*Revue, Informatique et Statistique dans les Sciences humaines*
XXII, 1-4, 1986, pp. 171-178.

# Alcune osservazioni sulle tendenze della lemmatizzazione latina

## Lessicografia e informatica

G. SPINOSA

La crescente utilizzazione dei calcolatori nella pratica della lessicografia ha avuto, tra gli altri effetti, quello di stimolare il dibattito intorno ai metodi che presiedono alla compilazione di dizionari e lessici. Sia da parte degli autori di tali strumenti, che da parte degli utenti, si è andata manifestando la tendenza ad una sempre maggiore esplicitazione di quei criteri che operavano implicitamente nella lessicografia tradizionale. Ciò ha portato maggiore chiarezza in molti ambiti e ha contribuito a ridare il giusto peso a non poche questioni.

Tra queste la lemmatizzazione è un tema intorno a cui il dibattito si è fatto più vivace, arricchendosi di molteplici apporti [1]. Una attenta considerazione delle varie proposte che sono state avanzate

1. Segnaliamo qui: R. Busa (a cura di), *Actes du séminaire international sur le dictionnaire latin de machine*, Pise, 27-28 mars 1968, in *Calcolo*, 1968, dedicato al confronto di due dizionari di macchina del latino, il Lessico Elettronico Latino di Liegi e quello di Gallarate; A. Duro, *Quelques suggestions dictées par l'expérience d'un vieux lexicographe*, Roma 1977, U. Keudel, *Bemerkungen zum Plan eines 'Thesaurus Mediae et Recentioris Latinitatis'*, Roma 1977, R. Busa, *Notes lexicologiques et lexicographiques sur lemme, formes et lemmatisation*, Roma 1978, comunicazioni presentate nel corso dei numerosi incontri e seminari organizzati dal *Lessico Intellettuale Europeo* sulla lessicografia latina medievale e in particolare sulla lemmatizzazione, nell'ambito del progetto del LIE per un *Thesaurus Mediae et Recentioris Latinitatis* di cui si dirà più avanti; A. Duro, *Quelques remarques à propos de la lemmatisation*, in *Lexicologie, een bundel opstellen voor F. De Tollenaere*, Groningen

../..

in questo campo permette d'altra parte di rilevare l'esistenza di una linea di convergenza delle varie tendenze. Sia le équipes che lavorano con i metodi manuali tradizionali, sia quelle che operano con il sussidio dell'elaboratore elettronico hanno infatti riconosciuto l'esigenza di evitare il pericolo di farsi sommergere e disorientare dalla vasta mole di attestazioni. Per fare solo alcuni esempi si pensi ai milioni di schede del *Thesaurus Linguae Latinae* o ai milioni di occorrenze raccolte su nastro presso il CETEDOC, o ancora al vastissimo *corpus* dell'*Index Thomisticus*. Tra queste iniziative si è andato di fatto affermando un modello di lemmatizzazione assai unificante rispetto alla potenziale moltiplicabilità dei lemmi autorizzata dalle leggi della morfologia e della sintassi latine. La riconduzione dei participi e degli aggettivi sostantivati rispettivamente sotto il verbo e sotto l'aggettivo, e degli avverbi sotto l'aggettivo o sotto il verbo operata nel TLL; la formulazione di una lemmatizzazione "di tipo formale" proposta da P. Tombeur; la classificazione dei lemmi in nominali, verbali e indeclinabili operata dal padre Busa, costituiscono la conferma di tale tendenza[1].

---

../..

1977, pp. 133-139; ID., *Pour un 'Thesaurus Mediae et Recentioris Latinitatis', Aspects techniques*, in *ORDO, II Colloquio Internazionale del LIE*, Roma 7-9 gennaio 1977, a cura di M. Fattori e M. Bianchi, Roma 1979, pp. 739-745; ID., *Un angoissant problème de lemmatisation: le traitement du participe*, in *Actes de la Seconde Table Ronde Internationale de Lexicographie Historique*, Leyde, mai 1977, Dordrecht-Cinnaminson 1980, pp. 117-142.

Nel retroterra di questi dibattiti si collocano gli apporti di glottologi e linguisti come S. D'A. Avalle che in *Bassa latinità. Le strutture morfologiche del nome nel latino medievale*, Torino 1971, dimostra con analitica documentazione come nei testi medievali la morfologia del nome, assai più di quella delle altre parti variabili del discorso, sia un settore aperto a mutamenti nella prospettiva del nuovo assetto della flessione romanza.

1. Il confronto operato da V. Lomanto sui primi 1000 lemmi del *Thesaurus Linguae Latinae*, del *Lexicon Totius Latinitatis* del Forcellini, e dell'*Ausführlisches Lateinisches-Deutsches Handwörterbuch* del Georges ha avuto, tra gli altri risultati, quello di confermare da un lato l'esistenza di notevoli divergenze negli strumenti lessicografici per quanto riguarda l'individuazione del lemma, e dall'altro la natura notevolmente unificante della lemmatizzazione operata dal *Thesaurus Linguae Latinae*. E' emerso infatti che il *Thesaurus* ha un maggior numero di sottolemmi (76) rispetto al Forcellini e al Georges; così come ha un maggior numero di 'lemmi secondari' (lemmi che seguono il principale con gli stessi caratteri tipografici) (25), rispetto al Forcellini (7) e al Georges (10). Mentre all'inverso, in rapporto complementare alla proporzione precedente, il *Thesaurus* ha un numero minore di lemmi di rimando (66) rispetto al Forcellini (104), e un numero uguale rispetto al Georges (66).

../..

La fertilità di queste proposte si rivela poi al massimo grado nel trattamento di un materiale lessicale che abbia subito molteplici trasformazioni, sia grafiche che morfologico-sintattiche (oltre che semantiche), come quello costituito da ampi *corpora* che raccolgono vocaboli afferenti al latino medievale e postmedievale. La necessità di un raccordo e di una sistemazione organica di tale materiale appare oggi un'esigenza primaria che può essere garantita dall'adozione di principi chiari e unitari, che in ogni caso non prescindano da una corretta impostazione filologica. Se è vero dunque che le strutture lessicali di fondo caratterizzanti la lingua latina si sono mantenute costanti nel corso della sua evoluzione, e che circa la metà del patrimonio lessicale mediolatino è costituita da vocaboli classici, mentre l'altra metà è costituita da neologismi che sono ora vocaboli formati da suffissi (*-ivus, -alis, -tas* ecc.) innestati in modo originale su radicali del latino classico; ora termini provenienti dalle lingue nazionali; ora termini scientifici (come gli imprestiti dal greco e dall'arabo); se è vero tutto ciò, si può concludere che tali proposte trovano una loro legittimazione anche da un punto di vista strettamente filologico-linguistico.

D'altra parte l'unificazione operata non intacca la molteplicità e la ricchezza della lingua latina, poichè il lemma più generale e formale che costituisce l'entrata si pone come una guida per una consultazione più analitica, percorribile attraverso lo studio dei sottolemmi. Esso permette di evitare, nella ricerca di un vocabolo, le ripetute operazioni di consultazione di entrate graficamente o morfologicamente diverse, operazioni che implicano spesso il rischio di trascurare una variante che può essere di particolare interesse per lo studioso. In ogni caso, la presenza di entrate di rinvio ai lemmi costituisce un'utile integrazione al sistema adottato.

---

../..

*Cf.* V. Lomanto, *Lessici latini e lessicografia automatica*, in "Memorie dell'Accademia delle scienze di Torino", Classe di scienze morali, storiche e filologiche, 1980, 2, pp. 111-270.

La constatazione della pluralità dei criteri di individuazione del lemma adottati in tali dizionari si pone come un'ulteriore prova della già riconosciuta impossibilità di rifarsi passivamente ad un singolo dizionario di riferimento in vista della costruzione di un lemmario latino che costituisca uno strumento per l'analisi, sia manuale che informatica, di *corpora* lessicali di vasta estensione cronologica.

In questa prospettiva, la proposta che si muove ad un più alto grado di formalizzazione e unificazione puo essere individuata nella lemmatizzazione "di tipo formale" formulata da P. Tombeur[1] che parte da un'esplicitazione dei criteri adottati dal *Thesaurus Linguae Latinae*. Il criterio che qui prevale nella individuazione del lemma è quello della struttura linguistico-lessicale della lingua latina, riservando ai caratteri morfologici, sintattici e semantici la determinazione dei sottolemmi. Ciò implica che i derivati lessicali, quali i sostantivi in *-us* della quarta declinazione, o i cosiddetti sostantivi mobili (*dominus, domina*) avranno un lemma autonomo. Al contrario, gli aggettivi sostantivati e i participi presenti e passati sostantivati, derivati morfologici degli aggettivi e dei participi, saranno ricondotti rispettivamente sotto l'aggettivo e sotto il verbo. Così gli avverbi derivati dagli aggettivi andranno sotto l'aggettivo, mentre gli avverbi derivati da usi aggettivali del participio andranno sotto il verbo. Per i lemmi omografi, interviene a individuarli non il criterio semantico, bensì il principio dell'etimologia. Dunque le oscillazioni e gli slittamenti da un significato ad un altro che non trovino giustificazione in una diversa etimologia non autorizzeranno a creare due lemmi omografi[2].

L'adozione del livello immediatamente sottostante a quello "formale", ovvero il livello morfologico (in una ideale scala dal generale al particolare) costituisce anch'esso una scelta che produce risultati notevolmente generalizzanti. E' quanto è riscontrabile nella classificazione morfologica dei lemmi del *Lexicon Totius Latinitatis*

1. *Cf.* P. Tombeur, *Instrumenta Lexicologica Latina, Pour un Thesaurus Patrum Latinorum*, in *Sancti Hilarii Pictaviensis episcopi De Trinitate, Corpus Christianorum, Series Latina*, LXII-LXIIA, Turnhout 1980, pp. 706-710. Tali criteri trovano applicazione negli *Instrumenta lexicologica Latina* del *Corpus Christianorum, Series Latina* e *Continuatio Mediaevalis*, fascicoli 1-21. Gli sviluppi di questa concezione sono stati messi a punto dallo stesso P. Tombeur nella relazione *Propositions nouvelles pour une lemmatisation unifiée du latin*, in *Proceedings of the Workshop "On the Possibilities and Limits of the Computer in Producing and Publishing Dictionaries"*, Pisa, maggio 1981, in corso di stampa.

2. Il principio dell'etimologia per l'individuazione dei lemmi sta alla base del lavoro del *Thesaurus Linguae Latinae* così come del *Mittellateinisches Wörterbuch*. A questo proposito si vedano le dettagliate osservazioni di Th. Payr sulla determinazione della forma del lemma : *Dictionnaire du Latin Médiéval. Remarques sur la méthode*, in *La lexicographie du latin médiéval et ses rapports avec les recherches actuelles sur la civilisation du Moyen-Age*, Paris 18-21 octobre 1978, Colloques Internationaux du CNRS, Paris 1981, pp. 477-478.

del Forcellini operata dal padre Busa, in vista del trattamento informatico del *corpus* dell'*Index Thomisticus*[1]. L'analisi morfologica infatti individua le caratteristiche del lemma in quanto parola a sé, isolata dal contesto; a prescindere quindi dalle ulteriori caratteristiche sintattico-semantiche, nelle quali la valutazione del contesto arricchirebbe il lemma stesso, fino ad autorizzarne eventuali sdoppiamenti. Per tale classificazione è stata sufficiente l'attribuzione di due codici alfabetici e/o numerici per ciascun lemma; non si è perciò distinto p. es. tra sostantivi e aggettivi, né tra participi e aggettivi; ma si è solo distinta, per l'ambito nominale, la declinazione di appartenenza, il genere o il numero delle uscite del nominativo. All'ambito dei codici nominali si sono affiancati quello dei codici verbali, degli indeclinabili e dei codici speciali. Siffatto sistema si rivela del tutto sufficiente ai fini della comprensione del lemma, in particolare in vista di una sua applicazione nel campo informatico[2].

Recentemente anche l'*Istituto di Linguistica Computazionale*, che ha curato per il *Lessico Intellettuale Europeo* l'aspetto informatico dell'*Aristoteles latinus* e che collabora al progetto per un *Thesaurus Mediae et Recentioris Latinitatis* ha avanzato una proposta di trattamento informatico dei lemmi che permetta di raccordarli a livello morfologico. La scelta del livello morfologico di formalizzazione, che implica la possibilità di considerare lemmi tutte le forme che hanno una individualità (come i participi, i comparativi ecc.), è bilanciata dall'adozione di un "iperlemma" che si pone ad un livello superiore di unificazione. Quest'ultima appare inoltre favorita dalla possibilità informatica di far ruotare la gerarchia dei lemmi e dell'iperlemma (raccordati biunivocamente), in modo da poter ottenere di volta in volta diversi livelli di lemmatizzazione[3].

Un esempio di applicazione di tali tendenze possiamo offrirlo qui, presentando la lemmatizzazione di alcuni vocaboli che

1. *Cf.* R.Busa, *The morphological categories of Forcellini's Latin lemmas*, in *Lexicologie*, cit., pp. 79-106.

2. Tali considerazioni sono state svolte a più riprese dal P.Busa anche nel corso delle riunioni organizzate dal LIE sulla lessicografia latina medievale. Presso lo stesso P.Busa, a Venezia, si svolse nell'aprile del 1979, un seminario sulla lemmatizzazione computerizzata con un'ampia partecipazione di specialisti.

3. *Cf.* A.Bozzi, *Progetto di organizzazione di un vasto repertorio lessicale automatico della lingua latina*, in *Maia*, N.S., fasc. II, XXXIV, 1982, pp. 167-172.

entreranno a far parte del TMRL[1] nell'ambito di una ricerca da me condotta presso il LIE[2]. Si tratta di termini scelti tra quelli registrati negli indici greco-latini della collana *Aristoteles latinus*, che raccoglie le versioni latine delle opere di Aristotele realizzate nei secoli VI e soprattutto XII e XIII. L'utilità e versatilità di tale lemmatizzazione per questo tipo di materiale lessicale si manifesta almeno per due ordini di motivi. Primo, perché tali vocaboli, corrispettivi di vocaboli greci, sono registrati negli indici della collana *Aristoteles latinus* in forme spesso non lemmatizzate, in funzione della loro correlazione alle forme greche; secondo, perché dei vocaboli presenti in tali indici verranno segnalati nel TMRL solo quelli che, non risultando registrati in quattro principali dizionari di latino classico e cristiano[3], si presentano come 'vocaboli nuovi', potenzialmente appartenenti al lessico postclassico, e tali da giovarsi al massimo grado di una formalizzazione lessicale.

---

1. Il progetto per un *Thesaurus Mediae et Recentioris Latinitatis* avanzato dal LIE nel 1976, si propone di raccogliere con sussidi informatici il vasto materiale lessicale latino medievale e recenziore, prodotto da sempre più numerosi Centri lessicografici e singoli studiosi, e non registrato nei maggiori dizionari di latino classico e cristiano. Il LIE si ripromette in questo modo di fornire uno strumento che renda possibile una consultazione rapida ed esaustiva, ottenibile grazie all'unificazione degli accessi lessicali. *Cf.* T. Gregory, *Pour un 'Thesaurus Mediae et Recentioris Latinitatis'*, in *ORDO*, cit., pp. 719-738; A. Bozzi – S. Emmanuele, *Thesaurus Mediae et Recentioris Latinitatis*, in *RES, III Colloquio Internazionale del LIE*, Roma 7-9 gennaio 1980, a cura di M. Fattori e M. Bianchi, Roma 1982, pp. 377-392; ID., *Sistema per la redazione semiautomatica delle voci, Tecnica di gestione interattiva di dati lessicali relativi al TMRL*, in *SPIRITUS, IV Colloquio Internazionale del LIE*, Roma 7-9 gennaio 1983, Roma 1984; G. Stabile, *Per un TMRL. Un' iniziativa del LIE*, in "Studi Medievali", 3a serie, XXIII, I, 1982, pp. 441-445.
2. Per il posto occupato dai vocaboli dell'*Aristoteles latinus* nell'ambito del progetto per un TMRL e per una più dettagliata descrizione delle caratteristiche di tale ricerca si rinvia a G. Spinosa, *Gli indici dell'Aristoteles latinus: esperienze di lavoro in vista di una loro inclusione nel Thesaurus*, in *SPIRITUS*, cit., pp. 545-566; ID., *Alcune traslitterazioni nelle versioni greco-latine di Aristotele*, in "Lexicon philosophicum", Quaderni di terminologia filosofica e storia delle idee, 1-1985, pp. 117-126.
3. I dizionari prescelti sono l'*Oxford Latin Dictionary*, Oxford 1982, il *Thesaurus Linguae Latinae*, Leipzig 1900-, il *Lexicon Totius Latinitatis*, