certo non realizzabile nel breve periodo, suggerisce Camps, consisterebbe nell'affidare il mantenimento e l'aggiornamento di un unico dizionario di riferimento esaustivo ad un'*équipe* di lessicografi.

[49] Diverso il caso dell'attributo SPEC=lat, poiché l'eventuale presenza di inserti latini nei testi franco-italiani è minima e non soggetta a ibridazione linguistica.

Ricordiamo che ci siamo sempre dovuti confrontare con l'impossibilità, da un lato, di applicare una doppia etichettatura e con il rischio, dall'altro, di sovraccaricare troppo l'annotazione, compromettendone la leggibilità. L'attenzione verso il dettaglio e il proposito di restituire tutte le informazioni contenute in un vocabolo scaturiscono anche dalla valutazione degli obiettivi che sottendono all'analisi lessicale di un testo come quello dell'*Entrée d'Espagne*: rendere progressivamente interrogabile l'intero *corpus* della letteratura franco-italiana; condividere con la comunità scientifica i testi annotati e lemmatizzati; aumentare la documentazione riguardante l'evoluzione diacronica e diatopica della *scripta* franco-italiana; raccogliere dati linguistici per un potenziale uso lessicografico volto alla creazione di concordanze, glossari, dizionari.

4. *L'esportazione del testo*

Una volta terminata la fase di post-correzione, utilizzando la funzione *Export tokens* di *Pyrrha*, è possibile esportare il testo analizzato in formato *csv* o *TEI*. Nel primo caso si avrà un *file* del tipo *comma-separated values* (*csv*), cioè un *file* in cui i dati sono organizzati all'interno di una struttura a tabella, dove le colonne (*Form*, *Lemma*, *POS*, *Morph*) riproducono l'interfaccia di lavoro di *Pyrrha*. Nel secondo caso si otterrà un file *XML* nel quale i risultati visualizzati rispetteranno il seguente schema di codifica:

```
<w xml:id="tn" n="n" lemma="lemma"
type="POS|annotazione_morfologica">forma</w>
```

Ogni forma è codificata come un elemento <*w*> (*word*), al cui interno vengono restituite le informazioni annotate in *Pyrrha* mediante un insieme di attributi:

- @xml:id: identificante *Pyrrha* del *token.*
- @n: numero d'ordine del *token* nel *corpus.*
- @lemma: lemma del dizionario di riferimento.
- @type: etichetta *POS* e *Morph* concatenate[50].

Il formato *XML-TEI*[51] permette di intervenire ulteriormente sul testo analizzato, rimaneggiando i dati ottenuti. Oltre a poter restituire l'impostazione del documento e la gerarchia delle parti che lo compongono, codificando ad esempio la suddivisione in versi e in lasse, esso consente di effettuare ulteriori approfondimenti su uno o più aspetti specifici e di destinarli a impieghi differenti. Dalle

[50] Se si tratta di una forma non flessa, il valore di *type* prevede solo la *POS*.
[51] Cfr. *TEI*.

indagini statistiche allo studio di fenomeni fonetici, morfologici, sintattici, dalle osservazioni stilistiche agli impieghi lessicografici, l'analisi lessicale trattata con il linguaggio *XML-TEI* mostra tutta la sua versatilità e ci conferma l'importanza di esaminare i testi franco-italiani per renderli maggiormente interrogabili e accessibili.

4.1. *Un esempio di visualizzazione dei dati*

Il primo formato proposto per l'*export* rivela appieno la sua praticità quando si vuole passare direttamente alla pubblicazione del *corpus* analizzato. Nel nostro caso, siamo infatti partiti dal file *csv* per caricare nella biblioteca digitale del *RIALFrI* il testo dell'*Entrée d'Espagne* annotato e lemmatizzato, restituendone la *mise en page* e la punteggiatura dell'edizione critica[52]. Accedendo al sito, l'utente ha la possibilità di consultare il testo interattivo dell'opera. Posizionando il cursore sopra una parola, essa viene evidenziata in giallo (fig. 6)

