Note di lavoro intorno alla creazione di una struttura di analisi lessicale (Roman de Troie Prose 2, ms. Grenoble BM 861)

Marta Materni

Francigena

8 (2022)

Note di lavoro intorno alla creazione di una struttura di analisi lessicale (*Roman de Troie Prose 2*, ms. Grenoble BM 861)

Marta Materni
(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/V8-008
DOI: 10.25430/2420-9767/V8-008

Note di lavoro intorno alla creazione di una struttura di analisi lessicale (*Roman de Troie Prose 2*, ms. Grenoble BM 861)*

Marta Materni
marta.materni@gmail.com

(Università degli Studi di Padova)

Asbtract:
Il contributo offre un resoconto dell'esperienza di progettazione di un'interfaccia di lemmatizzazione all'interno del progetto PRODIGI, il cui obiettivo è l'annotazione lessicale completa del testo di *Prose 2* trasmesso dal ms. di Grenoble BM 861. La costruzione dello strumento è stata l'occasione per riflettere sulle principali criticità e ambiguità del linguaggio di cui è necessario tener conto nel momento in cui la sua complessità viene setacciata dal filtro del formalismo informatico.

This work offers a report on the design experience of a lemmatization interface within the PRODIGI project, whose aim is the complete lexical annotation of *Prose 2* text as transmitted by the ms. Grenoble BM 861. The creation of the tool was an opportunity to reflect on the main criticalities and ambiguities of the language, that need to be taken into account when its complexity is sifted through the filter of digital formalism.

Parole chiave: *Roman de Troie en prose* – *Prose 2* – Lemmatizzazione – Annotazione linguistica – Lessicografia
Keywords: *Roman de Troie en prose* – *Prose 2* – Lemmatization – Linguistic annotation – Lexicography

> Per rendere un testo operabile al computer, bisogna prima di tutto che sia microanalizzato nei suoi passi elementari, millimetro per millimetro. Quando avessimo formulato e definito quali sono i passi elementari che fa la nostra mente, nel suo complesso, per riassumere, formalizzarli numericamente, bitizzarli in un computer, sarebbe un gioco. Ma la difficoltà sta proprio nel fatto che noi non conosciamo abbastanza le nostre operazioni mentali.
> Noi conosciamo le parole dai denti in fuori, ma dai denti in dentro siamo ancora un mistero per noi stessi. Per cui, l'industria dell'informatica ha bisogno – se parliamo di linguistica applicata – di più umanesimo. Se io chiamo umanesimo lo studio, la riflessione dell'espressione umana nei suoi vari aspetti, questo resta vero. Quel computer che agli inizi della cibernetica, tra gli anni '50 e '60, veniva presentato come la minaccia per l'intelligenza umana, oggi sfida l'intelligenza umana e dice: "Tu non conosci abbastanza te stessa e se io computer non riesco a servirti di più è perché tu non sai programmarmi, non sai alimentare di sufficienti dati i miei programmi[1].

* This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 886478.

[1] Busa 2001.

Le frasi citate, e tratte da uno dei tanti interventi informali di padre Roberto Busa, ben sintetizzano l'approccio al macrocosmo informatico, e in particolare al microcosmo trasversale (rispetto a una molteplicità di discipline) che è la cosiddetta 'informatica umanistica', adottato da parte di chi scrive e che è alla base dell'elaborazione di due progetti di ricerca finanziati da una Individual Fellowship Marie Curie: il progetto DIGIFLOR (n° 745821, *Digital Edition of the 'Roman de Florimont'*) e, più direttamente legato alle note che qui si propongono, il progetto PRODIGI (n° 886478, *'Prose 2' Digital Lemmatized Edition*). I due progetti, distinti ma complementari, hanno permesso di esplorare in maniera approfondita due degli approcci possibili all'oggetto Testo nell'ambito della filologia digitale: quello strettamente editoriale, rivolto al testo in quanto manifestazione di un'Opera tramite il vettore Documento (specificità che emerge soprattutto nel mondo manoscritto), e quello analitico (una delle tante possibili prospettive analitiche), rivolto al Documento del Testo in quanto *specimen* o *corpus* microcosmico di un determinato codice linguistico. Nel passare dal livello della forma documentaria a quello dell'analisi del codice linguistico si compie il passaggio da un problema di rappresentazione della conoscenza – del testo e dei metadati a esso collegati, attività al cuore dei linguaggi di *markup* o codifica testuale –, a un problema di produzione della conoscenza, ponendosi così al centro di quella che padre Busa insisteva nel definire l'informatica ermeneutica in contrapposizione all'informatica editoriale e a quella documentaria.

1. *Il progetto PRODIGI e ULA*

Fig. 1

Il progetto PRODIGI (entrato nella sua fase finale al momento della scrittura di questo contributo) ha avuto un duplice obiettivo: la valorizzazione linguistica del *corpus* manoscritto della seconda prosificazione del *Roman de Troie* o *Prose 2* e la realizzazione di uno strumento di annotazione linguistica funzionale alla gestione di un caso complesso come quello da cui ha preso spunto (siamo, come si specificherà oltre, nell'ambito del 'francese d'Italia'), senza però trasformarsi in uno strumento *ad hoc*, specificatamente modellato sulle esigenze proprie del testo in esame e utilizzabile esclusivamente in questo ambito. Da questa prospettiva deriva il nome attribuito allo strumento: ULA, per *Universal Language Annotation*, dove l'aggettivo *universal* è da intendersi nel senso di *cross-language*.

L'infrastruttura di lemmatizzazione ULA è a disposizione di qualsiasi utente la trovasse ergonomica e funzionale al proprio lavoro di annotazione, a prescindere dalla lingua su cui si sta lavorando, sebbene sia necessario rimanere nell'ambito

di lingue espresse in caratteri latini e fondate sulla nozione di parola. In questo senso, lo si ribadisce, l'aggettivo *universal* va inteso come *cross-language* ma non come ricerca di una metodologia a valore universale.

Una seconda precisazione è importante per inquadrare le note che seguono. L'annotazione lessicale[2] di *Prose 2* con strumenti informatici nasce all'interno di una prospettiva filologica con finalità essenzialmente lessicologiche/lessicografiche: se il progetto potesse continuare (lo si illustrerà nel paragrafo successivo), il suo obiettivo finale sarebbe sostanzialmente la creazione di un 'dizionario d'opera' orientato verso la valorizzazione della variazione. Progetto e relativo strumento vogliono quindi mantenersi coscientemente al di sotto, o al di qua, delle grandi tematiche e sfide della linguistica computazionale, facendo confluire i risultati piuttosto nel campo della linguistica storica. Il che non toglie che, per quel che riguarda i formati di codifica dei dati, siano state operate scelte che, si spera, possano garantire la loro riusabilità anche con altre prospettive, in quell'ottica di FAIR-izzazione dei dati che fa parte ormai della deontologia di qualsiasi progetto nato sotto il segno di un finanziamento europeo[3].

Progettare uno strumento significa inoltre riflettere su categorie e astrazioni; progettare uno strumento nell'ambito dell'applicazione dell'informatica alle tradizionali pratiche umanistiche significa in aggiunta riflettere su abitudini ormai diventate automatiche, rendere esplicito e sequenziale ciò che il nostro cervello elabora in un sol gesto, formalizzare, cioè esplicitare in ogni minimo dettaglio, i nostri processi di analisi. In questo senso, il contributo dell'informatica umanistica, al di là dei risultati, è anche, se non talvolta soprattutto, in questo esercizio intellettuale che va ben oltre gli aspetti meramente tecnici. Da un certo punto di vista, la costruzione di un modello informatico in ambito umanistico può anche fungere da quella che in statistica viene definita analisi fattoriale confermativa: l'ottenimento di un risultato in linea con quanto già noto, ma attraverso un processo completamente differente basato sulla formalizzazione, rappresenta una conferma tutt'altro che superflua ma che al contrario rafforza l'affermazione.

[2] Adotto la formula annotazione lessicale per riunire sotto un'unica espressione le due operazioni distinte di lemmatizzazione e annotazione linguistica (nei casi in cui si faccia riferimento a una sola delle due operazioni lo si espliciterà), seguendo la formula *étiquetage lexical* proposta nell'ottimo, per chiarezza espositiva, contributo di sintesi di Laporte 2000 (28): «Nous regrouperons sous le terme d'étiquetage lexical l'ensemble des techniques qui concourent à passer d'un texte brut, exempt d'informations linguistiques, à une séquence de mots étiquetés par des informations linguistiques, au premier rang desquelles les informations morphologiques et grammaticales. Cette définition inclut donc la délimitation des mots, la morphologie et la levée des ambiguités lexicales». Cfr. anche Laporte 1997 e Glessgen 2003.

[3] FAIR: *Findability, Accessibility, Interoperability and Reuse [of digital assets]* https://www.go-fair.org/fair-principles/ [cons. 24. VII. 2022]. Sempre in quest'ottica, i *files* creati durante il progetto (vale a dire i *files* contenenti i risultati dell'annotazione) saranno depositati nel *repository* GitHub https://github.com/digiflor, in una sezione appositamente dedicata al progetto attuale.

1.1. *PRODIGI in prospettiva*

La seconda prosificazione del *Roman de Troie*, *ante* 1298, è stata scelta per quest'esperienza perché tanto il testo in sé per sé quanto l'insieme dei suoi testimoni manoscritti presentano una serie di motivi di interesse. Il testo, che, fra le cinque prosificazioni del *Roman*, è l'unico ancora inedito[4] (ma non è in questo dato che risiede il suo interesse), appartiene infatti alla galassia, particolarmente in auge, della francofonia medievale *outside France*[5] (il riferimento è ai progetti del recentemente scomparso prof. Simon Gaunt), e, all'interno del sotto-insieme rappresentato dall'antica letteratura franco-italiana (secondo la definizione del RIALFrI, *Repertorio Informatizzato Antica Letteratura Franco-Italiana*)[6], alla ulteriore sotto-categoria delle 'opere originali scritte da autori italiani in francese' (secondo la nomenclatura proposta ancora dal RIALFrI). Il testo è tràdito da tre testimoni integrali, tutti e tre esemplati in Italia, dove nasce peraltro la stessa prosificazione, e, fortunatamente per noi dal punto di vista linguistico, in tre differenti aree in termini di varietà: Padova (Grenoble BM 861), 1298; Genova (Paris BnF n.a.fr. 9603), ultimo ventennio del XIII sec.; Verona (Oxford BL Douce 196), copiato nel 1323 dal notaio Pietro di Bonaventura Scacchi. L'insieme della tradizione manoscritta di questo testo si converte quindi in un interessantissimo *corpus* linguistico di osservazione dei fenomeni di interferenza tra la *langue d'oïl* e le varietà italiane-settentrionali, mostrando il medesimo testo attraverso tre prismi con caratteristiche differenti. Al tempo stesso, la lunghezza dell'opera, 132 cc. nel ms. di Grenoble (cifre esatte circa il numero di occorrenze e di lemmi potranno essere fornite al termine del progetto, previsto per gennaio 2023), è tale da offrire un'ampia selezione lessicale. Infine, questa prosificazione è stata oggetto di un fedelissimo volgarizzamento in toscano ad opera di Binduccio dello Scelto[7], testimoniato dal ms. Firenze BN II-IV-45. Il *corpus* nella sua interezza permette quindi, al di là dei risultati immediati legati al biennio finanziato dalla borsa Marie Curie, di pianificare un'evoluzione futura che potrebbe portare alla creazione di un dizionario di *Prose 2* costruito a partire dai tre testimoni (tutti lemmatizzati), teso a valorizzare non solo un particolare bagaglio lessicale, nella forma quindi di un dizionario d'autore (o, meglio, d'opera), ma anche ad apprezzare e quantificare la variazione di questa selezione lessicale attraverso tre aree dialettali distinte. L'esistenza della traduzione di Binduccio permetterebbe poi di completare il tutto con una sorta di 'dizionario bilingue', ovviamente limitato agli usi e interpretazioni

[4] Per uno *status questionis* aggiornato sulle prosificazioni del *Roman de Troie* si rimanda all'introduzione di Anne Rochebouet alla sua recente edizione di *Prose 5* (2021).
[5] *Medieval Francophone Literary Culture Outside France* (https://medievalfrancophone.ac.uk/) [cons. 24. VII. 2022].
[6] https://www.rialfri.eu/ [cons. 24. VII. 2022].
[7] Cfr. Carlesso 1966; Binduccio dello Scelto, *Storia di Troia*, (ed. Gozzi 2000, ed. Ricci 2004).

che di *quel* bagaglio lessicale vengono dati in *quel* testo da *quell'*autore. L'utilizzo di trascrizioni di manoscritti, e non di edizioni critiche, per creare il *corpus* linguistico fornirebbe risposta alla problematica segnalata ad es., fra gli altri, da Martin Glessgen:

> Toute la lexicologie aussi bien que la lexicographie de l'ancien français repose sur les formes plus ou moins normalisées des éditions critiques, en dépit du jugement des philologues qui sont bien conscients de ce problème[8].

1.2. *PRODIGI oggi*

Alla data attuale, per porre le basi dei possibili sviluppi illustrati nel paragrafo precedente, il progetto PRODIGI si è concretizzato nella creazione dello strumento di annotazione lessicale ULA e nella realizzazione dell'annotazione lessicale del ms. di Grenoble. Si tratta del manoscritto meno studiato fra i tre, esplicitamente datato (1298), coevo di quello parigino, con il quale forma la coppia degli *antiquiores,* e firmato dal copista, Johannes de Stennis[9]. La scelta del testimone esula da valutazioni di tipo stemmatico relative alla lezione di *Prose 2* in quanto testo di un'Opera, dal momento che nella nostra prospettiva ciò che interessa è la dimensione del testo del singolo manoscritto in quanto cristallizzazione di un codice linguistico; in questo caso i tre testimoni sono dunque fra loro in una posizione di parità e di pari interesse, dato che ognuno fa riferimento a un contesto linguistico differente. A livello di manufatto il ms. Grenoble appartiene a quell'interessante categoria delle opere trascritte da 'copisti-prigionieri' ampiamente studiata soprattutto nell'ambito pisano-genovese[10]. Ragioni quindi di carattere puramente storico e non testuale, oltre all'occasionale legame fra la città di Padova, dove il manoscritto è stato esemplato, e quella di Grenoble, dove è conservato, legame speculare al filo di continuità tematica che unisce i due progetti DIGIFLOR, ospitato a Grenoble, e PRODIGI, ospitato a Padova, hanno portato alla scelta di questo manoscritto, per dare avvio alla costituzione del *corpus Prose 2*.

2. *Lo scheletro del sistema*

La progettazione di uno strumento non può prescindere ovviamente da un'operazione di 'scavo archeologico', sia fra gli strumenti già in uso sia fra le proposte

[8] Glessgen 2003b: 58. Cfr. anche Korfanty 1999 e Glessgen 2003a. Per un panorama sulla questione della tipologia di edizione su cui fondare un nuovo *corpus* ampliato dell'*ancien français*, cfr. Tittel 2015.

[9] Un primo studio dettagliato è quello di Fois 2021.

[10] Dopo il classico Cigni 2006, cfr. la più recente rassegna di Cambi 2016. Cfr. anche Cursi 2009 per un altro *atelier* di copia.

progettuali già avanzate, con l'obiettivo di far tesoro delle soluzioni più riuscite e che si sono rivelate efficaci, entrando così nella pratica comune. In questo senso, nel paragrafo successivo si analizzeranno nel dettaglio le caratteristiche di due strumenti adottati nell'ambito di progetti dedicati a testi analoghi a quello al centro di PRODIGI, i quali hanno rappresentato perciò un punto di riferimento e una base di partenza per l'elaborazione dell'interfaccia ULA. I due strumenti risultano particolarmente interessanti da comparare in quanto rappresentano esattamente i due poli opposti nell'approccio possibile alla questione annotazione lessicale in ambito informatico: lo 'storico' sistema LGeRM, elaborato da Gilles Sovay in seno all'ATILF (*Analyse et Traitement Informatique de la Langue Française*)[11], utilizzato fra l'altro nella costruzione del DMF (*Dictionnaire du Moyen Français*)[12]; e il 'nuovo' sistema *Pyrrha*, frutto di un lavoro d'*équipe* con base all'École des Chartes[13].

Accanto al lavoro di analisi delle soluzioni in uso, è necessario anche individuare quali siano i nodi, gli elementi costitutivi di cui si deve tener conto sia nella valutazione del già esistente che nell'elaborazione del nuovo; elementi costitutivi del sistema che rappresentano altrettanti potenziali punti sensibili suscettibili di far emergere criticità. Questa la griglia di analisi elaborata, che ha funzionato al tempo stesso da promemoria o scheletro per una nuova elaborazione progettuale:

1. principi di gestione del processo di lemmatizzazione;
2. formato di *input* dei dati;
3. interfaccia di lavoro/immissione dei dati;
4. interfaccia di correzione;
5. formato di archiviazione dei dati;
6. formato di *output* dei dati;
7. modalità di visualizzazione dei dati;
8. modello di interrogazione dei dati.

[11] https://www.atilf.fr/ [cons. 24. VII. 2022].
[12] http://zeus.atilf.fr/dmf/ [cons. 24. VII. 2022].
[13] Fra i vari sistemi esistenti, si è scelto di illustrare nel dettaglio questi due perché ampiamente diffusi per l'analisi linguistica di testi francesi medievali; recentemente *Pyrrha* è stato adottato anche dal gruppo RIALFrI (si veda al proposito il rendiconto Ceresato 2021). Si trattava inevitabilmente del confronto più ovvio dato il testo su cui si sta lavorando oltreché il personale ambito di formazione linguistica, dal momento che questo retroterra permetteva di avere maggiore coscienza delle problematiche linguistiche specifiche e delle soluzioni adottate rispetto a strumenti dedicati ad altri ambiti linguistici. Si consideri ad es. la panoramica offerta da Gleim *(et alii)* 2019 con riferimento al tedesco e al latino.
[14] http://www.atilf.fr/LGeRM [cons. 24. VII. 2022].

2.1. *LGeRM (Lemmes Graphies et Règles Morphologiques)*[14]

Fig. 2

Sviluppato inizialmente per il *moyen français* (1350-1500), LGeRM è stato successivamente adattato al francese del XVI-XVII sec. ed è in grado di gestire anche la lemmatizzazione del francese moderno. La bibliografia, perlopiù a firma Gilles Sauvay e Sylvie Bazin-Tacchella, è liberamente consultabile nella sua interezza sulle rispettive pagine HAL dei due studiosi, alle quali si rimanda[15].

LGeRM non è mai stato implementato per un utilizzo in locale (benché nella pagina web di accesso alla piattaforma si possa leggere: «Une version utilisable en local sous Windows est en cours de développement»)[16].

Col nome di LGeRM si indica ancor più precisamente l'intera piattaforma di lemmatizzazione (la definizione estesa e completa dello strumento è infatti *Plateforme de lemmatisation de la variation graphique des états anciens du français*), dietro la quale si collocano i due pilastri strutturali del sistema: da una parte il lemmatizzatore vero e proprio, e dall'altro i lessici morfologici, uno pertinente al *moyen français* e l'altro al francese del XVI-XVII sec. I lemmi per il *moyen français* sono ricavati dal DMF, quelli del XVI-XVII sec. provengono dal TLF (*Trésor de la Langue Française*).

Si tratta di un modello di lemmatizzatore basato su regole (6.500 regole di cui 4.500 solo per la flessione verbale) e quindi *language dependent*, e il suo principale punto di forza risiede nella capacità di gestire la variazione grafica, dove per variazione grafica si deve intendere tanto la variazione grafica *stricto sensu*, una caratteristica che potremmo definire storico-contestuale, che la flessione, proprietà intrinseca alla tipologia linguistica, una lingua flessionale. Il lemmatizzatore ricerca le forme nei lessici morfologici senza tener conto del contesto; quando si trova di fronte a una forma sconosciuta, applica delle regole morfologiche per ricondurla a una forma nota.

Altro punto a favore di LGeRM è la sua elasticità di fronte ai dati di *input*: è infatti capace di gestire una trascrizione diplomatica riconoscendo i «caractères non modernisés u-v-i-j, s long, tilde de nasalisation. Il est capable de gérer les

[15] https://hal.archives-ouvertes.fr/ [cons. 24. VII. 2022].

[16] Alla voce del menu *Exécutable* leggiamo inoltre: «Distribution de l'exécutable. L'exécutable n'est pas officiellement distribué car pas assez testé et pas de procédure d'installation automatique fiable… Néanmoins vous pouvez me contacter pour en obtenir une version, il faudra que je vous aide à l'installer. Il fonctionne uniquement sous Windows et sur les caractères ISO-8859-1».

agglutinations standards des prépositions, adverbes et articles»[17]. Tuttavia il *tagset* LGeRLM ha una granularità di analisi piuttosto bassa dal momento che per ciascun lemma viene indicata solo la categoria grammaticale (POS).

Il *Formulaires en ligne* (si ricorda che LGeRM non è utilizzabile in locale) offre tre possibilità di utilizzo:

1. studio di una singola forma;
2. lemmatizzazione di un testo (copiato nello spazio specifico);
3. lemmatizzazione di un file <100 Kb.

Il formato dei dati di *input* può essere:

1. testo puro;
2. testo codificato in XML-TEI;
3. tabellare con etichette CATTEX, PRESTO, FRANTEXT[18].

Il formato di *output* a sua volta può essere:

1. file .CSV;
2. tabellare con etichette;
3. codifica XML-TEI nella forma
   <w lemma="" type="(POS)">
   ad es. <w lemma="chevalier" type="subst"> .

La piattaforma LGeRM propone anche un *Outil Glossaire* per la costruzione appunto di un glossario. Tuttavia, le condizioni d'utilizzo sono abbastanza restrittive:

> L'outil glossaire est libre d'accès dans le cadre d'une collaboration, formelle ou non, avec l'équipe du DMF. Pour utiliser l'outil, il faut nous transmettre un texte, nous le lemmatisons, nous créons un compte pour accéder aux résultats de la lemmatisation et à l'ébauche du glossaire. […] La version actuelle ne permet pas une autonomie complète de l'utilisation de l'outil[19].

[17] Si cita dalla pagina *Lemmatiseur/Introduction/Texte normalisé ou diplomatique* a partire dal Menu principale.
[18] Per CATTEX v. oltre. PRESTO (*Projet ANR/DFG: l'évolution du système prépositionnel du français*): progetto franco-tedesco coospitato fra il 2013 e il 2017 all'università di Lyon 2 e all'università di Köln; ha avuto per obiettivo «l'étude diachronique de l'emploi, des valeurs sémantiques et discursives des prépositions français *à*, *en*, *par*, *contre*, *dès*, *devant*, *entre*, *pour*, *sans*, *sur*, *sous*, *vers*, *dans*, de l'ancienne langue jusqu'au français contemporain», http://presto.ens-lyon.fr/ [cons. 24. VII. 2022]. FRANTEXT: base testuale sviluppata all'ATILF e che si caratterizza per il fatto di presentare un *corpus* costituito da testi distribuiti dal IX al XXI sec., permettendo ricerche cumulative per forme, lemma e categoria grammaticale, su un totale di, attualmente, 264.000.000 di occorrenze, https://www.frantext.fr/ [cons. 24. VII. 2022].
[19] Si cita dalla pagina *Plateforme/Conditions d'accès* del Menu principale.

L'ultima versione del lessico morfologico LGeRM, che alla data del 17. IX. 2021 conta 955.000 forme, è codificata in formato XML nella forma:

```
<lemmatizedForm>
    <orthography target="DM">aaiser</orthography>
    <grammatical Category>verbe</grammatical Category>
    <orthography target="TL">aaisier</orthography>
    <orthography target="AND"[20]>eiser1</orthography>
</lemmatizedForm>
```

Fig. 3

2.2. *Pyrrha*[21]

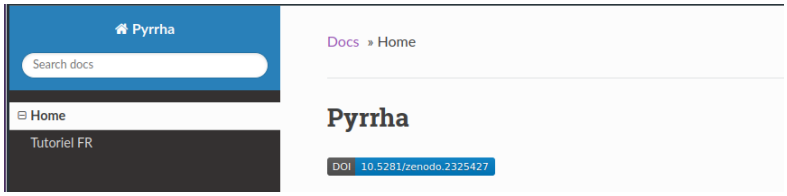


Fig. 4

Si tratta di un sistema sviluppato principalmente da Julien Pilla et Thibault Clérice all'École Nationale des Chartes. Benché a un primo approccio *Pyrrha* si presenti ufficialmente come una WebApp (l'utilizzo al momento è solo *on-line*) di post-correzione di *corpora* lemmatizzati e annotati morfo-sintatticamente, in realtà la piattaforma permette anche di lemmatizzare/annotare un testo avvalen-

[20] TL: *Tobler-Lommatzsch*; AND: *Anglo-Norman Dictionary.*

[21] Tutorial: https://pyrrha.readthedocs.io/en/latest/; documentazione: https://zenodo.org/record/2325428; versione di sviluppo: https://dev.chartes.psl.eu/pyrrha; repository: https://github.com/hipster-philology/pyrrha [cons. 24. VII. 2022].

dosi del *tagger* PIE (*A Framework for Joint Learning of Sequence Labeling Tasks*) sviluppato da Enrique Manjavacas e Mike Kestemont e rivolto, come possiamo leggere nella pagina di presentazione, a «variation-rich languages»[22]. Si tratta di un 'approccio' *machine learning* e quindi predittivo[23], che colloca la piattaforma *Pyrrha*, la quale si avvale del *tagger* PIE, all'estremo opposto dell''approccio' a regole della piattaforma LGeRM, benché entrambe abbiano lo stesso obiettivo: la gestione della variazione.

> Lemmatization of standard languages is concerned with (i) abstracting over morphological differences and (ii) resolving token-lemma ambiguities of inflected words in order to map them to a dictionary headword. [*Nel caso di*] non-standard historical languages […] the difficulty is increased by an additional aspect (iii): spelling variation due to lacking orthographic standards[24].

Per procedere alla correzione è necessario caricare il testo annotato e scegliere fra le *Listes de contrôle* disponibili (*ancien français*, francese moderno, LASLA per il latino) o caricarne di nuove; tali liste concernono lemmi, POS (*Part Of the Speech*) e MSD (*Morpho-Syntactic Description*). La presenza delle liste di controllo facilita la correzione in due modi: facendo emergere etichette non presenti (o perché del tutto assenti o perché scritte in modo erroneo); aiutando l'inserimento con dei suggerimenti di autocompletamento.

Il testo già annotato, in formato .TSV (*Table-Separated Value*) viene trasferito nello spazio di lavoro *on line* con un'operazione di copia-incolla e deve essere codificato secondo lo schema:

Form	*Lemma*	POS	*Morph*
son	son4	NOMCom	PERS.=3|NOMB.=s|GENRE=m|CAS=r

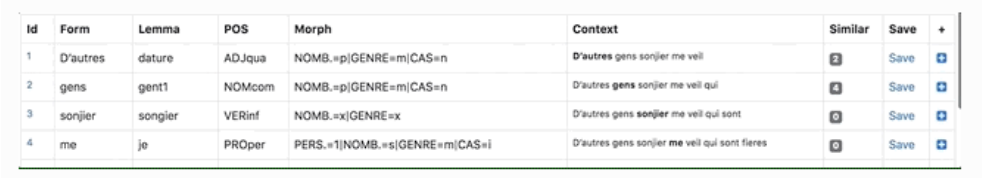
Nel caso in cui invece si utilizzi la piattaforma per lemmatizzare, si farà un copia-incolla del testo puro, il quale sarà sottoposto dal sistema a tokenizzazione e restituito sotto forma di lista, ogni *token* su una singola linea in una tabella predisposta con i campi obbligatori *Form*|*Lemma*|POS|*Morph*. Il servizio di lemmatizzazione è attualmente limitato al latino e al francese medievale, e si appoggia sui rispettivi modelli PIE, chiamati *Deucalion*[25]. Per quel che riguarda il contesto francese, di più diretto interesse per queste note, *Deucalion* utilizza i lemmi del TL, mentre il *tagset* è quello di CATTEX (v. oltre).

[22] https://doi.org/10.5281/zenodo.4572585 [cons. 24. VII. 2022]. Per la sua descrizione cfr. Manjavacas (*et alii*) 2019.

[23] Per altri esempi di applicazione di questa metodologia per affrontare la questione della lemmatizzazione, cfr. Eger (*et alii*) 2016 (latino e tedesco medievali), Dereza 2018 (antico irlandese), Kestemont (*et alii*) 2016 (tedesco medievale).

[24] Manjavacas (*et alii*) 2019: 1493.

[25] https://zenodo.org/record/2707476 [cons. 24. VII. 2022], per il latino; https://zenodo.org/record/3237455 [cons. 24. VII. 2022], per il francese medievale.



Id	Form	Lemma	POS	Morph	Context	Similar	Save	+
1	D'autres	dature	ADJqua	NOMB.=p|GENRE=m|CAS=n	**D'autres** gens sonjier me veil	2	Save	
2	gens	gent1	NOMcom	NOMB.=p|GENRE=m|CAS=n	D'autres **gens** sonjier me veil qui	4	Save	
3	sonjier	songier	VERinf	NOMB.=x|GENRE=x	D'autres gens **sonjier** me veil qui sont	0	Save	
4	me	je	PROper	PERS.=1|NOMB.=s|GENRE=m|CAS=i	D'autres gens sonjier **me** veil qui sont fieres	0	Save	

Fig. 5

L'interfaccia di correzione si presenta come una struttura a colonne, tre delle quali sono modificabili dall'utente per la correzione: *Lemma*, POS, *Morph*. La chiave del sistema è rappresentata dalla prima colonna, quella della *Form*, che nella schermata *Pyrrha* coincide col *token*, cioè con l'occorrenza, a differenza di quanto farà ULA (v. oltre). A livello di correzione è possibile anche usufruire dell'opzione di correzione per blocchi (*lots*): blocchi cioè di tutti i *token* identici, eventualmente da raffinare filtrandoli attraverso i campi *Lemma*, POS e *Morph* (opzione a cui far ricorso soprattutto in caso di omografia).

I dati così corretti (o, a seconda delle situazioni, lemmatizzati e corretti) possono essere esportati in due formati: .TSV, secondo il già citato schema *Form*|*Lemma*|POS|*Morph*; e XML-TEI nella forma:

<w xml:id="t6" n="6" lemma="qui" type="POS=PROrel|NOMB.=s|GENRE=m|CAS=n">**qui**</w>

(si fa notare che si opta per il generico attributo @type in luogo del sistema più specifico previsto dalle *Guidelines* e articolato in @pos @msd).

Rispetto alle opzioni offerte dai due strumenti analizzati, varie esigenze del progetto PRODIGI e del modello di analisi lessicale che si voleva proporre per *Prose 2* restavano insoddisfatte. Il principale elemento a sfavore dell'utilizzo di LGeRM è stato rappresentato dalla sua condizione di 'strumento chiuso', il cui utilizzo, per un testo di così ampie dimensioni, richiede il processo di lemmatizzazione assistita dall'équipe LGeRM. Quanto alle condizioni di utilizzo, entrambi gli strumenti, LGeRM e *Pyrrha*, sono concepiti solo per un uso *on line*: personalmente sono convinta dell'importanza che gli strumenti di lavoro siano utilizzabili (anche) in locale, demandando allo spazio Internet il ruolo di canale di diffusione ma non di spazio di lavoro esclusivo (altro discorso è quello dell'organizzazione di una rete di lavoro da remoto, che rappresenta una delle maggiori opportunità offerte dall'infrastruttura Internet).

Su un altro piano, nel caso di LGeRM risultava molto insoddisfacente il livello di dettaglio dell'annotazione, limitato sostanzialmente al POS. Benché quello di *Pyrrha* fosse più articolato, non coincideva tuttavia con il modello di analisi che si voleva proporre, e, come per altri strumenti, l'utente non può modificare il *tagset*. Parimenti, non è possibile intervenire autonomamente sulle voci del dizionario, per le quali è necessario inoltrare richiesta di modifica all'équipe. E i principi di tokenizzazione di *Pyrrha* non permettevano quel trattamento delle polirematiche che verrà illustrato nei paragrafi successivi.

La possibilità di un controllo diretto sul processo e di un adattamento degli elementi di analisi, sia a livello di *tagset* che di dizionario, oltre che l'esigenza di un utilizzo in locale, sono state fra le principali motivazioni che hanno spinto alla creazione dell'infrastruttura ULA, che si è dimostrata più ergonomica per chi scrive, e, si spera, potenzialmente anche per altri.

3. *Desiderata: annotare e correggere*

Qualsiasi operazione di analisi lessicale automatica comporta inevitabilmente una molteplicità, più o meno accentuata, di fasi successive di correzione. A meno che non si stia utilizzando un *corpus* di dati di dimensioni e con finalità tali, per esempio statistiche, che il rumore[26] derivante dall'errore sia ampiamente compensato dalla quantità stessa delle informazioni. Ma nel momento in cui i dati così ricavati diventano oggetto di studi prettamente linguistici, ancor più precisamente lessicografici, è evidente che la correttezza del dato stesso si rivela imprescindibile.

Nel caso specifico, ma non unico, di *Prose 2*, inteso come esponente del 'francese d'Italia', il forte grado di interferenza con le varietà dialettali italiane, scontrandosi con dizionari 'esterni' alla situazione linguistica in esame (cioè il dizionario dell'*ancien français* per *Pyrrha*), finiva per aumentare notevolmente il congenito e inevitabile rumore nei risultati. Sira Rodeghiero (2021), nel suo resoconto sull'esperienza di utilizzo di *Pyrrha* nella lemmatizzazione delle *Enfances Bovo*, ha sottolineato alcune delle difficoltà nate dallo scontro fra uno strumento rispetto ad alcuni aspetti 'chiuso' e testi così imprevedibili, concludendo che:

> 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 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[27].

[26] Su questa nozione di rumore rispetto ai risultati dell'analisi si consideri Laporte 2000: 33: «Pour l'étiquetage lexical comme pour de nombreuses autres activités consistant à associer à des éléments connus (le texte) des éléments pris parmi un stock (les étiquettes), les résultats obtenus ne correspondent pas toujours exactement aux résultats désirés. Cet écart se mésure par le bruit e le silence. On peut définir le taux de bruit comme la proportion d'étiquettes non désirées parmi les étiquettes présentées, et le taux de silence comme la proportion d'étiquettes non présentées parmi les étiquettes désirées. [*Si può inoltre definire*] le taux de précision (proportion d'étiquettes désirées parmi les étiquettes présentées), complémentaire du taux de bruit, et le taux de rappel (proportion d'étiquettes présentées parmi les étiquettes désirées), complémentaire du taux de silence».

[27] Ivi: 337-338.

La convivenza fra la complessità dell'oggetto linguistico e un dizionario/*tagset* immodificabile può talvolta rivelarsi una convivenza scomoda e insoddisfacente. Da qui la scelta che è stata fatta, nel caso di *Prose 2*, a favore di una lemmatizzazione parzialmente manuale (sottolineo parzialmente perché comunque, come si vedrà, andando avanti nel lavoro la percentuale di 'manualità' diminuisce notevolmente a favore dell''automaticità'), ma soprattutto a favore di un dizionario costruito in corso d'opera, interno al testo stesso piuttosto che esterno preesistente, e di un *tagset* liberamente modificabile dall'utente. Perlomeno in questa fase in cui non esiste ancora un dizionario di riferimento del 'francese d'Italia'.

Prose 2 presenta in sostanza una duplice problematicità: alla problematicità di gestione insita in una lingua còlta nelle sue fasi di pre-standardizzazione moderna si aggiunge la problematicità ulteriore di una varietà linguistica *de facto* esistente ma al tempo stesso difficilmente sistematizzabile come è quella del vasto bacino del/i francese/i copiato/i e/o scritti in Italia.

3.1. *Dal lemmatizzatore all'interfaccia di lemmatizzazione*

Come si è appena detto, possibilità di adattamento al contesto ed ergonomia sono state le due parole d'ordine che hanno guidato la concezione di ULA, oltreché le due esigenze specifiche legate alla lemmatizzazione di un testo come *Prose 2*. In questa prima fase, il lavoro si è pertanto concentrato sull'efficienza dell'interfaccia, e definirei più esattamente lo strumento messo a punto un'interfaccia di annotazione lessicale più che un lemmatizzatore vero e proprio. Sicuramente, dal punto di vista strettamente informatico, ma anche linguistico, il lavoro su un algoritmo di annotazione lessicale è una sfida intellettuale particolarmente appassionante, tanto quanto essa è complessa. Si tratta infatti di ricostruire e svelare i meccanismi che articolano il nostro discorso, riconducendo la varietà del reale a un nucleo fondante di regole chiaramente applicabili. In questo senso l'esercizio intellettuale di categorizzazione ed esplicitazione del reale che è alla base di un lemmatizzatore a regole non è meno intellettualmente affascinante dell'analisi dei risultati, sia in termini di riuscita che di errore, generati da un sistema probabilistico. In entrambi i casi, per vie opposte – a partire dall'origine o risalendo all'origine attraverso i risultati – si seguono gli auspici di padre Busa: si cerca di scoprire ciò che c'è «dai denti in dentro».

Da questo punto di vista, allo stato attuale ULA può considerarsi, sotto il profilo informatico, uno strumento 'semplice': esso si limita a creare un dizionario-macchina di associazioni forma-lemma che, avanzando nel lavoro, vengono poi riproposte dalla macchina precompilando una parte dei campi di analisi, in presenza di un nuovo testo. In caso di ambiguità da sciogliere, lo strumento propone le varie alternative (come LGeRM); un'evoluzione futura potrebbe permettere di ridurre il rumore di queste proposte alternative fino a, in alcuni casi, risolvere automaticamente e autonomamente le ambiguità sulla scorta dei

contesti delle occorrenze (con l'ovvio margine di errore insito in questo genere di operazioni).

Lo strumento ULA vuole ad oggi risultare utile soprattutto per chi avesse necessità di trattare dati linguistici complessi e variegati come è il caso del contesto linguistico a partire dal quale è stato elaborato e rispetto al quale risultava imprescindibile poter correggere con facilità e avere un certo margine di manovra nei confronti dei parametri del dizionario e del *tagset* di analisi.

4. *Guidelines ULA-PRODIGI*

4.1. *L'interfaccia di immissione dati ULA*

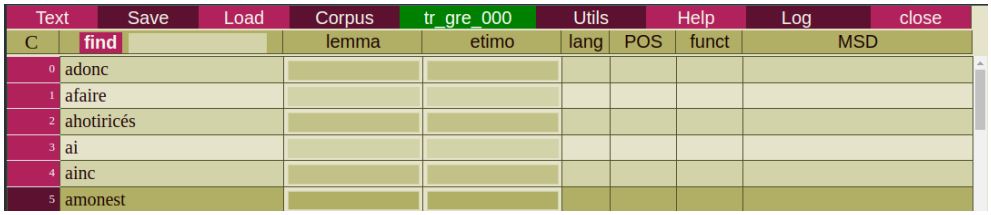


Fig. 6

Nell'elaborare l'interfaccia si è cercato di creare una sintesi fra i vantaggi che si scorgevano nei due sistemi di immissione/post-correzione possibili: quello a testo (che segue la linea testuale originale), che definirei 'contestuale', e quello tabellare (che riduce il testo a liste di *tokens*), che definirei 'destrutturante' (della linea testuale) o 'acontestuale'. Nel primo caso si lavora di volta in volta sulle singole occorrenze; nel secondo caso si lavora a livello delle forme e l'annotazione della forma viene automaticamente estesa a tutte le occorrenze da essa riassunte. Il primo si impone come fondamentale laddove sia necessario risolvere delle ambiguità, il secondo accelera in modo esponenziale il lavoro di annotazione. Si è visto che lo strumento *Pyrrha* prevede le due opzioni ma in quest'ordine: sistema contestuale per l'immissione e correzione dei dati, sistema tabellare come opzione possibile per la correzione (definita in questo caso correzione a blocchi o *lots*) accanto a quella contestuale.

L'interfaccia ULA propone come sistema di immissione dati e prima correzione un sistema esattamente inverso:

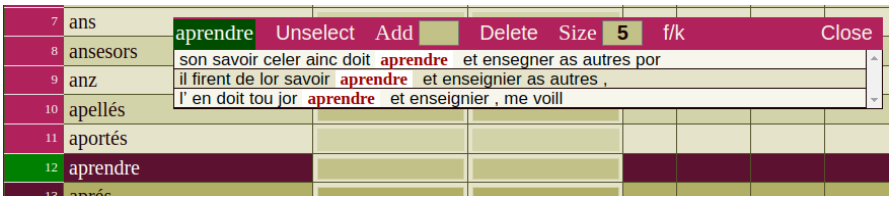


Fig. 7

un'interfaccia tabellare, immediatamente mitigata però dalla possibilità di visualizzare i contesti o concordanze delle occorrenze di ogni singola forma, partendo dall'idea che, astraendo e formalizzando, un testo può in ultima analisi essere definito come un insieme di occorrenze (*tokens*) di forme di lemmi, sequenza questa dove i due termini estremi, occorenza *vs.* lemma, rappresentano i due poli opposti fra dato reale e noto e dato astratto frutto dell'analisi, non noto a priori e da assegnare. Il punto di partenza del sistema ULA è quindi rappresentato da una lista delle forme.

4.2. *Il formario*

Per quel che riguarda la forma, categoria che si pone fra l'occorrenza e il lemma, si possono avere due approcci differenti. Alcuni progetti, in presenza di varianti grafiche, introducono il concetto di forma standardizzata o normalizzata, una categoria quindi intermedia fra la reale esistenza del dato-occorrenza e la totale astrazione del dato-lemma. Traducendo l'introduzione di questa categoria in una rappresentazione informatica, il risultato della tokenizzazione, cioè della creazione della lista delle occorrenze, sarà sottoposto al primo filtro rappresentato dalle forme, ottenendo quindi una lista di tutte le forme esistenti nel testo; questa lista di tutte le forme esistenti sarà quindi sottoposta a un secondo filtro 'normalizzante' che raggrupperà sotto una singola forma standard le molteplici forme con varianti grafiche, ottenendo quindi una lista delle forme normalizzate a cui è infine associato il lemma[28].

[28] Si muoveva in questo senso ad es. il prototipo di lemmatizzatore (solo ed esclusivamente un lemmatizzatore, non è prevista etichettatura morfosintattica) espressamente «conçu pour les bésoins philologiques des textes médiévaux de langue romane» da Martin-Dietrich Glessgen 2003, in collaborazione con Matthias Kopp dell'università di Tübingen: dopo aver prodotto una lista di tutti i *tokens* sotto forma di indice KWIC (*Key Words In Context*), cioè di concordanze, il sistema passa a ridurre gli elementi della lista risolvendo «les cas d'"équivalences graphiques" qui dépendent du système graphématique du langage étudié» (ivi: 68). Per fare questo si deve dapprima impostare una lista di equivalenze grafiche (es. doppia consonante/consonante semplice; doppia vocale/vocale semplice; omofoni *en*/*an*, *y*/*i*, *ngni*/*ngn*/*ng*; variazioni regionali, *ei*/*é*, *w*/*g*, ecc.), un complesso di regole che viene poi applicato dal sistema. Si segnala anche l'interessante prototipo *Graphist*, «un logiciel d'indexation, de lemmatisation et de modernisation automatique pour les textes allant du XVI^{e} siècle à nos jours» (Catach 1996: § 1), sviluppato da Laurent Catach, sotto la guida fra gli altri anche della madre Nina, in seno all'équipe HESO (*Histoire Et Structure de l'Ortographe*) del CNRS: «Du fait de la prise en compte des textes anciens, il faut en réalité opérer une *double* lemmatisation: lemmatisation *flexionnelle* (forme fléchie > forme vedette) et lemmatisation des *graphies anciennes* (forme ancienne > forme moderne). Cela suppose une *analyse morphologique* du mot, permettant d'identifier la flexion (féminins et pluriels pour les substantifs, formes conjuguées pour les verbes) et simultanément une *analyse des variantes graphiques*, celles-ci pouvant porter à la foi sur le lemme lui-même et sur la marque de flexion» (*ibid.*).

Date le caratteristiche linguistiche dei testi che a vario titolo (cioè con gradi diversi di mescidanza) appartengono alla galassia franco-italiana, la lemmatizzazione PRODIGI non applica il concetto di forma standardizzata, tanto più difficile da definire nei casi di forte contaminazione fra le due sfere linguistiche, rispetto ai quali, talvolta persino a livello di lemma, risulta difficile definire quale delle due componenti sia prevalente e quale sia, quindi, la forma standardizzata di riferimento.

Nel caso di introduzione del concetto di forma standard, a prescindere dalle forme presenti nel testo avremo quindi una singola forma, per es. *chevalier* come forma unica del singolare del lemma '*chevalier*'. Nel caso della lemmatizzazione PRODIGI, si considera come chiave unificatrice e univoca fra forme-varianti non già una artificiale forma standard ma la particolare combinazione di lemma/POS/MSD: la combinazione POS/MSD nome comune maschile singolare del lemma '*chevalier*' si esprime quindi non esclusivamente attraverso la forma standard *chevalier*, che raccoglie sotto di sé a un secondo livello le forme possibili *quevalier*, *chivalier*, *cavalier* ecc., ma direttamente attraverso queste forme in modo paritario (le quali, cambiando prospettiva, sono legate tra loro appunto dalla comunanza della combinazione lemma '*chevalier*', POS nome comune, MSD maschile singolare).

In conclusione quindi il testo da annotare immesso in ULA viene tokenizzato e restituito sotto l'aspetto di lista delle forme grafiche. Tuttavia, proprio in previsione di casi in cui si voglia applicare il concetto di forma standardizzata, la colonna delle *Form* è affiancata da una seconda colonna, *FormS* (*Form Standard*), attualmente resa non visibile in quanto nel caso specifico rappresenterebbe un semplice doppione della prima, ma che, se attivata, accoglierebbe le forme standardizzate.

4.3. *Principi di tokenizzazione*

Per quel che riguarda la tokenizzazione[29], come principio generale le parole vengono identificate come tali attraverso la loro delimitazione fra due spazi bianchi. Il tokenizzatore riconosce poi come entità separate le parole graficamente unite da apostrofo (') o punto mediano/*middle dot* (·): in questo caso, oltre a separare le due parole, si conserva il segno diacritico associandolo rispettivamente alla parola di sinistra (per l'apostrofo) e a quella di destra (per il *middle dot*).

Negli ultimi anni non mancano le iniziative volte all'automazione della normalizzazione grafica di lingue non standard «using rule-based, statistical and neural string-transduction models» (Manjavacas [*et alii*] 2019: 1493). Cfr. Pettersson (*et alii*) 2014 e Bollman-Søgaard 2016. Per una discussione sulla nozione di standard in una prospettiva francese cfr. Grübl 2013 e Roger 2017.

[29] Molto diverso lo 'stile' di tokenizzazione di *Pyrrha*, per il quale si veda (con riferimento al francese moderno) Gabay (*et alii*) 2020.

l'autre	*e·l*
l'	e
autre	**·l**

Il tokenizzatore ULA attualmente elimina i segni di punteggiatura, tanto più 'rumorosi' nel caso dell'edizione di un testo medievale in quanto per la maggior parte frutto dell'attività editoriale. Il tokenizzatore conserva però, se applicati, due caratteri che si è scelto di utilizzare nella preparazione del file .TXT (il formato di input dei dati in ULA): il trattino '-' per separare e lemmatizzare separatamente i pronomi clitici posposti o enclitici:

andando-se-ne
andando
-se
-ne

l'*underscore* '_', per unire i componenti di una locuzione o poliremática (o MWEs, *Multi World Expressions*, v. oltre), per es. *por ce que* = *por_ce_que*, o grafie di forme, future canoniche univerbate, che oscillano ancora fra separazione e agglutinazione, per es. le grafie *de denz* = *de_denz* o *le quel* = *le_quel*.

A questo proposito si coglie l'occasione per ricordare l'origine del primo testo lemmatizzato nel contesto di creazione di ULA, il testo cioè del *Roman de Troie Prose 2* come trasmesso dal ms. di Grenoble. Il testo .TXT su cui si è lavorato è quello di un'edizione 'di manoscritto', oggetto anche di una codifica XML-TEI che ne permette la visualizzazione in modalità diplomatica e in modalità interpretativa, realizzata secondo i principi di codifica già applicati al *Roman de Florimont* nell'ambito del precedente progetto DIGIFLOR[30].

Si tratta quindi di un testo su cui sono state operate alcune modernizzazioni – lo scioglimento delle abbreviazioni, la traduzione in chiave ortografica moderna delle lettere ramiste, l'inserimento di apostrofi, accenti e altri segni diacritici – ma che rispetta fondamentalmente la grafia del manoscritto. Si conserva perciò, ritornando all'esempio citato, l'alternanza delle grafie *de denz* e *dedenz*, codificandole però entrambe in unico tag <w> (*word*) (<w>*de denz*</w>, <w>*dedenz*</w>), in quanto si fa differenza fra il concetto di parola grafica e quello di lessìa. Uno degli obiettivi del progetto PRODIGI – che potrà dirsi ampiamente giustificato in relazione ai risultati ottenuti nel caso in cui si potesse in futuro proseguire il lavoro lemmatizzando anche gli altri due testimoni (dei quali alla data odierna è già stata realizzata la trascrizione e, parzialmente, la codifica XML-TEI) –, è infatti, lo si ripete, quello di offrire un *corpus* lemmatizzato di testi manoscritti e non di edizioni critiche, di materiale grafico-linguistico di prima mano o documentario, nei limiti dei pochi trattamenti necessari a una sua corretta fruizione e analisi.

[30] http://digiflorimont.huma-num.fr/ [cons. 24. VII. 2022]. Cfr. Materni 2020 e 2021.

Per questa prima versione di ULA il formato di *input* dei dati da annotare è il formato .TXT o testo puro. E, sia per ridurre il rumore (prodotto ad esempio dalle stesse forme con o senza abbreviazioni), sia per rendere i dati facilmente assorbibili da altri progetti analoghi, si è scelto di lavorare sul testo nella sua veste grafica di edizione interpretativa. Una prima possibile evoluzione dello strumento potrebbe consistere nell'accettazione dei dati testuali di *input* sotto forma di codifica XML-TEI, dando la possibilità di superare così la dimensione stessa di distinzione fra diplomatica e interpretativa, a condizione ovviamente che la codifica del testo rispetti alcune regole. Il progetto informatico portato avanti da chi scrive a partire dall'esperienza DIGIFLOR si basa infatti – considerandola come una scelta fondamentale per pensare in modo differente l'oggetto-edizione testuale, con un cambio radicale di prospettiva nel passaggio dal paradigma cartaceo al paradigma informatizzato – sulla codifica del testo a livello di singoli componenti, che si è scelto di far coincidere con le parole, ciascuna identificata in modo univoco attraverso un xml:id[31]. La trascrizione e la codifica avvengono a livello diplomatico, cioè documentario, e l'edizione interpretativa viene prodotta in modo automatico sulla scorta della semantica della codifica: ad es., nella parola *cheualier* si isola il carattere *u* specificando che si tratta di una lettera ramista con valore consonantico, che andrà quindi sostituita dal carattere *v* nella versione interpretativa. Ciascuna parola, delimitata da un tag <w> (*word*) racchiude dunque in sé tanto la forma grafica diplomatica quanto, a un secondo livello, la sua forma modernizzata. Nel momento in cui l'oggetto della lemmatizzazione fosse rappresentato dalla stringa XML-TEI, il *corpus* delle forme lemmatizzate potrebbe essere visualizzato e consultato, a seconda degli interessi, in entrambe le forme, rendendo ad es. manifesto l'uso delle abbreviazioni. Si tratta di una delle evoluzioni possibili dello strumento su cui si sta lavorando. Il problema in questo caso non è di natura tecnica, quanto nell'elaborazione a monte del modello concettuale: nel caso in cui si utilizzasse lo stesso file XML-TEI sia per l'edizione sia per l'annotazione linguistica, così da creare un oggetto editoriale complesso, si dovrebbe infatti decidere in che modo gestire una parte delle informazioni documentarie, ad es. le correzioni del copista, di interesse per il piano editoriale ma sovrabbondanti, se non fuorvianti, sul piano dell'analisi linguistica.

4.4. *Questioni di* tagset

Si è detto che un altro degli obiettivi perseguiti dallo strumento ULA, accanto a quello di (tentare di) offrire un ambiente di lavoro ergonomico/*user friendly*, è

[31] L'xml:id utilizzato nel progetto DIGIFLOR si compone di: sigla dell'opera/sigla del manoscritto/numero dell'episodio/numero di lassa/numero di verso/numero della parola all'interno del verso, es. Fl_G_1_1_1_1. Quello utilizzato dal progetto PRODIGI si compone invece di: sigla dell'opera/sigla del manoscritto/numero dell'episodio/numero di capitolo/numero di paragrafo/numero di frase/numero della parola all'interno della frase, es. P2_G_1_1_1_1_1.

stato quello di garantire una maggiore accessibilità da parte dell'utente al *tagset* e al dizionario, così da poter operare modifiche/adattamenti resi inevitabilmente necessari dalla varietà del fenomeno linguistico.

La questione della personalizzazione richiede però, prima di procedere, alcune precisazioni. La personalizzazione cui si fa riferimento dovrebbe intendersi, in uno scenario ideale, come possibilità di apportare le necessarie modifiche contestualizzanti ma, e si sottolinea ma, a partire da uno standard, condiviso e riconosciuto come tale. La rigida standardizzazione, infatti, lungi dal rappresentare un'imposizione soffocante, si rivela invece un presupposto imprescindibile per il dialogo fra elementi eterogenei nati in contesti differenti, un metodo di costruzione condiviso che permette di creare un edificio uniforme globale nonostante l'eterogeneità dei materiali costruttivi. Fuor di metafora, parametri di analisi condivisi e standardizzati permettono l'osmosi fra, nel caso specifico, *corpora* linguistici risultato di molteplici progetti. Lo standard chiuso si rivela così presupposto fondamentale per la scienza aperta.

Nel mondo delle edizioni digitali, un primo passo in questo senso, almeno teoricamente, è stato fatto attraverso la TEI (*Text Encoding Initiative*)[32], benché da una certa prospettiva il punto debole della TEI stessa resti proprio il fatto di proporre le sue *Guidelines* come raccomandazioni ma non come un vero e proprio standard, lasciando un ampio margine di autonomia nella realizzazione: ciò che colpisce esplorando il dietro le quinte dei progetti, vale a dire i *files* XML che registrano la codifica dei testi, è per l'appunto la grande varietà di scelte nella creazione dei modelli di codifica che pur fanno riferimento alla stessa tipologia testuale, agli stessi principi filologici, e che utilizzano lo stesso linguaggio di *markup*. In questo senso una delle applicazioni più fruttuose della codifica TEI, fruttuosa in quanto sotto il segno della organicità, è l'iniziativa EpiDoc[33], nata nell'ambito della TEI ma rivolta esclusivamente a documenti antichi (epigrafi e papiri) e quindi calibrata sulle necessità specifiche di questi ultimi, con una positiva rigidità che trasforma le EpiDoc *Guidelines* in autentico standard e non raccomandazioni.

Nell'ambito dell'annotazione linguistica la discussione è ancora aperta. Se è un dato acquisito il fatto che, dopo la lemmatizzazione, l'annotazione morfosintattica o *grammatical tagging* consiste nell'attribuzione di etichette appartenenti rispettivamente alla categoria definita POS e MSD, la nomenclatura dei rispettivi *tagset* è invece regno dell'eterogeneità. La diversità delle forme di etichette e delle regole con cui tali categorie si esprimono rendono difficile l'integrazione fra *corpora* e la loro interrogazione cumulativa. Problema questo al quale, dopo alcuni tentativi di elaborazione di uno standard unico, si sta cercando di rispondere con

[32] https://tei-c.org/ [cons. 24. VII. 2022].
[33] https://sourceforge.net/p/epidoc/wiki/Home/ [cons. 24. VII. 2022].

una differente strategia, quella cioè dell'armonizzazione fra elementi eterogenei attraverso la mediazione delle tecniche del Semantic Web, leggi delle Ontologie, in un'ottica *linked data*:

> In the standardization approach, terminology harmonization is achieved by aggregation and consolidation across all possible user communities within a centralized, monolithic repository. An alternative approach […] was the idea of distributed terminology harmonization by creating links among independent, self-contained domain terminologies and between them and one or multiple 'upper models', especially by means of ontology and Semantic Web technologies. Different user communities formalize and provide their respective terminology in a stand-alone, self-contained ontology, and these ontologies are subsequently linked with each other by means of designated relations […] between identical or near-identical concepts[34].

Tuttavia, prima di armonizzarsi con gli altri occorre ad ogni modo, nel momento in cui si dà avvio a un nuovo progetto, scegliere il proprio *tagset* di riferimento, e questa scelta richiede un'attenta riflessione.

A seguire, dopo aver descritto due iniziative che hanno tentato di creare uno standard, si descriverà il *tagset* adottato per *default* da PRODIGI. Tuttavia, la strutturazione dell'interfaccia ULA è tale che, nel caso si trovasse ergonomica l'interfaccia ma si dissentisse con i criteri di annotazione lessicale utilizzati da PRODIGI, il *tagset* può essere completamente sostituito, continuando a sfruttare l'interfaccia di annotazione: i valori di POS e MSD, ai quali si aggiunge un campo *Lang*, per le annotazioni linguistiche, e il campo *Funct*, illustrato in seguito, sono infatti organizzati in tabelle esterne autonomamente configurabili dall'utente.

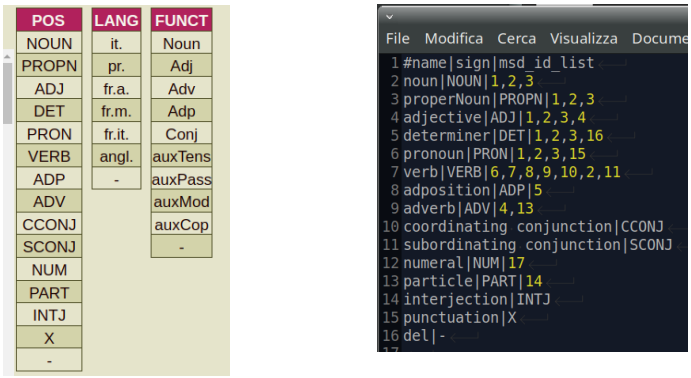


Fig. 8

4.5. *Tentativi di standardizzazione sotto il segno dell'Europa*

Negli anni Novanta sono stati realizzati due importanti progetti, finanziati

[34] Chiarcos (*et alii*) 2020: 5668.

dalla Comunità Europea, orientati proprio verso la costituzione di un *tagset* di annotazione linguistica con l'aspirazione di imporsi come standard. Purtroppo entrambi i progetti non hanno superato la soglia dei primissimi anni 2000 in termini di aggiornamento, benché la documentazione sia ancora perfettamente accessibile. Si tratta dei progetti EAGLES e MULTEXT.

EAGLES (*Expert Advisory Group on Language Engineering Standards*)[35]

Fig. 9

Si è trattato di un'iniziativa della Commissione Europea lanciata nel febbraio 1993 con la sigla LRE-61-100, e coordinata da Antonio Zampolli insieme a Nicoletta Calzolari (Istituto di Linguistica Computazionale, Pisa) e Judith McNaught (CCL); il progetto ha coinvolto anche l'Instituto Cervantes di Madrid, il Deutsches Forschungszentrum für Künstliche Intelligenz (Saarbrücken, Germany), il Center for Sprogteknologi (Copenaghen) e la società Vocalis Ltd. (Great Shelford, UK). L'ultimo aggiornamento della documentazione, fra cui le EAGLES *Guidelines*[36], risale al giugno 1996.

Le parole di presentazione del progetto, nella pagina introduttiva del sito web, ben esprimono sinteticamente e chiaramente tutti i benefici della creazione/adozione di uno standard; purtroppo, queste parole, che datano alla metà degli anni '90, sono ancora attuali nel 2022, nel senso che l'auspicio in esse formulato e rispetto al quale EAGLES rappresentava una prima risposta, rimane fondamentalmente un auspicio dato che uno standard non si è ancora consolidato:

> Recently, many researchers, language engineers and technology planners have become aware of the idea of reusability, and of its crucial role in facilitating the development of practical language technology products that respond to the need of users.
> However, reusability in the language technology field relies, as it does in other fields of technology, on the existence of common practices, guidelines, standards and compatible framework. Standards, whether these be de facto standards or national and international standards, are the necessary key to true reusability.
> With widely known and broadly accepted standards, interchangeability of language technology components becomes feasible, tools can be built to accept input or produce output be mapped into a standard form, products of one type can be compared, if they adhere to relevant standards[37].

[35] http://www.ilc.cnr.it/EAGLES96/home.html [cons. 24. VII. 2022].
[36] http://www.ilc.cnr.it/EAGLES96/browse.html [cons. 24. VII. 2022].
[37] *Introduction to the EAGLES initiative* (http://www.ilc.cnr.it/EAGLES/edintro/edintro.html) [cons. 24. VII. 2022]. Cfr. Calzolari (*et alii*) 1995, Walker (*et alii*) 1995.

Il *tagset* di EAGLES[38] si articola in *Obligatory Attributes*, cioè POS,

1. N [noun]	2. V [verb]	3. AJ [adjective]
4. PD [pronoun/determiner]	5. AT [article]	6. AV [adverb]
7. AP [adposition]	8. C [conjunction]	9. NU [numeral]
10. I [interjection]	11. U [unique/unassigned]	12. R [residual]
13. PU [punctuation]		

Fig. 10

e *Recommended Attributes*, cioè MSD. Si riporta come esempio il sistema di annotazione proposto per il verbo:

(i)	Person:	1. First	2. Second	3. Third	
(ii)	Gender:	1. Masculine	2. Feminine	3. Neuter	
(iii)	Number:	1. Singular	2. Plural		
(iv)	Finiteness:	1. Finite	2. Non-finite		
(v)	Verb form / Mood:	1. Indicative	2. Subjunctive	3. Imperative	4. Conditional
		5. Infinitive	6. Participle	7. Gerund	8. Supine
(vi)	Tense:	1. Present	2. Imperfect	3. Future	4. Past
(vii)	Voice:	1. Active	2. Passive		
(viii)	Status:	1. Main	2. Auxiliary		

Fig. 11

Il progetto EAGLES è stato proseguito negli anni successivi sotto l'etichetta ISLE (*International Standards for Language Engineering*), sempre coordinato da Antonio Zampolli, assistito da Paola Baroni. L'ultimo aggiornamento della documentazione risale in questo caso al marzo 2004[39].

MULTEXT (*Multilingual Text Tools and Corpora*)

MULTEXT

Multilingual Text Tools and Corpora

Fig. 12

Parallelamente al progetto EAGLES e al tempo stesso nel solco di EAGLES, fra il 1994 e il 1996, sotto la direzione dell'Université de Provence (in particolare del *Laboratoire Parole et Language*) in collaborazione con il CNRS, è stato finanziato, ancora dalla Comunità Europea, con la sigla LRE-62-050, il progetto

[38] http://www.ilc.cnr.it/EAGLES96/annotate/annotate.html [cons. 24. VII. 2022].

[39] «The aim of ISLE is to develop HLT standards within an international framework, in the context of the EU-US International Research Cooperation Initiative. […] Its objectives are to support national projects, HLT RTD projects and the language technology industry in general by developping, disseminating and promoting de facto HLT standards and guidelines for language resources, tools and products» (http://www.ilc.cnr.it/EAGLES96/isle/ISLE_Home_Page.htm) [cons. 24. VII. 2022].

MULTEXT[40]. L'iniziativa, coordinata da Nancy Ide e Jean Véronis, è stata applicata a sei lingue – italiano, tedesco, spagnolo, francese, olandese e inglese –, perseguendo sostanzialmente gli stessi obiettivi di EAGLES:

> MULTEXT's general aim is to develop tools for corpus annotation which contribute to the standardization of this kind of work in an academic and an industrial environment. These tools will be provided with resources from six different languages to ensure their validity[41].

Se l'iniziativa MULTEXT, al pari di EAGLES, si è sostanzialmente arenata, producendo documentazione completa all'epoca ma purtroppo non più aggiornata (basti pensare che siamo ancora, a livello di codifica testuale, in epoca SGML), da essa si è però staccato un ramo che risulta invece ancora produttivo: si tratta dell'iniziativa MULTEXT-East, *Multilingual Text Tools and Corpora for Central and Eastern European Languages*[42]; in questo caso, l'ultimo aggiornamento della documentazione risale a giugno del 2021. Attualmente questo standard di annotazione, che si autodefinisce come «EAGLES-based morphosyntactic specification», a sottolineare la continuità fra queste iniziative nate tutte sotto il segno dell'Europa e nello stesso lasso di tempo, si applica a 16 varietà linguistiche dell'Est europeo.

4.6. *L'opzione UD*

Una delle scelte che si sono rese necessarie parallelamente alla progettazione dell'interfaccia ULA è stata quindi quella del *tagset* di analisi. La scelta più difficile è stata decidere se adottare o meno il sistema EAGLES (come si è visto il sistema MULTEXT non è altro che un sistema *EAGLES-based*): la visione della problematica degli standard portata avanti da quel progetto è infatti esattamente quella a cui ho scelto di aderire nel momento in cui ho cominciato a formulare progetti nell'ambito dell'applicazione dell'informatica allo studio dei testi; in secondo luogo, l'autorevolezza di chi ha coordinato il lavori dell'iniziativa EAGLES, Antonio Zampolli, oltre al fatto che si è trattato di un'iniziativa promossa dalla Comunità Europea, rappresentavano sicuramente elementi che sbilanciavano pesantemente la scelta verso questo standard *in potentia* (*de facto* non è mai assurto, purtroppo, a tale status). Ma fra i contro si imponeva il fatto che sostanzialmente il progetto si è arenato verso la fine degli anni '90, continuando indirettamente solo attraverso il ramo collaterale di MULTEXT-Est. Ricordo che il CES (*Corpus Encoding Standard*), «designed to be optimally suited for use

[40] https://www.issco.unige.ch/en/research/projects/MULTEXT.html [cons. 24. VII. 2022]. Cfr. Ide-Véronis 1993 e 1994, Calzolari-Monachini 1995.
[41] Si cita dalla pagina di introduzione del sito.
[42] http://nl.ijs.si/ME/ [cons. 24. VII. 2022].

in language engineering research and applications, in order to serve as a widely accepted set of encoding standards for corpus-based work in natural language processing applications»[43], componente delle EAGLES *Guidelines* e rivolto appunto alla preparazione dei documenti del *corpus*, è ancora formulato in SGML. Nel caso di questo progetto, visto il contesto importante in cui esso è nato, sarebbe auspicabile riprendere in mano l'iniziativa e procedere a un aggiornamento, operazione che non compete certo all'iniziativa unilaterale del singolo ma che necessita di quella analoga rete di collaborazioni internazionali che ha presieduto alla sua concezione e sviluppo. Per il momento, non si è giudicato fruttuoso applicare *tout-court* delle *Guidelines* ferme al 1996 e non condivise attualmente da alcun progetto in corso.

La seconda valutazione ha riguardato l'eventualità dell'adozione o meno del *tagset* CATTEX. Questo *tagset*, elaborato da Céline Guillot, Sophie Prévost, Serge Heiden e Alexei Lavrentiev[44], cioè dalla stessa équipe responsabile del progetto BFM, è stato concepito espressamente per il francese medievale e adottato in vari progetti francesi, oltre a rappresentare il *tagset* scelto da *Pyrrha* (che lo ha esteso anche al francese del XVI-XVII sec.[45]). Anche CATTEX è stato però escluso come prima scelta (in futuro, fra le evoluzioni di ULA si potrà anche mettere in conto di allestire un sistema di equivalenze e conversioni fra *tagsets* differenti, così da garantire interoperabilità dei dati in un sistema di risorse ancora tanto eterogeneo) e questo per due motivi: in primo luogo, e soprattutto, per una certa insoddisfazione rispetto alla nomenclatura del *tagset* stesso e alle regole di attribuzione delle etichette (che corrispondono anche a un diverso sistema di tokenizzazione rispetto a quello proposto da ULA); in secondo luogo, per il suo carattere eccessivamente nazionale e monolinguistico: il *tagset* CATTEX è utilizzato esclusivamente da progetti francofoni, prevalentemente francesi, è stato tagliato su misura sulla lingua francese e si esprime in francese. Lungi dal voler consacrare l'inglese come lingua della comunicazione accademica, è però innegabile che, al di là delle descrizioni analitiche di commento per le quali si auspica invece un doveroso plurilinguismo, a livello di linguaggi formali, ridotti quindi alla dimensione di stringhe di caratteri quasi simbolici quali sono ad es. le etichette morfosintattiche o i *tags* della TEI, l'inglese funziona obiettivamente da lingua universale. CATTEX è stato pensato in Francia, in francese, per testi francesi: questa sua dimensione, oltre ripeto ad alcune scelte rispetto all'annotazione, mal si coniugava con lo spirito con cui è stata creata l'interfaccia ULA, solo accidentalmente testata per la prima volta su un testo della francofonia medievale.

[43] Si cita dalla pagina introduttiva del sito, https://www.cs.vassar.edu/CES/ [cons. 24. VII. 2022].
[44] Prévost (*et alii*) 2009.
[45] Gabay (*et alii*) 2020.

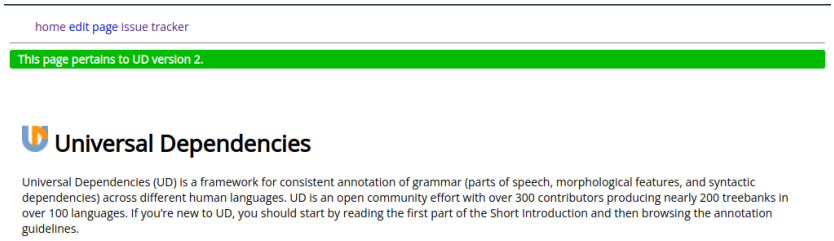


Fig. 13

La scelta è ricaduta dunque su un progetto che condivide lo stesso spirito di EAGLES e MULTEXT ma che è attualmente attivo e costantemente aggiornato. Si tratta del *framework* UD (*Universal Dependencies*)[46]:

> *Universal Dependencies* (UD) is at the same time a framework for cross-linguistically consistent morphosyntactic annotation, an open community effort to create morpho-syntactically annotated corpora for many languages, and a steadily growing collection of such corpora[47].

Per fornire qualche cifra, alla data dell'ultimo aggiornamento, risalente al 15. IX. 2021, il *corpus* UD contava 29.074.543 *tokens* per 183 *treebanks*, 104 varietà linguistiche e 416 ricercatori coinvolti. Il progetto è attualmente coordinato da Joakim Nivre, professore di linguistica computazionale e decano della Facoltà di lingue dell'Università di Uppsala. La scelta di questo *tagset* come punto di riferimento è stata confortata dal fatto che anche il progetto, sempre lionese, TXM (*TeXtoMétrie*)[48], legato alla BFM, a partire dal 2018 propone una tabella di conversione del nativo *tagset* CATTEX in *tagset* UD[49].

Avvicinarsi al *framework* UD significa entrare in un sistema complesso di analisi linguistica che prevede sia il livello lessicale che quello grammaticale (*Universal POS tag*)[50], morfologico (*Universal feature inventory*)[51] e sintattico (*Universal dependency relation*)[52], il tutto codificato come testo puro nel formato CONLL-U[53]. Il tempo a disposizione per la realizzazione del progetto PRODIGI (24 mesi) non ha permesso un'adesione *in toto* al sistema UD: in particolare ci si

[46] Zeman (*et alii*) 2021; l'intero sistema è ampiamente descritto in de Marneffe (*et alii*) 2021.
[47] Dalla pagina di presentazione del progetto, https://universaldependencies.org/ [cons. 24. VII. 2022].
[48] http://txm.ish-lyon.cnrs.fr/ [cons. 24. VII. 2022].
[49] Bertrand (*et alii*) 2019.
[50] https://universaldependencies.org/u/pos/index.html [cons. 24. VII. 2022].
[51] https://universaldependencies.org/u/feat/index.html [cons. 24. VII. 2022].
[52] https://universaldependencies.org/u/dep/index.html [cons. 24. VII. 2022].
[53] https://universaldependencies.org/format.html [cons. 24. VII. 2022]; cfr. anche Bucholz-Marsi 2006.

è visti costretti per il momento a rinunciare all'annotazione sintattica. Ma, si sottolinea, per il momento: l'impostazione dei dati secondo un determinato schema rende infatti perfettamente fattibile il completamento successivo di questo ulteriore livello di annotazione – che si affiancherebbe ai precedenti senza interferire con essi, permettendo una stratificazione di livelli di analisi per aggiunte successive –, e la possibilità quindi di proporre l'integrazione del *corpus Prose 2* fra le *treebanks* UD, accanto all'unico rappresentante dell'antico francese, cioè la *treebank* derivante dalla parziale conversione a UD del corpus SRCM (*Syntactic Reference Corpus of Medieval French*)[54]: dieci testi distribuiti dal IX al XIII sec. per un totale di circa 200.000 *tokens*.

Tuttavia (e il *framework* UD prevede un campo in cui è possibile utilizzare delle specifiche relative alla singola varietà linguistica), il *tagset* UD non è stato adottato in blocco senza alcune piccole modifiche, come si spiegherà.

Il *tagset* UD *Universal POS* si articola in tre classi:

Open class words	Closed class words	Other
ADJ	ADP	PUNCT
ADV	AUX	SYM
INTJ	CCONJ	X
NOUN	DET	
PROPN	NUM	
VERB	PART	
	PRON	
	SCONJ	

Fig. 14

1. *Open class words*: ADJ: adjective; ADV: adverb; INTJ: interjection; NOUN: noun; PROPN: proper noun; VERB: verb
2. *Closed class words*: ADP: adposition; AUX: auxiliary; CCONJ: coordinating conjunction; DET: determiner; NUM: numeral; PART: particle; PRON: pronoun; SCONJ: subordinating conjunction
3. *Other*: PUNCT: punctuation; SYM: symbol; X: other

La tokenizzazione ULA, come si è detto, esclude la punteggiatura e non è quindi stato adottato il POS 'PUNCT'. Il POS 'X' è stato invece riservato ai casi di parole incorrette, etichettate come <sic> nel testo (che, si ricorda, è un'edizione di manoscritto, una trascrizione diplomatica in forma interpretativa nella quale non sono state operate correzioni se non minime in presenza di evidenti e banali errori paleografici, per es. *armane* per *armaire*).

[54] http://srcmf.org; https://github.com/UniversalDependencies/UD_Olf_French_SRCMF/tree/master [cons. 24. VII. 2022].

Un primo punto rispetto al quale ci si è discostati dal *tagset* UD, perché quest'ultimo risultava in disaccordo con l'interpretazione grammaticale adottata per costruire il sistema ULA (disaccordo intellettuale che però, per garantire l'omogeneità dei dati e la ricerca cumulativa all'interno dell'intera *treebank* UD, sarà risolto a livello formale tramite la doppia etichettatura prevista da UD POS/XPOS, dove XPOS rappresenta un valore *language-specific* da far corrispondere a un UD POS[55]) riguarda il POS 'AUX' (*auxiliary*):

> An auxiliary is a function word that accompanies the lexical verb of a verb phrase and expresses grammatical distinctions not carried by the lexical verb, such as person, number, tense, mood, aspect, voice or evidentiality. It is often a verb (which may have non-auxiliary uses as well) but many languages have nonverbal TAME[56] markers and these should also be tagged AUX. The class AUX also include copulas (in the narrow sense of pure linking words for nonverbal predication)[57].

Considerando il contesto specifico delle lingue romanze, a cui si fa comunque riferimento, mi è sembrata una forzatura eccessiva introdurre una categoria grammaticale autonoma 'ausiliare' separandola da quella di 'verbo'. L'ausiliare romanzo è appunto un verbo, con una, anche, sua vita propria e autonoma, con tutti i caratteri morfologici del verbo, contestualmente utilizzato con la funzione di ausiliare. Oltre a separare a livello di POS l'ausiliare dal verbo, il sistema UD mette anche a disposizione la *feature VerbType* con i valori: *Aux* (*auxiliary verb*), *Cop* (*copula verb*), *Mod* (*modal verb*)[58] (*feature* che si somma alla scelta precedente e non la sostituisce come alternativa, in quanto il principio di analisi resta l'indipendenza dell'ausiliare dalle altre categorie). Ma, ripeto, mi sembra che lo schema mal si adatti alla realtà dell'ausiliare romanzo, il quale non è un tipo di verbo ma un verbo autonomo utilizzato contestualmente in *funzione* di ausiliare; in altre parole, la dimensione di ausiliare, copula ecc., non mi sembra propriamente una *feature*, vale a dire una caratteristica morfologica. Il valore linguistico espresso dal POS 'AUX' è stato quindi sostituito dal valore *aux* di un nuovo campo proposto dal *tagset* PRODIGI accanto a quelli canonici POS e MSD, vale a dire il campo *Funct*, con le specifiche: *auxTens* (*Tense auxiliary*), *auxPass* (*Passive auxiliary*), *auxMod* (*Modal auxiliary*), *auxCop* (*Verbal copulas*). Il campo *Funct*, come si vedrà successivamente, viene utilizzato anche per risolvere la questione della discrepanza fra forma e uso, che non è evidenziata nella sua ambiguità da UD, dove si indica

[55] «[…] the XPOS optionally contains a language-specific part-of-speech tag, normally from a traditional, more fine-grained tagset. If the XPOS field is used, the treebank-specific documentation should define a mapping from XPOS to UPOS values» (https://universaldependencies.org/format.html [cons. 24. VII. 2022]).

[56] TAME: *Tense, Aspect, Modality, Evidentiality markers*.

[57] https://universaldependencies.org/u/pos/AUX_.html [cons. 24. VII. 2022].

[58] https://universaldependencies.org/u/feat/all.html#al-u-feat/VerbForm [cons. 24. VII. 2022].

invece di optare esclusivamente per una delle due categorie a livello di POS: «Participles are words forms that may share properties and usage of any of adjectives, nouns, and verbs. Depending on the language and context, they may be classified as any of ADJ, NOUN or VERB»[59].

Un altro elemento rispetto al quale ci si discosta dal *tagset* UD, perché la nomenclatura mi sembra eccessivamente verbosa in relazione alle caratteristiche delle lingue romanze, è il trattamento degli articoli. L'articolo appartiene alla categoria dei Determinanti (DET, *Art*), insieme agli aggettivi possessivi, interrogativi, relativi, dimostrativi, negativi e indefiniti. Una ripartizione questa ormai comunemente accettata:

> I determinanti sono parole grammaticali con la funzione di unirsi al nome per formare un SN[60]; morfologicamente, quando non sono invariabili, sono di norma caratterizzati dall'accordo in genere e numero con il nome (come gli aggettivi); da un punto di vista semantico trasmettono informazioni sulla identificabilità del SN, generica (gli articoli definiti) o specifica nello spazio, nel tempo, nel discorso (i dimostrativi), sulla sua quantificazione, generica (individuandone solo l'esistenza, cioè l'articolo indefinito) o specifica (i quantificatori suddivisibili nella sottoclassi dei numerali cardinali e degli indefiniti), o sulla sua individuazione (deittica o anaforica)[61].

A sua volta, nello schema UD, esso è interessato dalla *feature* della *definitiness or state*, con i valori *definite/indefinite*, valori che nelle lingue romanze però si applicherebbero solo agli articoli, dal momento che per aggettivi e pronomi indefiniti è proposta una categoria a parte. Per snellire l'annotazione sono stati dunque introdotti direttamente i due valori *ArtDef* e *ArtInd*, più consoni alla comune pratica di analisi linguistica in ambito romanzo. Analogamente è stato modificato il sistema di annotazione dei determinanti (ex aggettivi) possessivi: in UD, il POS 'DET' va accompagnato dalla *feat*. *Prs* (*personal or possessive personal pronoun or determiner*) e dall'opzione booleana *Poss* (*possessive*) con valore *Yes* nel caso si tratti di un possessivo[62]: mi è sembrato che in questo caso il sistema forzasse troppo le usuali convenzioni di analisi in ambito romanzo, e ho utilizzato il valore *Poss* direttamente come *feature* tanto dei determinanti che dei pronomi, distinguendo questi ultimi dai pronomi personali.

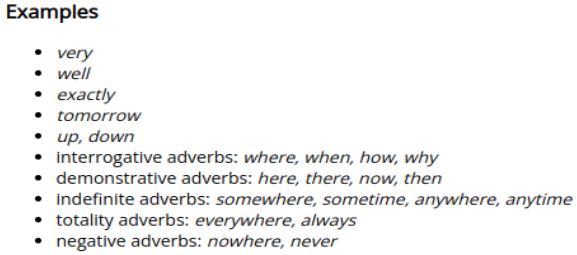
Sul modello di altre *treebanks* confluite in UD e che presentano anch'esse alcune piccole modifiche[63], per la categoria ADP (*adposition*), che riunisce senza

[59] https://universaldependencies.org/u/pos/ADJ.html [cons. 24. VII. 2022].
[60] SN: sintagma nominale.
[61] Dotto 2012: 359.
[62] «While many tagset would have "possessive" has one of the various pronoun types, this feature is intentionally separate from PronType, as it is orthogonal to pronominal types. Several of the pronominal types can be optionally possessive, and adjectives can too» (https://universaldependencies.org/u/feat/all.html#al-u-feat/Poss) [cons. 24. VII. 2022].
[63] La lista di tutti gli attributi e valori presenti nella totalità delle *treebanks* confluite in UD è

distinzioni preposizioni e posposizioni, si è sentita la necessità di introdurre la *feat*. *Prep*[64], nella forma dei valori *PrepS* e *PrepArt* per *simple preposition* e *articulated preposition*.

Nell'analisi verbale, è stata introdotta: la diatesi riflessiva[65] (*Rfl*) accanto a quelle attiva e passiva, aspetto rappresentato dalla *feat*. *Voice*; e la *feat*. *Property* con i valori *Intr* (*intransitive*) e *Trans* (*transitive*), aspetto questo non contemplato da nessuna delle *treebanks* presenti in UD.

Complessa, e non risolta ancora soddisfacentemente, l'annotazione degli avverbi. Le *Guidelines* UD propongono di applicare agli avverbi i valori della *feat*. *PronType*, comune anche a pronomi e determinanti. Seguendo gli esempi proposti, e tratti dall'inglese, avremo dunque


Examples

- *very*
- *well*
- *exactly*
- *tomorrow*
- *up, down*
- interrogative adverbs: *where, when, how, why*
- demonstrative adverbs: *here, there, now, then*
- indefinite adverbs: *somewhere, sometime, anywhere, anytime*
- totality adverbs: *everywhere, always*
- negative adverbs: *nowhere, never*

Fig. 15

Come si vede però l'analisi presenta una dissimmetria: i primi cinque esempi risultano 'liberi', privi di una categoria. Se si consulta la lista della totalità delle *features* presenti nelle varie *treebanks*, si vede come da alcuni progetti siano stati introdotti valori che fanno riferimento alla classica distinzione 'tematica' degli avverbi in avverbi di *manner* (*Man*)[66], *time* (*Tim*)[65], *place* (*Loc*)[68], *degree* (*Deg*)[69], a cui a questo punto si propone di aggiungere anche *frequency* (*Freq*). Trattandosi di un soggetto assai complesso, al quale sostanzialmente tutti i *tagsets* consultati hanno dato risposte differenti, si è scelto per il momento di conservare entrambi i sistemi di classificazione proposti, rinviando a un approfondimento futuro un affinamento di questa parte del *tagset*, al pari di quel che concerne le congiunzioni,

consultabile alla pagina https://universaldependencies.org/ext-feat-index.html [cons. 24. VII. 2022].

[64] Come nelle *treebanks* di «Afrikaans, Arabic, Armenian, Czech, Danish, Estonian, Finnish, Galician, German, Kurmanji, Latin, Lithuanian, Low Saxon, Polish, Romanian, Skolt Sami, Slovak, Western Armenian» (*ibid.*).

[65] Come nelle *treebanks* di «Turkish, Turkish German» (*ibid.*).

[66] «Erzya, Komi Permyak, Komi Zyrian, Moksha» (*ibid.*).

[67] «Apurina, Catalan, Classical Chinese, Erzya, Komi Zyrian, Latin, Moksha, Skolt Sami, Spanish» (*ibid.*).

[68] «Erzya, Komi Permyak, Komi Zyrian, Latin, Moksha» (*ibid.*).

[69] «Bhojpuri, Classical Chinese, Erzya, Hindi, Komi Permyak, Komi Zyrian, Moksha, Urdu» (*ibid.*).

per ora sommariamente classificate come coordinanti o subordinanti senza ulteriori specificazioni.

L'altro elemento rispetto al quale ci si discosta dallo schema UD è il trattamento delle polirematiche o MWEs (*Multi Words Expressions*), di cui si discuterà più nel dettaglio nel paragrafo successivo. Le MWEs, di tutte le tipologie, sono contemplate in UD ma la loro segnalazione è riservata alla fase di annotazione sintattica delle *dependencies*. I singoli elementi di una MWEs sono tokenizzati, lemmatizzati e annotati ciascuno in modo indipendente e legati fra loro successivamente dalla relazione di dipendenza *fixed* (con l'eccezione dei casi di diatesi passiva e di tempo verbale composto dove la relazione di dipendenza è di tipo *aux*).

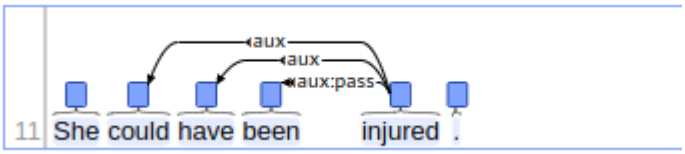


Fig. 16

La lemmatizzazione PRODIGI opta invece per una lessicalizzazione delle MWEs in quanto tali, nel loro complesso e non destrutturate nei singoli elementi, processo auspicabile per tutte le categorie di polirematiche ma di fatto, come si vedrà, applicata solo ad alcune, per ragioni più pragmatiche che teoriche.

5. *Criticità del linguaggio*

Nel costruire il sistema ULA ci si è scontrati inevitabilmente con tre problematiche linguistiche, di cui due, la disambiguazione delle forme omografe e la discrepanza fra forma e uso, fanno riferimento a un'universale problematicità intrinseca al linguaggio stesso (che convive, in proporzione maggiore o minore a seconda dei casi, con l'omofonia, la quale però non interessa la lemmatizzazione in quanto appartenente al livello della fonazione); mentre la terza, il trattamento delle polirematiche, attiene invece alla sfera metodologica, dal momento che la rilevazione della loro esistenza e le modalità della loro annotazione dipendono da una scelta di analisi.

5.1. *Disambiguazione*

Dal punto di vista lessicografico, per quel che riguarda i criteri di distinzione degli omografi, si è cercato di evitare il ricorso a disambiguatori di tipo semantico[70], appoggiandosi esclusivamente sull'annotazione morfosintattica. Il problema

[70] Si è cercato cioè di mantenersi il più possibile a quel livello di formalizzazione dell'informazione

dell'esistenza però di casi limite in cui si ha coincidenza esatta anche di forma, grafia del lemma, POS e MSD, ad es. i due sostantivi con grafia *pesca* nell'esempio che segue:

pesca : pesca (=pèsca), nome com., femm. sing.
pesca : pesca (=pésca), nome com., femm. sing.
pesca : pescare, verbo, ind. pres. 3° sing.
pesca! : pescare, verbo, imp. pres. 2° sing.

è stato risolto aggiungendo ai campi *Form* e *Lemma* il campo *Etimo*. Si tratta di una soluzione parzialmente simile a quella adottata nel DiVo (*Dizionario dei Volgarizzamenti*)[71], parzialmente perché, nel DiVo, essa non è sistematica ma dipende dalla situazione: gli omografi non omofoni sono distinti dall'accento e, per quel che riguarda l'etimo (che può anche essere sostituito dal significato, il tutto riportato in un campo chiamato disambiguatore)[72], esso è normalmente associato esclusivamente al lemma meno attestato e solo in caso di parità di uso a entrambi i lemmi[73]. Ragioni di omogeneità di trattamento da una parte e di sicurezza dell'informazione dall'altra (il carattere accentato può rappresentare un punto debole per facile svista dell'operatore) hanno fatto optare per un uso sistematico dell'etimo come disambiguatore in casi estremi, quelli cioè di omografia del lemma, a prescindere dalla resa fonetica.

L'introduzione dell'etimologia (nel caso specifico di PRODIGI compilata con riferimento al FEW, *Französisches Etymologisches Wörterbuch*), rappresenta sicuramente, al di là della questione dell'omografia, un ulteriore arricchimento del *corpus* ma anche, chiaramente, un aggravio che pesa sul lavoro del compilatore. Tuttavia si tratta di uno di quei casi in cui maggiormente si apprezza la possibilità offerta dal formato digitale di compilare i dati per step successivi (oltreché di correggere facilmente il già fatto): l'etimologia può essere infatti aggiunta sistematicamente in una fase successiva a quella prioritaria di assegnazione del lemma e dei POS/MSD, limitandosi nella prima fase ad inserirla solo laddove essa risulti

che ne rende possibile una gestione completamente automatica (avendo sempre come prospettiva la definizione stessa di 'informatica', termine che altro non significa se non scienza dell'informazione automatizzata), permettendo così di distinguere nettamente e senza commistioni fra un dizionario 'umano' e un dizionario elettronico: «La rédaction d'un dictionnaire éditorial s'adresse à un lecteur humain: elle fait appel à son intelligence et à son intuition pour rétablir l'information implicite par analogie, par application de règles générales, qu'il est inutile de formuler précisement, par suggestion à partir d'exemples, ou tout simplement à partir de sa connaissance préalable de la langue. Au contraire, le contenu d'un dictionnaire éléctronique est destiné à l'exploitation informatique directe, et n'est constitué que d'infomations codées et explicites; les exemples éventuels sont à usage interne et ont un statut comparable à celui des commentaires dans le code source des programmes» (Laporte 2000: 36).

[71] http://divoweb.ovi.cnr.it/ [cons. 24. VII. 2022].

[72] Per la lemmatizzazione manuale del DiVo, realizzata all'interno del software di gestione testuale GATTO, cfr. Iorio-Fili 2007.

[73] Dotto 2012: 346.

necessaria per la gestione degli omografi (o per qualsiasi altro motivo). A questo proposito, più in generale, risulta utile concepire la struttura di archiviazione dei dati nel modo più 'frammentario' possibile, così da poter gestire ogni campo autonomamente e in tempi diversi.

5.2. *Le polirematiche*

La seconda criticità affrontata riguarda il trattamento delle polirematiche. A tal proposito è innanzitutto necessario decidere a monte se includere o meno tale nozione fra le categorie di analisi[74]. In secondo luogo è necessario tener presente che esistono più tipologie di polirematiche, alcune (quelle nominali, aggettivali e verbali) afferenti anche alla sfera semantica, altre (quelle congiuntive, avverbiali e preposizionali) afferenti a una sfera sintattica, funzionale. Includere la categoria delle polirematiche sicuramente aumenta per certi versi il carico di lavoro, in questo caso a livello di preparazione del testo per l'importazione dal momento che le parole grafiche separate (v. sopra principi di tokenizzazione) devono essere associate dal sistema così da formare un *token* unico corrispondente alla polirematica[75]. Tuttavia, la loro esistenza linguistica è un dato di fatto altrettanto innegabile. Valutando pro e contro[76], benefici e aggravi, nel progetto di annotazione lessicale PRODIGI si è scelto di prendere in considerazione le sole locuzioni del secondo tipo, ovvero congiuntive/avverbiali/preposizionali, quelle cioè che non coinvolgono anche la sfera semantica bensì soltanto quella grammaticale, rimandando la segnalazione dell'esistenza delle prime a una sede-dizionario: ricadiamo infatti in questo caso prevalentemente nell'ambito degli usi e dei significati, ambito che da un certo punto di vista, ma è ovviamente opinabile, si può anche considerare estraneo a quello dell'annotazione linguistica *stricto sensu*. Altro discorso è quello che riguarda invece le locuzioni o polirematiche prese in considerazione, in cui, di fatto, i singoli elementi risultano per l'appunto – come l'espressione 'per l'appunto' appena utilizzata – componenti di un insieme inscindibile ed estremamen- te ricorrente. Lemmatizzare a livello dei componenti

[74] Una delle trattazioni più dettagliate è quella di Giuliani 2008, relativa al TLIO (*Tesoro della Lingua Italiana delle Origini*). Il TLIO propone un repertorio delle polirematiche riconosciute (ttp://ovipc44.ovi.cnr.it/Tliopoli/ [cons. 24. VII. 2022]), assai utile come punto di riferimento soprattutto per prendere decisioni circa il lemma sotto il quale collocare la polirematica stessa.

[75] Un'evoluzione possibile dell'infrastruttura ULA sarà il riconoscimento automatico delle sequenze polirematiche, sollevando progressivamente l'operatore, con l'avanzare del lavoro, dall'onere della preparazione del file di *input*.

[76] Raramente le polirematiche sono introdotte come categoria fra i criteri di lemmatizzazione. Fra i progetti che le hanno incluse sistematicamente si segnala il progetto COLFIS (*Corpus e Lessico di Frequenza dell'Italiano Scritto*) (http://www.ge.ilc.cnr.it/corpus_lem.php [cons. 24. VII. 2022]), nel quale esse vengono definite 'forme sintagmatiche' (Bambini-Trevisan 2012).

locuzioni diffusissime come, nel caso dell'antico francese, *por ce que* o *tout ce*, può inoltre significare aumentare notevolmente il rumore di fondo che circonda inevitabilmente i risultati della lemmatizzazione, falsando il quadro rispetto all'uso proprio e autonomo dei singoli elementi (per es. il pronome dimostrativo *ce*).

Due casi di polirematica hanno imposto una rinuncia un po' sofferta, e la rinuncia è dovuta a motivazioni differenti nelle due situazioni. Il primo caso è quello dei verbi polirematici (*andar via*, *uscire fuori*), assai più frequenti delle polirematiche aggettivali e nominali; il secondo, quello che riguarda i tempi composti. Le due categorie ovviamente si intersecano sul terreno comune della forma verbale. Lemmatizzare come polirematiche queste forme verbali avrebbe però comportato una serie di problemi: una dissimmetria di analisi, nel primo caso, nel momento in cui si fosse scelto di segnalare i verbi polirematici a discapito delle altre polirematiche semantiche; un problema tecnico, sia nel primo che nel secondo caso, perché, a differenza delle componenti delle polirematiche che si è deciso di analizzare, le singole componenti di quelle qui citate possono risultare separate (*è spesso uscito fuori*, *è stato molto amato*) (dislocazione delle componenti che può ulteriormente accentuarsi nel dettato poetico), rendendo assai complesso, se non eventualmente macchinoso, il sistema di gestione della tokenizzazione automatica, che segue inevitabilmente la linearità della linea testuale.

In ULA la polirematica viene dunque tokenizzata in unico *token* preparando il file .TXT tramite l'uso dell'*underscore* per legare i singoli elementi: *aprés_ce_que*. Il problema che si pone successivamente, a livello di analisi, è duplice: a quale lemma associare la locuzione? quale POS attribuire?

Per quel che riguarda la prima questione, due sono le soluzioni possibili: l'associazione della locuzione con un lemma rappresentato dalla locuzione stessa, *aprés ce que*, trasformando, in altre parole, la polirematica in lemma; oppure, come di fatto si ritrova abitualmente in qualsiasi dizionario, associandola con una voce 'significativa' fra quelle che la compongono. Quale sia questa voce è tuttavia il risultato di una convenzione da stabilire a priori, una vera e propria regola del gioco. Si riportano come esempio le indicazioni fornite dalla *Norme per la redazione del TLIO*:

> Tutte le polirematiche di vario tipo […] si schedano in ordine di priorità nella voce
> - del primo sostantivo (*cavallo di fiume* s.v. *cavallo*);
> - in mancanza, in quella del primo aggettivo (*tenere caldo* s.v. *caldo*);
> - in mancanza in quella del primo verbo (*andare dietro* s.v. *andare*);
> - in mancanza in quella del primo avverbio (*per annanzi* s.v. *annanzi*)[77].

Nel caso specifico di PRODIGI, si fa riferimento ai criteri lessicografici del DMF (peraltro in linea con la citazione precedente dei criteri del TLIO): la polirematica *aprés ce que* sarà dunque lemmatizzata sotto la voce *aprés*.

[77] Beltrami (*et alii*) 2013: 81.

La seconda questione, conseguente, riguarda l'attribuzione del POS. Proseguendo con l'esempio citato, la poliremática *aprés ce que* è una locuzione congiuntiva creata a partire dalla preposizione *aprés*. Essendo la poliremática associata al lemma dell'elemento più significativo, in questo caso la preposizione *aprés*, essa deve per forza ereditarne anche il POS, pena la creazione di una chimera grammaticale e falso omografo come risulterebbe essere un lemma *aprés* con POS 'SCONJ' come entrata della forma *aprés ce que*. Il sistema PRODIGI procede allora: 1. segnalando dapprima, attraverso la *feat*. '*MWEs*', che la forma, in questo caso *aprés*, a cui è attribuito il suo POS d'origine, è utilizzata all'interno di una locuzione; 2. indicando successivamente la funzione sintattica della locuzione attraverso un valore del campo *Funct.* La locuzione *aprés ce que* sarà dunque associata al lemma '*aprés*', di cui erediterà POS 'ADP' e MSD '*PrepS*/*MWEs*', e sarà poi determinata dal valore *Funct* '*sConj*'.

5.2.1. I tempi composti (*ho amato*) e la diatesi passiva (*sono amato*/*sono stato amato*)

Come si è detto, in questi casi per scelta pragmatica non si è adottato il concetto di poliremática ma si è lasciato che la tokenizzazione rompesse i legami fra le componenti. Al tempo stesso si è voluto però cercare di mantenere l'informazione circa il legame esistente, così da poter distinguere tra l'uso proprio dei verbi *essere* e *avere* e il loro uso come ausiliari (insieme ai verbi modali e alla copula). Gli ausiliari di tempo e di diatesi sono quindi etichettati come verbi, VRB, con gli MSD del caso, e associati con la funzione *auxTens* per i tempi composti e *auxPass* per la diatesi passiva. Nel caso di tempo composto di diatesi passiva (*sono stato amato*), per convenzione si seguono le linee guida ('linee guida', una poliremática nominale!) dettate dal *framework* UD, punto di riferimento per la nomenclatura del *tagset* ULA: il primo ausiliare, *sono*, è considerato portatore della funzione temporale; il secondo, *stato*, della funzione di marca della diatesi passiva. Parallelamente (e questo non è contemplato da UD), dal lato del participio passato si è inserito, fra i *tags* MSD, il valore *MWEs*, a cui corrisponde una funzione *compTense* (*compound Tense*, sottinteso 'componente di tempo composto') e *compPass* per il passivo, così da distinguere eventuali usi assoluti del participio.

5.3. *Forma grammaticale e uso sintattico*

Il campo *Funct* e il meccanismo appena visto di distribuzione dell'informazione sono stati utilizzati per risolvere un ultimo problema, intrinseco ancora una volta alle ambivalenze proprie della lingua, cioè l'uso di una forma, ad es. il participio presente, con un valore grammaticale differente da quello di origine (si pensi ad es. all'uso del participio presente in funzione di aggettivo: le stelle brillanti in cielo *vs.* un uomo brillante; ma anche di sostantivo: docente, cantante), non di rado

con notevoli difficoltà nel definire quale delle due categorie, quella d'origine o quella d'uso, sia realmente preponderante[78]. In questi casi, si annoterà l'occorrenza con il suo POS di origine, adottando un criterio strettamente morfologico, e si segnalerà l'uso secondario nel campo *Funct*. La terminologia fa riferimento alle *Guidelines* EAGLES e alle specifiche per *Multiple tagging practices. Form-function and lemmatisation*:

> Sometimes the need is felt to assign two different tags to the same word: one representing the formal category, and the other the functional category, e.g.: A word with the form of a past participle but the function of an adjective; A word with the form of an adjective but the function of an adverb. In principle, it can be argued that two tags should be assigned to each of these word types, and should be distinctly encoded. In practice, tagging schemes up to the present have tended to give priority of one criterion over another (i.e. giving priority to function over form or vice versa). The annotation scheme for a given tagged corpus should clearly state the use of such criteria[79].

Uno degli impegni della lemmatizzazione PRODIGI è appunto nella sistematicità dell'annotazione riguardante questo aspetto. Farà poi parte delle 'convenzioni di lemmatizzazione' stabilite a priori in fase di definizione dei parametri di analisi il fatto di lemmatizzare in modo diverso, direttamente sotto la categoria grammaticale acquisita, termini cristallizzati nell'uso come *studente*.

Anche in questo caso si adotta una soluzione parzialmente analoga a quella del DiVo, nel quale: a prescindere dall'uso (come in ULA) si lemmatizzano i participi presenti come tali, come una categoria a parte (al contrario di ULA), attribuendo tuttavia come lemma l'infinito; nel già citato campo 'disambiguatore' si inseriscono poi le categorie *v.* o *agg.* per «separare le occorrenze con valore sicuramente verbale da quelle con valore sicuramente aggettivale. Sarà invece obbligatorio lasciare vuoto il disambiguatore per tutte le occorrenze ambigue, anche quelle debolmente ambigue»[80]. Il ruolo svolto dal campo disambiguatore, che quindi in DiVo finisce per accogliere informazioni eterogenee (v. sopra per gli omografi) viene invece in ULA svolto dal campo *Funct*. Inoltre il sistema ULA estende la metodologia a tutti gli altri casi di trasposizione: del verbo in sostantivo, dell'aggettivo in sostantivo, dell'aggettivo in avverbio, adottando come criterio generale e uniforme la distribuzione delle informazioni fra POS e *Funct*. Si tratta sostanzialmente, al netto della terminologia, della soluzione adottata dal *tagset* CATTEX, in cui si affiancano due etichette, l'*étiquette morphologique* (M) e l'*étiquette morphosyntaxique* (Ms): un infinito sostantivato sarà dunque annotato tramite la combinazione M <VERinf>/ Ms <NOMcom>.

[78] A tal proposito si consideri la lunga discussione sui test per distinguere fra uso aggettivale e uso participiale esposta nei *Criteri di lemmatizzazione* del COLFIS (http://www.ge.ilc.cnr.it/corpus_lem.php [cons. 24. VII. 2022]).
[79] http://www.ilc.cnr.it/EAGLES96/annotate/node24.html [cons. 24. VII. 2022].
[80] Dotto 2012: 351.

Quello che segue è dunque lo schema delle combinazioni POS-MSD utilizzate dal progetto PRODIGI, seguito dalla classificazione delle *features* MSD impiegate. Nello schema vengono sottolineate le etichette introdotte o modificate dal progetto.

1	NOUN	Masc/Fem/Neut Sing/Plur Nom/Acc MWEs
2	PROPN	Masc/Fem/Neut Sing/Plur Nom/Acc
3	ADJ	Masc/Fem/Neut Sing/Plur Nom/Acc Pos/Comp/Sup/Abs MWEs
4	DET	ArtDef/ArtInd/Poss/Rel/Dem/Neg/Ind/Int/Exc Masc/Fem/Neut Sing/Plur Nom/Acc MWEs
5	PRON	Prs/Poss/Rel/Dem/Neg/Ind/Int Masc/Fem/Neut Sing/Plur Nom/Acc MWEs
6	VERB	Fin/Ind Inf/Part/Ger/Ind/Imp/Cnd/Sub Past/Pres/Fut/Imp Act/Pass/Rfl Intr/Trans 1/2/3 Sing/Plur Masc/Fem/Neut MWEs
7	ADP	PrepS/PrepArt MWEs
8	ADV	Dem/Ind/Neg/Int/Tot Man/Loc/Tim/Deg/Freq Pos/Comp/Sup/Abs MWEs
9	CCONJ	MWEs
10	SCONJ	MWEs
11	NUM	Card/Ord/Mult
12	PART	Pos/Neg/Int/Exc
13	X	

a. Gender: *Masc* (masculine), *Fem* (feminine), *Neut* (neuter)
b. Number: *Sing* (singular), *Plur* (plural)
c. Person: *1* (first), *2* (second), *3* (third)
d. Case: *Nom* (nominative/direct), *Acc* (accusative/oblique)[81]
e. Degree: *Pos* (positive), *Cmp* (comparative), *Sup* (superlative), *Abs* (absolute superlative)
f. PronType[82]: *ArtDef* (article definite), *ArtInd* (article indefinite), *Prs* (personal), *Reflex* (reflexive), *Poss* (possessive), *Rel* (relative), *Dem* (demonstrative), *Neg* (negative), *Ind* (indefinite), *Int* (interrogative), *Exc* (exclamative), *Tot* (total/collective); [solo avverbi] *Man* (manner), *Loc* (place), *Tim* (time), *Deg* (degree), *Freq* (frequency)
g. VerbForm: *Fin* (finite verb), *Inf* (infinitive), *Part* (participle, verbal adjective), *Ger* (gerund)
h. Mood: *Ind* (indicative or realist), *Imp* (imperative), *Cnd* (conditional), *Sub* (subjunctive/conjunctive)
i. Tense: *Past* (past tense/preterite/aorist), *Pres* (present/non-past tense), *Fut* (future tense), *Imp* (imperfect)
j. Voice: *Act* (active or actor-focus voice), *Pass* (passive or patient-focus voice), *Rfl* (reflexive)
k. Property: *Intr* (intransitive), *Trans* (transitive)
l. NumType: *Card* (cardinal number), *Ord* (ordinal number), *Mult* (multiplicative number)
m. Polarity: *Pos* (positive, affermative), *Neg* (negative)

Funct(ion): *Noun*, *Pron*, *Adj*, *Adv*, *Adp*, *Cconj*, *Sconj*, *auxTens*, *auxPass*, *auxMod*, *auxCop*, *compTens*, *compPass*, *compPassTens*

6. *Lo spazio di lavoro ULA*

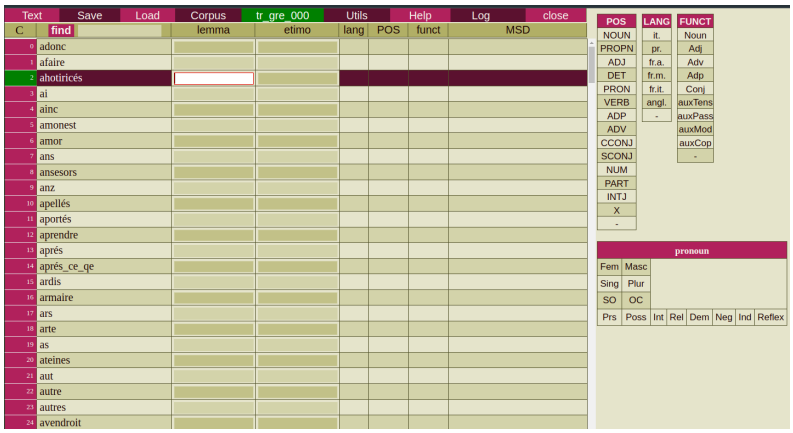


Fig. 17

[81] «If the language has only two cases, which are called "direct" and "oblique", the direct case will be marked Nom. […] the oblique case will be marked Acc» (https://universaldependencies.org/u/feat/all.html#al-u-feat/Case) [cons. 24. VII. 2022].
[82] «This feature [*pronominal type*] typically applies to pronouns, pronominal adjectives (determiners), pronominal numerals (quantifiers) and pronominal adverbs» (https://universaldependencies.org/u/feat/PronType.html) [cons. 24. VII. 2022].

Si è detto che, per velocizzare il lavoro, l'interfaccia ULA ha cercato di comporre i vantaggi del sistema tabellare con quelli del sistema contestuale. La prima fase di lavoro consiste dunque nella lemmatizzazione e *tagging* delle Forme o *Type* (di contro all'Occorrenza o *Token*). La colonna delle forme rappresenta quindi il baricentro dello spazio di lavoro ULA.

Il numero a sinistra di ciascuna forma permette di aprire una finestra delle concordanze con tutte le occorrenze di quella specifica forma grafica, consentendo, tramite la contestualizzazione dell'occorrenza, di procedere a quella che costituisce la parte principale del lavoro di correzione, vale a dire la disambiguazione degli omografi.

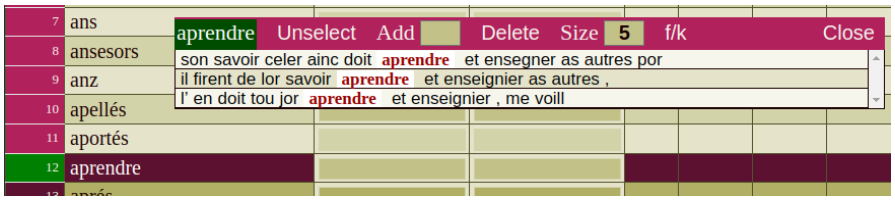


Fig. 18

A destra della colonna delle forme si apre lo spazio di lavoro vero e proprio. A sua volta questo spazio si divide in due sottosezioni (corrispondenti anche a due sottosezioni concettuali e, in realtà, a due differenti operazioni: lemmatizzazione da una parte, annotazione linguistica dall'altra): le colonne *Lemma* e *Etimo*, a *input* libero; e le colonne POS, MSD, *Funct*, *Lang*, che, per ridurre al minimo l'errore umano[83], vengono compilate in modo automatico selezionando i valori nelle tabelle (modificabili) poste a destra dell'area. Per il salvataggio di una voce è sufficiente, ma necessario, riempire i campi *Lemma* e POS.

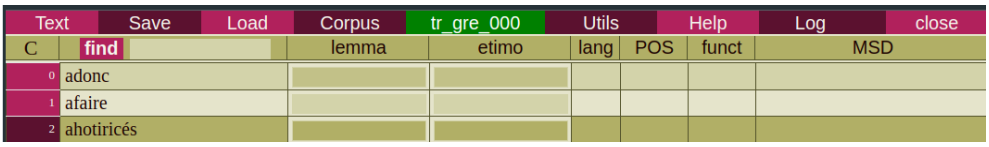


Fig. 19

La prima opzione del menu, *Text*, permette di affinare la fase di correzione, di passare cioè alla modalità di visualizzazione del testo in quanto tale, nella sua forma lineare originaria e non destrutturata come lista di forme in ordine alfabetico.

[83] Ovviamente, in un sistema di immissione dati l'errore non è mai eliminabile del tutto ma solo riducibile: in questo caso si elimina l'errore di digitazione, ma certamente si può ingenerare un nuovo errore, quello di selezione di un valore sbagliato. Questo sistema a selezione però, valutando pro e contro, risultava assai più rapido, trattandosi di una lemmatizzazione che, nelle sue prime fasi, è completamente manuale.

Scorrendo ciascuna riga, si vedrà apparire una tabella con le informazioni di lemmatizzazione e, nel caso di errori, sarà possibile ritornare allo spazio di lavoro per effettuare la correzione.

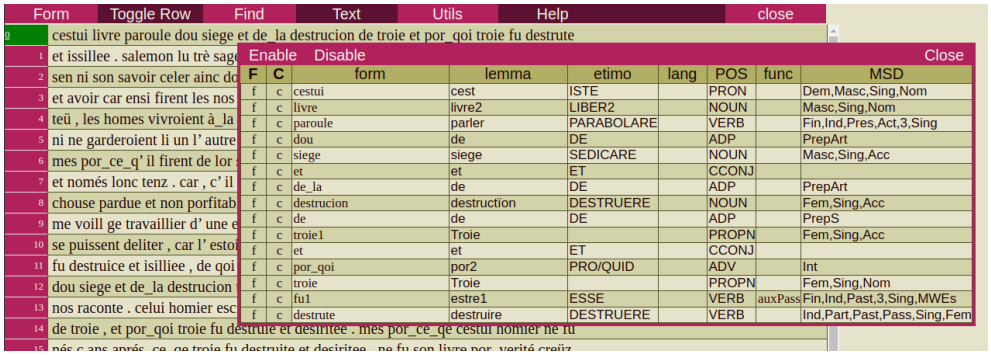


Fig. 20

Sempre nella schermata *Text* è possibile attivare il campo di ricerca: per *Form*, *Lemma*, *Etimo*, POS, MSD e LANG. La visualizzazione in modalità *Text* funziona quindi al tempo stesso da complemento della fase di correzione e da provvisorio *output* dei risultati dell'annotazione. Si tratta dell'aspetto più delicato del sistema su cui si sta continuando a lavorare: attualmente, infatti, la ricerca è possibile solo sui singoli testi del *corpus* (corrispondenti in questo caso ai singoli episodi di *Prose 2*), e il prossimo passo sarà renderla cumulativa sull'intero *corpus* o su una selezione di testi.

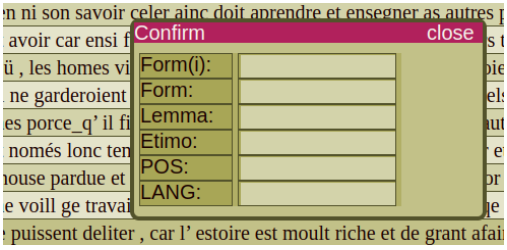


Fig. 21

7. *La sequenza di lavoro ULA*

Diversamente dagli strumenti citati, LGeRM e *Pyrrha*, il lemmatizzatore ULA è stato concepito come strumento da utilizzare in locale. Esso è compatibile con qualsiasi sistema operativo e richiede come unico presupposto la pre-installazione di Python3. Una volta scaricato e installato, per lavorare con ULA sarà sufficiente digitare la linea di comando ulaserver.py in un terminale e aprire una porta dedicata nel browser. La compatibilità di ULA con qualsiasi sistema è ottenuta infatti anche dall'utilizzo del browser come interfaccia.

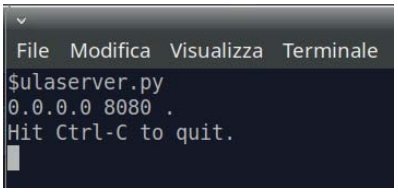
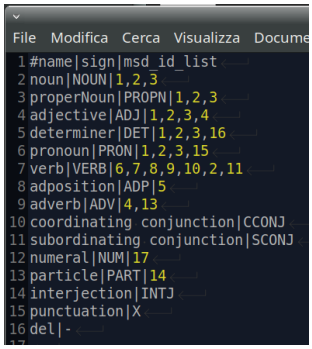


Fig. 22

1. Il primo passo è l'eventuale preparazione del file .TXT, necessaria, lo si ricorda, solo nel caso si volessero lemmatizzare in modo specifico alcuni elementi, cioè le poliremathiche e i clitici. Al di là dell'applicazione specifica ('_' per le poliremathiche, '-' per i clitici), in generale: l'*underscore* '_' unisce due elementi; il trattino '-' li separa. L'utente, in base al contesto linguistico e ai *desiderata* dell'analisi, sarà libero di applicarli alle realtà che ritiene opportune.

2. Contestualmente, a inizio progetto, l'utilizzatore può apportare anche le eventuali necessarie modifiche ai parametri del *tagset* (obbligatori: POS, MSD; opzionali: *Funct, Lang*) agendo sui *files* di configurazione delle tabelle: il *tagset* può anche essere completamente sostituito rispetto a quello proposto per default. Il legame fra la tabella dei POS e quella degli MSD si realizza attraverso una chiave numerica: a ciascuna combinazione di *features*, ad es. quella del genere, viene attribuito un numero, che costituisce poi la chiave da associare al POS.



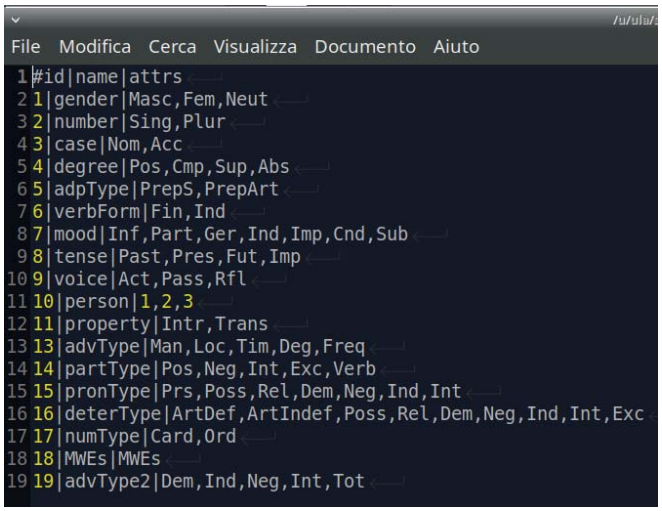


Fig. 23

3. Una volta caricato il testo nel sistema, è buona norma controllare il risultato della tokenizzazione per correggere eventuali errori sfuggiti in fase di preparazione. Chiaramente, l'importanza di un testo corretto al cento per cento contro un testo rispetto al quale si adotta un certo grado di tolleranza dell'errore dipende sia dagli obiettivi finali del processo di analisi, sia dall'entità stessa del *corpus*: ovviamente, quanto maggiore è la mole dei dati, tanto minore è l'impatto della presenza di

errori sulla significatività dei risultati. La correzione di eventuali errori, anche banalmente errori di digitazione che possono essere sfuggiti all'ennesimo controllo preventivo, può comunque realizzarsi anche in corso d'opera perché il sistema permette di correggere direttamente il testo tokenizzato, sia a livello di forma che di occorrenza, senza dover intervenire a monte sul *file* testo originale.

Al di là del contesto specifico, l'utilizzo di uno strumento che esporta il testo sotto forma di lista di forme, accompagnate però dalle occorrenze, può in realtà rivelarsi assai utile anche in un lavoro di 'semplice' edizione, come ulteriore strumento di correzione, permettendo ad esempio di mettere in evidenza eventuali attestazioni uniche che potrebbero essere frutto di un errore di trascrizione, o di evidenziare incongruenze nello scioglimento delle abbreviazioni.

4. A questo punto inizia la fase di lavoro vero e proprio. ULA non si avvale di un dizionario pre-caricato[84], almeno al suo primo utilizzo, ma costruisce il dizionario – un dizionario-macchina si badi bene, cioè un insieme di associazioni forma-lemma – di pari passo con il progredire dell'annotazione stessa del testo. Ovviamente, in un'ottica di accumulazione virtuosa del lavoro, il dizionario ottenuto da un primo lavoro potrà essere successivamente riutilizzato. Il primo testo caricato, al primo utilizzo in assoluto di ULA in un determinato contesto linguistico, dovrà quindi essere annotato manualmente nella sua interezza; a partire dal secondo testo il lavoro di immissione dei dati da parte dell'utente diminuirà sempre più in favore di quello di controllo e correzione, dal momento che il sistema riempirà automaticamente i campi corrispondenti alle forme già note.

1683	ver						
1684	veraie	verai	*VERACUS		ADJ		Fem,Sing,Acc
1685	veraiement	veraiement	*VERACUS		ADV		
1686	verais						
1687	verdoiant						
1688	vergogneuse						
1689	verité	verité	VERITAS		NOUN		Fem,Sing,Nom
1690	verra						
1691	verrais						

Fig. 24

Un testo molto lungo come quello di *Prose 2* è stato spezzato in episodi così da poter lavorare progressivamente.

Lemma ed *Etimo*, lo si ricorda, sono i soli due campi che richiedono una digitazione diretta da parte dell'utente, mentre tutte le altre informazioni (POS/MSD/*Lang*/*Funct*) sono selezionate attraverso le tabelle laterali.

Il grosso del lavoro in ULA si realizza dunque operando a livello delle forme, che nella maggior parte dei casi sono univoche; e dalle forme l'annotazione è

[84] Così come il prototipo di lemmatizzatore concepito da Glessgen 2003.

automaticamente trasferita a tutte le occorrenze. Come si è già detto, cliccando sul numero a sinistra della forma è possibile aprire la lista delle concordanze, con la possibilità di ampliare il contesto, fissato per default a cinque parole a sinistra e a destra, fino a nove parole,

Fig. 25

fondamentale soprattutto quando ci si trova di fronte a forme grafiche che possono celare occorrenze omografe. In questo caso si creeranno allora delle forme-*alias*, la forma più un numero progressivo (forma, forma1, forma2, ecc.), ciascuna con la sua specifica combinazione *Lemma/Etimo/POS/MSD*, e si selezioneranno e assoceranno le occorrenze interessate. Nei testi analizzati successivamente al primo in cui è emersa l'esistenza di un'omografia, le forme interessate rispetto alle quali sono già stati creati degli *alias* risulteranno al centro di un conflitto di attribuzione dal momento che il dizionario-macchina registra più opzioni, che saranno perciò graficamente segnalate. Contestualmente, insieme al *pop-up* delle occorrenze, verrà aperto un secondo *pop-up* con l'elenco dei casi già incontrati e annotati, vale a dire delle forme-*alias* già create: l'utente potrà allora o scegliere fra le opzioni già disponibili oppure, se necessario, aggiungere ulteriori forme-*alias*.

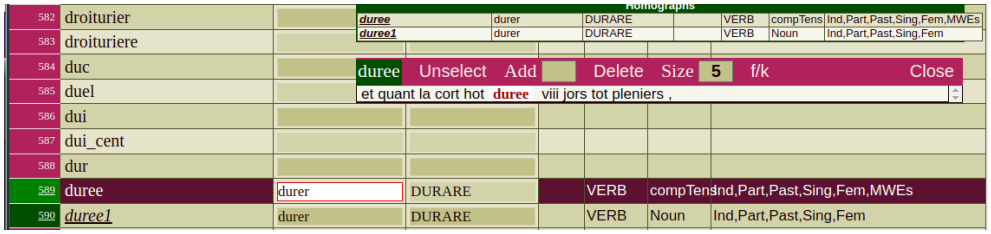
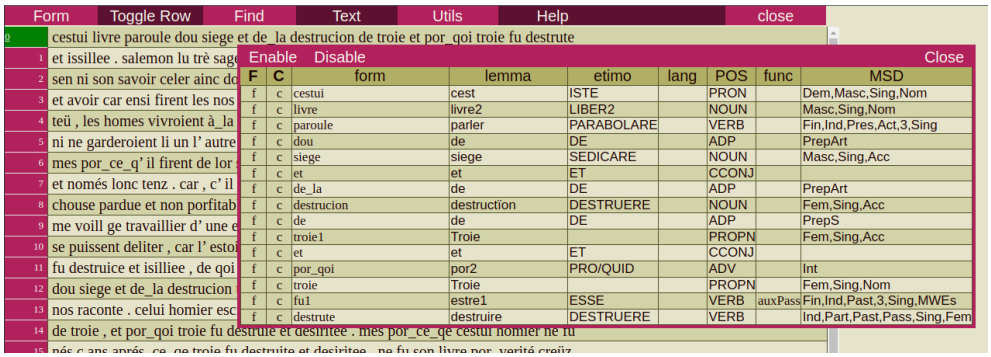


Fig. 26

5. Una volta terminato il lavoro sulle forme, sarà possibile realizzare un ulteriore controllo dei dati passando alla visualizzazione contestuale, la visualizzazione cioè del testo nella sua forma originale e non destrutturata (voce *Text* del Menu). Scorrendo riga per riga il testo, apparirà una finestra con la stessa riga in versione lemmatizzata e annotata. A questo punto si possono riscontrare eventuali errori o a livello di forma o a livello di associazione forma-occorrenza: tramite due comandi appositi (*F*/Forma, *C*/*Context*) è allora possibile puntare direttamente alla forma o all'occorrenza interessata nello spazio di lavoro (non è infatti possibile correggere direttamente nello spazio *Text*).



F	C	form	lemma	etimo	lang	POS	func	MSD
f	c	cestui	cest	ISTE		PRON		Dem,Masc,Sing,Nom
f	c	livre	livre2	LIBER2		NOUN		Masc,Sing,Nom
f	c	paroule	parler	PARABOLARE		VERB		Fin,Ind,Pres,Act,3,Sing
f	c	dou	de	DE		ADP		PrepArt
f	c	siege	siege	SEDICARE		NOUN		Masc,Sing,Acc
f	c	et	et	ET		CCONJ		
f	c	de_la	de	DE		ADP		PrepArt
f	c	destrucion	destruction	DESTRUERE		NOUN		Fem,Sing,Acc
f	c	de	de	DE		ADP		PrepS
f	c	troie1	Troie			PROPN		Fem,Sing,Acc
f	c	et	et	ET		CCONJ		
f	c	por_qoi	por2	PRO/QUID		ADV		Int
f	c	troie	Troie			PROPN		Fem,Sing,Nom
f	c	fu1	estre1	ESSE		VERB	auxPass	Fin,Ind,Past,3,Sing,MWEs
f	c	destrute	destruire	DESTRUERE		VERB		Ind,Part,Past,Pass,Sing,Fem

Fig. 27

6. Attualmente i dati della lemmatizzazione sono esportati in un file .CSV, in attesa di ulteriori implementazioni dello strumento. Così come sarebbe auspicabile la possibilità di inserire i dati di *input* sotto forma di codifica TEI, così si vorrebbe offrire la possibilità anche di esportarli in questo stesso formato. Nel momento in cui l'implementazione di ULA renderà possibile questa opzione, lo schema di codifica proposto sarà strettamente TEI conforme: <w lemma="" pos="" msd="" >. Sarà possibile aggiungere come opzioni: @lemmaref, un puntatore per una risorsa esterna, in primo luogo un lessico o un dizionario; @xml:lang, nel caso in cui in fase di lemmatizzazione fossero stati inseriti anche i dettagli linguistici; @function, questo attributo accoglierebbe invece l'eventuale livello sintattico dell'annotazione; mentre l'attributo generalista @ana permetterà di inserire le informazioni contenute nel campo *Funct* dell'annotazione lessicale.

8. *Un primo bilancio*

Come si è detto all'inizio, il progetto PRODIGI prevede l'annotazione lessicale completa del testo di *Prose 2* così come trasmesso dal manoscritto di Grenoble. Per raggiungere questo obiettivo è stato creato uno strumento *ad hoc*, l'interfaccia di lemmatizzazione ULA, che rappresenta una sorta di progetto all'interno del progetto. Nei mesi futuri, al termine del processo di annotazione, tutt'ora in corso, sarà possibile tracciare il quadro lessicale del testo troiano-padovano (si ricorda che il ms. è stato copiato nelle carceri padovane), rispondendo a tutta una serie di quesiti che vanno dal piano statistico al piano linguistico vero e proprio: per un testo costituito da X occorrenze/*tokens*, quante forme sono individuabili? quanti lemmi? di quale ampiezza cioè è il bagaglio lessicale di un testo in prosa di tale estensione? quali sono le tipologie di interferenza linguistica, fra varietà italiana-settentrionale e *langue d'oil*? come si distribuiscono fra il piano fonetico, morfologico, sintattico e lessicale? ecc. ecc.

In attesa di poter procedere, tutti i dati alla mano, a questo più corposo bilancio, le note qui presentate hanno rappresentato un bilancio di *midterm*, presentando lo strumento concepito per poter raggiungere l'obiettivo principale di PRODI-

GI. In questo caso, tuttavia, l'interesse non è stato tanto nel risultato finale – uno strumento che può essere funzionale in alcuni casi, con determinate esigenze – quanto nel lavoro di riflessione a cui obbliga sempre la progettazione di un *tool* informatico al servizio della ricerca umanistica. In una certa prospettiva proprio l'attività di modellizzazione in sé per sé, al di là della realizzazione stessa del modello, potrebbe essere considerata il cuore pulsante della disciplina denominata Informatica Umanistica. Progettare un'interfaccia digitale di annotazione lessicale, per quanto rudimentale essa sia, ha rappresentato quindi l'occasione per prendere coscienza, misurandosi sul campo, delle principali problematiche che emergono dallo scontro fra l'intrinseca ambiguità del linguaggio umano e l'intrinseca 'disambiguità' del linguaggio binario, affinando anche, attraverso questa esperienza, in un circolo virtuoso, la capacità di valutazione degli altri strumenti e delle possibilità e soluzioni offerte da ciascuno.

Bibliografia

I. Manoscritti

Grenoble BM 861	Bibliothèque Municipale		861
Oxford BL Douce 196	Bodleian Library	Douce	196
Paris BnF n.a.fr. 9603	Bibliothèque nationale de France	nouv. acq. fr.	9603

II. Opere

Prose 5

Le Roman de Troie en prose. Prose 5, édité par Anne Rochebouet, Paris, Classiques Garnier, 2021 («Textes littéraires du Moyen Age», 59).

Binduccio dello Scelto, *Storia di Troia*

La storia di Troia, a cura di Maria Gozzi, Roma, Carocci, 2000 («Biblioteca medievale», 77).

Storia di Troia, a cura di Gabriele Ricci, Parma, Guanda, 2004.

III. Studi e strumenti

Artale 2013

Elena Artale, *Funzioni grammaticali e valore verbale in lessicografia. Alcuni casi di gerundio nel TLIO: lemmatizzazione e redazione*. Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas (6-11 septiembre 2010),

Emili Casanova Herrero and Cesáreo Calvo Rigual (éds), 8 voll., Berlin-Boston, De Gruyter, 2013, vol. VIII, pp. 43-54.

Bambini – Trevisan 2012
Valentina Bambini, Marco Trevisan, *EsploraCOLFIS: Un'interfaccia web per le ricerche nel Corpus e Lessico di Frequenza dell'Italiano Scritto (COLFIS)*, in «Quaderni del laboratorio di linguistica», 11 (2012), pp. 1-16.

Beltrami (*et alii*) 2013
Opera del Vocabolario Italiano. Norme per la redazione del Tesoro della Lingua Italiana delle Origini, a cura di Pietro G. Beltrami, con la collaborazione dei redattori e revisori del TLIO, Versione aggiornata 2013, https://www.dilass.unich.it/sites/st06/files/normetlio.pdf [cons. 24. VII. 2022].

Bertrand (*et alii*) 2019
Tutoriel TXM pour la BFM, Version 3.1, 15 mai 2019, créé par Lauranne Bertrand, Céline Guillot, Serge Heiden, Alexei Lavrentiev, Bénédicte Pincemin, http://bfm.ens-lyon.fr/IMG/pdf/tutoriel_txm_bfm_v3.1.pdf [cons. 24. VII. 2022].

Bollmann – Søgaard 2016
Marcel Bollmann, Anders Søgaard, *Improving historical spelling normalization with bi-directional LSTMs and multi-task learning*, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, Osaka, The COLING 2016 Organizing Commitee, December 2016, pp. 131-139, https://aclanthology.org/C16-1013.pdf [cons. 24. VII. 2022].

Bucholz-Marsi 2006
Sabine Bucholz, Erwin Marsi, *CoNLL-X Shared Task on Multilingual Dependency Parsing*, in *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, New York City, Association for Computational Linguistics, June 2006, pp. 149-164, https://aclanthology.org/W06-2920.pdf [cons. 24. VII. 2022].

Busa 2001
Roberto Busa S.J., *Le tre informatiche*, in «La città del secondo rinascimento», 2 (settembre 2001) [= *I nuovi media dell'arte, dell'impresa, della finanza*], http://www.ilsecondorinascimento.it/Pages/TxtBUSA.htm [cons. 24. VII. 2022].

Calzolari (*et alii*) 1995
Towards a Network of European Reference Corpora: Report of the Nerc Consortium Feasibility Study, edited by Nicoletta Calzolari, Mona Baker, Johanna G. Kruyt, Pisa, Giardini, 1995 («Linguistica Computazionale», XI-XII).

Calzolari – Monachini 1995
Nicoletta Calzolari, Monica Monachini, *EU LRE Project 62-050. MULTEXT. Work package 1. Milestone B. Deliverable D1.6.1B. Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets*, edited by Nuria Bel, Nicoletta Calzolari and Monica Monachini, March 1995, https://nl.ijs.si/ME/Vault/V3/msd/related/msd-multext/ [cons. 24. VII. 2022].

Cambi 2016
Matteo Cambi, *"In carcere Ianuentium". Fonti e nuovi documenti sul "milieu" carcerario genovese (1284-1300)*, in «Aevum», 90/2 (Maggio-Agosto 2016), pp. 401-416.

Carlesso 1966
Giuliana Carlesso, *La versione sud del "Roman de Troie en prose" e il volgarizzamento di Binduccio dello Scelto*, in «Atti dell'Istituto veneto di scienze, lettere ed arti. Classe di scienze morali, lettere ed arti», 124 (1966), pp. 519-560.

Catach 1996
Laurent Catach, *Graphist: Logiciel de lemmatisation, indexation et modernisation automatique de textes anciens*, in «Digital Studies / Le champ numérique», 4 (1996), http://doi.org/10.16995/dscn.215 [cons. 24. VII. 2022].

Ceresato 2021
Floriana Ceresato, *L'analisi lessicale dell'"Entrée d'Espagne": bilancio di una prima sperimentazione*, in «Francigena», 7 (2021), pp. 355-381.

Chiarcos *et alii* 2020
Christian Chiarcos, Christian Fäth, Frank Abromeit, *Annotation Interoperability for the Post-ISOCat Era*, in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020) (May 11-16, 2020, Palais du Pharo, Marseille, France)*, European Language Resources Association, 2020, pp. 5668-5677, https://aclanthology.org/2020.lrec-1.696.pdf [cons. 24. VII. 2022].

Cigni 2006
Fabrizio Cigni, *Copisti prigionieri (Genova, fine sec. XIII)*, in *Studi di filologia romanza offerti a Valerio Bertolucci Pizzorusso*, a cura di Pietro G. Beltrami, Maria Grazia Capusso, Fabrizio Cigni, Sergio Vatteroni, 2 voll., Pisa, Pacini Editore, 2006, vol. I, pp. 425-439.

Cursi 2009
Marco Cursi, *"Con molte sue fatiche": copisti in carcere alle Stinche alla fine del Medioevo (secoli XIV-XV)*, in *In uno volumine. Studi sul libro e il documento in*

età medievale offerti a Cesare Scalon, a cura di Laura Pani, Udine, Forum Edizioni, 2009, pp. 151-192.

Dereza 2018
Oksana Dereza, *Lemmatization for Ancient Languages: Rules or Neural Networks?*, in *Artificial Intelligence and Natural Language*. 7th International Conference, AINL, 2018 (St. Petersburg, 17-19 October 2018), Cham, Springer, 2018, pp. 35-47.

Dotto 2012
Diego Dotto, *Note per la lemmatizzazione del corpus DiVo*, in «Bollettino dell'Opera del Vocabolario Italiano», 17 (2012), pp. 339-366.

Eger (*et alii*) 2016
Steffen Eger, Rüdiger Gleim, Alexander Mehler, *Lemmatisation and Morphological Tagging in German and Latin: A Comparison and a Survey of the state-of-the-art*, in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC '16)*, Portoroz, European Language Resources Association (ELRA), May 2016, pp. 1507-1513, https://aclanthology.org/L16-1239 [cons. 24. VII. 2022].

Fois 2021
Jacopo Fois, *Un capitolo dell'espansione della leggenda troiana in Italia: note sul manocritto di "Prose 2" Grenoble, Bibliothèque Municipale, 861 (263 Rés.)*, in «Carte Romanze», 9/1 (2021), pp. 199-223, https://doi.org/10.13130/2282-7447/15199 [cons. 24. VII. 2022].

Gabay (*et alii*) 2020
Simon Gabay, Jean-Baptiste Camps, Thibault Clérice, *Manuel d'annotation linguistique pour le français moderne (XVIe-XVIIIe siècles), Version A ("Absidale Argélie")*, 12 mai 2020, https://hal.archives-ouvertes.fr/hal-02571190v1 [cons. 24. VII. 2022].

Giuliani 2008
Mariafrancesca Giuliani, *Le polirematiche nel TLIO: pratiche lessicografiche, dati e criteri di classificazione*, in *Proceedings of the XIII EURALEX International Congress (Barcelona, Universitat Pompeu Fabra, 15-19 July 2008)*, edited by Elisenda Bernal e Janet DeCesaris, Barcelona, Insitut Universitari de Linguistica Aplicada Universitat Pompeu Fabra, 2008, («Sèrie activitats», 20), pp. 1123-1138, https://euralex.org/publications/le-polirematiche-nel-tlio-pratiche-lessicografiche-dati-e-criteri-di-classificazione/ [cons. 24. VII. 2022].

Gleim (*et alii*) 2019
Rüdiger Gleim, Steffen Eger, Alexander Mehler, Tolga Uslu, Wahed Hemati,

Andy Lücking, Alexander Henlein, Sven Kahlsdorf, Armin Hoenen, *A practitioner's view: a survey and comparison of lemmatization and morphological tagging in German and Latin*, in «Journal of Languages Modelling», 7/1 (2019), pp. 1-52.

Glessgen 2003a

Martin-Dietrich Glessgen, *L'élaboration philologique et l'étude lexicologique des "Plus anciens documents linguistiques de la France" à l'aide de l'informatique*, in *Frédéric Godefroy*. Actes du X^e^ colloque international sur le moyen français (Metz, 12-14 juin 2002), textes réunis et présentés par Frédéric Duval, Paris, École des chartes, 2003, pp. 371-386.

Glessgen 2003b

Martin-Dietrich Glessgen, *La lemmatisation de textes d'ancien français: méthodes et recherches*, in *Ancien et moyen français sur le web. Enjeux méthodologique et analyse du discours*, sous la direction de Pierre Kunstmann, France Martineau, Danielle Forget, Ottawa, Les Éditions David, 2003 («Voix savantes», 20), pp. 55-75.

Gozzi 2000

Maria Gozzi, *Dal "Roman de Troie" di Benoît de Sainte-Maure al "Libro de la storia di Troia" di Binduccio dello Scelto: metamorfosi di un testo*, in *La lotta con Proteo: metamorfosi del testo e testualità della critica*. Atti del XVI Congresso AISLLI, Associazione Internazionale per gli Studi di Lingua e Letteratura Italiana (Los Angeles, University of California-UCLA, 6-9 ottobre 1997), a cura di Luigi Ballerini, Gay Bardin e Massimo Ciavolella, 2 voll., Fiesole, Cadmo, 2000, vol. I, pp. 457-464.

Grübl 2013

Klaus Grübl, *La standardisation du français au Moyen Age: point de vue scriptologique*, in «Revue de Linguistique Romane», 307-308/77 (2013), pp. 343-383.

Ide – Véronis 1993

Nancy Ide, Jean Véronis, *What next after the Text Encoding Initiative? The need for text software*, in «ACH Newsletter», (1993), pp. 1-12.

Ide – Véronis 1994

Nancy Ide, Jean Véronis, *MULTEXT: Multilingual Text Tools and Corpora*, in *COLING 1994, Volume 1: The 15^th^ conference on Computational Linguistics*, Kyoto, August 1994, pp. 581-592, https://aclanthology.org/C94-1097.pdf [cons. 24. VII. 2022].

Ide – Veronis 1995
Text Encoding Initiative: Background and Context, edited by Nancy Ide, Jean Veronis, Dordrecht-Boston-London, Kluwer, 1995.

Iorio-Fili 2007
Domenico Iorio-Fili, *Breve storia, stato attuale e prospettive del software GATTO*, in «Bollettino dell'Opera del Vocabolario Italiano», 12 (2007), pp. 365-386.

Iorio-Fili 2012
Domenico Iorio-Fili, *GATTO. Versione 3.3. Manuale d'uso*, CNR, Istituto Opera del Vocabolario Italiano, 2012.

Kestemont (*et alii*) 2016
Mike Kestemont, Guy De Pauw, Renske van Nie, Walter Daelemans, *Lemmatization for variation-rich languages using deep learning*, in «Digital Scholarship in the Humanities», 32/4 (26 August 2016), pp. 797-815, https://doi.org/10.1093/llc/fqw034 [cons. 24. VII. 2022].

Korfanty 1999
Sylvie Korfanty, *Lexicography et glossographie du français du XVI^e^ siècle. Prolegomenes à un dictionnaire du français preclassique*, sous la direction de Claude Buridant, soutenue en 1999 à Strasbourg 2, Thèse dactylographié, UMB Strasbourg, 1999.

Laporte 1997
Eric Laporte, *Les mots, un demi-siècle de traitements*, in «TAL», 38/2 (1997), pp. 47-68.

Laporte 2000
Eric Laporte, *Mots et niveau lexical*, in *Ingénierie des langues. Série Information, Commande, Communication*, sous la direction de Jean-Marie Pierrel, Paris, Hermès, 2000, pp. 25-49.

Lenci – Montemagni 2002
Alessandro Lenci, Simonetta Montemagni, *La Treebank sintattico-sematica dell'italiano di SI-TAL*, in *Matématicas y tratamiento de corpus*. Actas del segundo seminario de la Escuela interlatina de altos estudios en lingüíistica aplicada (San Millán de la Cogolla, 19-23 de septiembre de 2000), Fundación San Millán de la Cogolla, 2002, pp. 221-244.

Manjavacas (*et alii*) 2019
Enrique Manjavacas, Akos Kadar, Mike Kestemont, *Improving Lemmatization of Non-Standards Languages with Joint Learning*, in *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Minneapolis (Minnesota), 2 voll., 2019, vol. I, pp. 1493-1503, https://aclanthology.org/N19-1153.pdf [cons. 24. VII. 2022].

de Marneffe (*et alii*) 2021
Marie-Catherine de Marneffe, Chistopher D. Manning, Joakim Nivre, Daniel Zeman, *Universal Dependencies*, in «Computational Linguistics», 47/2 (2021), pp. 255-308, https://doi.org/10.1162/coli_a_00402 [cons. 24. VII. 2022].

Materni 2020
Marta Materni, *DigiFlorimont: une occasione de réflexion philologique et numérique autour de la représentation de la complexité textuelle médiévale*, in *Autour du "Roman de Florimont". Approches multidisciplinaires à la complexité textuelle médiévale*, Padova, 2020 («Quaderni di Francigena», 2), pp. 165-184, https://phaidra.cab.unipd.it/o:453784 [cons. 24. VII. 2022].

Materni 2021
Marta Materni, *Les Guidelines TEI à l'épreuve de la complexité textuelle et graphique médiévale*, in «Magnificat. Revista de Cultura i Literatura Medievals», 8 (2021), pp. 1-32, https://ojs.uv.es/index.php/MCLM [cons. 24. VII. 2022].

Pettersson (*et alii*) 2014
Eva Pettersson, Beáta Megyesi, Joakim Nivre, *A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text*, in *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, Gothenburg, Association for Computational Linguistics, 2014, pp. 32-41, https://aclanthology.org/W14-0605 [cons. 24. VII. 2022].

Prévost (*et alii*) 2009
Sophie Prévost, Céline Guillot, Alexei Lavrentiev, Serge Heiden, *Jeu d'étiquettes morphosyntaxiques CATTEX 2009*, *Version 2.0. 2013-04-08*, http://bfm.ens-lyon.fr/IMG/pdf/Cattex2009_2.0.pdf [cons. 24. VII. 2022]

Rodeghiero 2021
Sira Rodeghiero, *Strumenti e criteri per la lemmatizzazione del franco-italiano: verso la costruzione di un* corpus *lemmatizzato della* Geste Francor, in «Francigena», 7 (2021), pp. 305-348.

Roger 2017
Geoffrey Roger, *Les scriptae régionales du moyen français. Pour l'analyse transversale des sources du DMF*, in *La mise à l'écrit et ses conséquences*. Actes du troisième

colloque *Repenser l'histoire du français* (Université de Neuchatel, 5-6 juin 2014), éd. par Andres M. Kristol, Tübingen, A. Francke Verlag, 2017, pp. 109-152.

Tittel 2015
Sabine Tittel, *Les exigences d'une lexicographie de corpus de l'ancien français à grande échelle: l'établissement d'un corpus de référence et d'un étiquetage sémantique*, in *Quelle philologie pour quelle lexicographie.* Actes de la section 17 du XXVII[e] Congrès international de linguistique et philologies romanes, édités par Sthephen Dörr, Yan Greub, Heidelberg, Universitätsverlag Winter, 2015, («Studia Romanica», 197), pp. 129-148.

Walker (*et alii*) 1995
Automating the Lexicon. Research and Practice in a Multilingual Environment, edited by Donald E. Walker, Antonio Zampolli, Nicoletta Calzolari, Oxford, Oxford University Press, 1995.

Zeman (*et alii*) 2021
Daniel Zeman *et alii*, *Universal Dependencies 2.9*, LINDAT/CLARIAH-C7 digital library at the Institute of Formal and Applied Linguistics (UFAL), Faculty of Mathematics and Physics, Charles University, 2021, https://lindat.cz/repository/xmlui/handle/11234/1-4611 [cons. 24. VII. 2022].