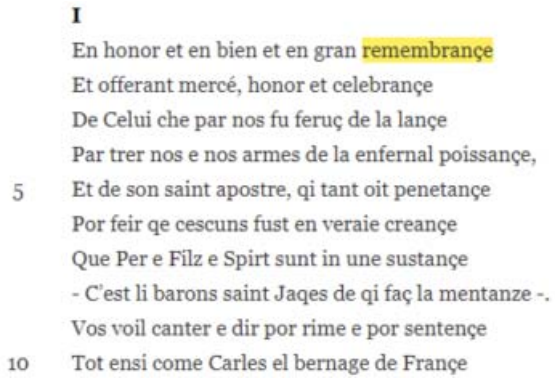
I
En honor et en bien et en gran remembrançe
Et offerant mercé, honor et celebrançe
De Celui che par nos fu feruç de la lançe
Par trer nos e nos armes de la enfernal poissançe,
Et de son saint apostre, qi tant oit penetançe
Por feir qe cescuns fust en veraie creançe
Que Per e Filz e Spirt sunt in une sustançe
- C'est li barons saint Jaqes de qi faç la mentanze -.
Vos voil canter e dir por rime e por sentençe
Tot ensi come Carles el bernage de Françe

Fig. 6. Esempio di visualizzazione del testo e di evidenziazione di una parola nel *RIALFrI*.

e, cliccando, compare un *pop-up* con il lemma di riferimento e lo scioglimento dell'annotazione morfosintattica (fig. 7). Come si può notare dall'immagine sottostante, l'attributo SPEC=it, ove presente, nella visualizzazione viene reso con 'italianismo'.

[52] Per il supporto informatico si ringrazia ancora una volta Luigi Tessarolo, responsabile tecnico del *RIALFrI*.

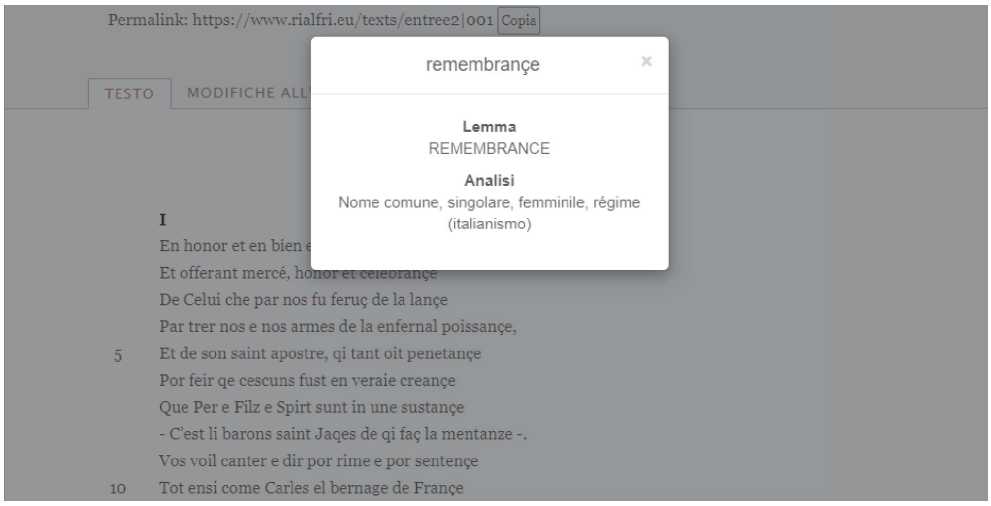

Fig. 7. Esempio di visualizzazione dell'analisi lessicale di un vocabolo nel *RIALFrI*.

L'utente può inoltre servirsi del motore di ricerca della piattaforma (fig. 8) per impostare una ricerca per forme (estesa a tutti i testi della banca dati) o per lemmi (limitata ai testi annotati e lemmatizzati):

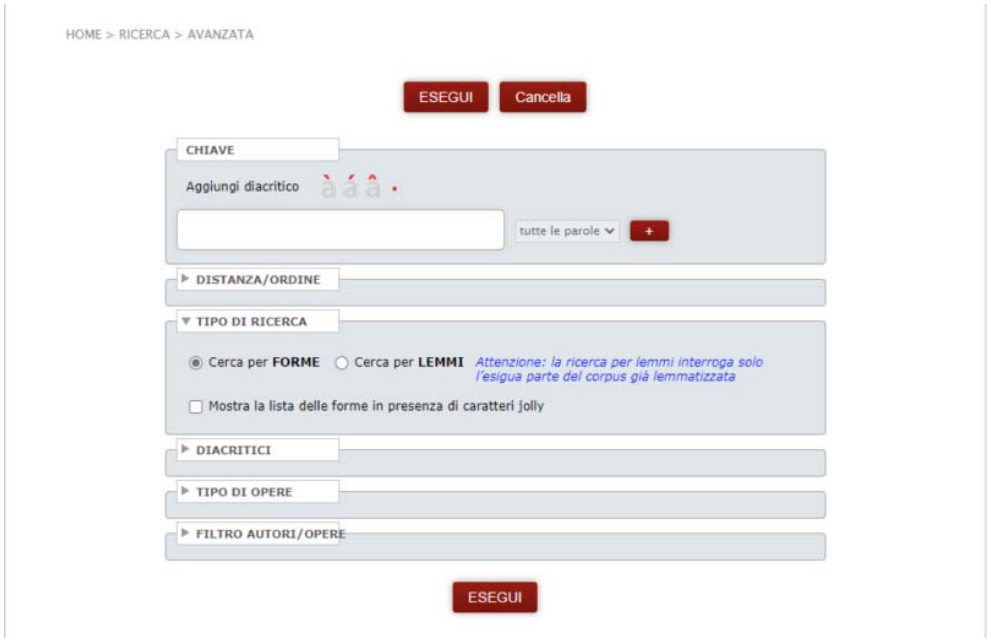

Fig. 8. Schermata di ricerca per forme e per lemmi nel *RIALFrI*.

5. *Conclusioni*

Grazie a questa sperimentazione abbiamo potuto apprezzare e valutare il contributo della tecnologia informatica e delle metodologie NLP (*Natural Language Processing*) nell'analisi lessicale di un testo. Abbiamo inoltre confermato quanto sia ancora fondamentale l'intervento dell'uomo, soprattutto nelle fasi di ideazione, sviluppo e controllo dei mezzi utilizzati. La sfida posta dal franco-italiano (lo scarto tra la realtà testuale concreta e la norma linguistica astratta) si moltiplica in modo esponenziale quando viene trasferita alla macchina. Essa, infatti, non è ancora in grado di contestualizzare e di discriminare come fa l'uomo; tende invece ad omologare, oltre che funzionare sulla base di regole predefinite. Il punto cruciale che abbiamo voluto evidenziare attraverso il nostro lavoro è che la multiformità del franco-italiano spinge gli strumenti informatici verso il punto di rottura.

Sulla base dell'esperienza provata con l'*Entrée d'Espagne*, riteniamo che la configurazione dell'attuale modello di analisi lessicale per l'antico francese sia adeguata anche per il franco-italiano. Tuttavia, abbiamo dimostrato che alcuni elementi necessitano di essere riqualificati o ripensati profondamente, anche alla luce delle finalità e dell'accuratezza che si desiderano attribuire all'annotazione morfosintattica e alla lemmatizzazione. Si impongono un ampliamento della lista dei lemmi attraverso l'integrazione di più dizionari, una revisione delle etichette a seconda delle informazioni lessicali da codificare, un cambiamento nello schema di annotazione morfosintattica per consentire l'indicazione dei diversi gradi di mescidanza linguistica.

Benché si sia trattato di un semplice banco di prova, lo studio condotto sull'*Entrée d'Espagne* ha comunque fornito una cospicua quantità di dati strutturati e ha messo a disposizione per la consultazione il testo interattivo dell'opera. In aggiunta, sono stati mossi i primi passi verso l'elaborazione di un nuovo modello di analisi lessicale costruito appositamente per il franco-italiano. Un obiettivo a medio termine che, verosimilmente, potrebbe lanciare una campagna di annotazione e lemmatizzazione su larga scala della produzione letteraria franco-italiana e, magari, favorire il raggiungimento di un altro obiettivo, stavolta a lungo termine: la realizzazione di un dizionario del franco-italiano.

Bibliografia

I. Opere

Entrée d'Espagne

Anonimo Padovano, *L'*Entrée d'Espagne*. Rolando da Pamplona all'Oriente*, a cura di Marco Infurna, Roma, Carocci, 2011 («Biblioteca Medievale», 133).

*L'*Entrée d'Espagne, chanson de geste franco-italienne publiée d'après le manuscrit unique de Venise par Antoine Thomas, 2 voll., Paris, Didot, 1913 («Société des anciens textes français»).

II. Studi e strumenti

Abeillé 2003

Treebanks. Building and Using Parsed Corpora, edited by Anne Abeillé, Dordrecht, Kluwer Academic Publishers, 2003.

ATILF

Analyse et Traitement Informatique de la Langue Française, CNRS – Université de Lorraine, https://www.atilf.fr/ [cons. 15. VII. 2021].

BFM

Base de Français Médiéval, Lyon, École Normale Supérieure, http://bfm.ens-lyon.fr/ [cons. 15. VII. 2021].

Camps 2016

Jean-Baptiste Camps, *Geste: un corpus de chansons de geste*, avec la collaboration d'Elena Albarran, Alice Cochet et Lucence Ing, Paris, 2016-, http://github.com/Jean-Baptiste-Camps/Geste [cons. 15. VII. 2021].

Camps – Clérice – Pinche 2020

Jean-Baptiste Camps, Thibault Clérice, Ariane Pinche, *Stylometry for Noisy Medieval Data: Evaluating Paul Meyer's Hagiographic Hypothesis*, 2020, http://arxiv.org/abs/2012.03845.

Clérice – Pilla – Camps 2019

Thibault Clérice, Julien Pilla, Jean-Baptiste Camps, *hipster-philology/pyrrha: 2.1.0*, 2019, http://doi.org/10.5281/zenodo.3524771 [cons. 15. VII. 2021].

DEAF

Dictionnaire étymologique de l'ancien français, fondé par Kurt Baldinger; avec la collaboration de Jean-Denis Gendron et Georges Straka; [puis] publié sous la direction philologique de Frankwalt Möhrenéd, Québec – Tübingen – Paris, PU Laval – Niemeyer – Klincksieck, 1974-2016, online al sito: http://www.deaf-page.de

DMF

Dictionnaire du Moyen Français (1330-1500), version 2020 (DMF 2020), ATILF – CNRS – Université de Lorraine, http://www.atilf.fr/dmf/ [cons. 15. VII. 2021].

FEW
Französisches etymologisches Wörterbuch. Eine Darstellung der galloromanischen Sprachschatzez, von Walther von Wartburg, continué sous la direction de Jean-Pierre Chambon et Jean-Paul Chauveau, 25 voll., Bonn – Heidelberg – Leipzig-Berlin –Bâle, Klopp –Winter –Teubner – Zbinden, 1928-2002.

Gdf
Frédéric Godefroy, *Dictionnaire de l'ancienne langue française et de tous ses dialectes du IX^e^ au XV^e^ siècle*, 10 voll., Paris, Vieweg, 1881-1902 [New York, Kraus Reprint, 1961], online al sito: http://micmap.org/dicfro/search/dictionnaire-godefroy/.

Geste
Geste: un corpus de chansons de geste, https://dev.chartes.psl.eu/elec/geste/ [cons. 15. VII. 2021].

Guillot – Lavrentiev – Prévost 2013a
Céline Guillot, Alexei Lavrentiev, Sophie Prévost, *Principes d'annotation Cattex09*, Version 2.0, Lyon, BFM – Base de Français Médiéval – École Normale Supérieure (Laboratoire ICAR), 2013, http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_principes_2.0.pdf.

Guillot – Lavrentiev – Prévost 2013b
Céline Guillot, Alexei Lavrentiev, Sophie Prévost, *Manuel de référence du jeu Cattex09*, Version 2.0, Lyon, BFM – Base de Français Médiéval – École Normale Supérieure (Laboratoire ICAR), 2013, http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_manuel_2.0.pdf.

Lenci – Montemagni – Pirrelli 2020
Alessandro Lenci, Simonetta Montemagni, Vito Pirrelli, *Testo e computer. Elementi di linguistica computazionale*, Roma, Carocci, 2020 («Aulamagna», 12).

LGeRM
Lemmatisation de la variation graphique des états anciens du français et lexiques morphologiques, ATILF – CNRS – Université de Lorraine, http://www.atilf.fr/LGeRM/ [cons. 15. VII. 2021].

Manjavacas – Kádár – Kestemont 2019
Enrique Manjavacas, Ákos Kádár, Mike Kestemont, *Improving Lemmatization of Non-Standard Languages with Joint Learning*, in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, Minneapolis, Association for

Computational Linguistics, 2019, vol. I, pp. 1493-1503, https://arxiv.org/pdf/1903.06939.pdf.

Manjavacas – Clérice – Kestemont 2021
Enrique Manjavacas, Thibault Clérice, Mike Kestemont, *emanjavacas/pie v0.3.7c*, 2 marzo 2021, http://doi.org/10.5281/zenodo.4572585 [cons. 15. VII. 2021].

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*, édités par Timothée Premat, Ariane Pinche, Lyon, Diachronies Contemporaines, 2019, pp. 48-58.

Pyrrha
Pyrra. A language independant post correction app for POS and lemmatization. Development environment, https://dev.chartes.psl.eu/pyrrha/ [cons. 15. VII. 2021]; *production environment*, https://dh.chartes.psl.eu/pyrrha/ [cons. 15. VII. 2021].

RIALFrI
Repertorio Informatizzato Antica Letteratura Franco Italiana, diretto da Francesca Gambino, Dipartimento di Studi Linguistici e Letterari, Università degli Studi di Padova, https://www.rialfri.eu/rialfriWP/ [cons. 15. VII. 2021].

TEI
TEI P5: Guidelines for Electronic Text Encoding and Interchange, versione 4.2.2, by the TEI Consortium, http://www.tei-c.org/Guidelines/P5/ [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.

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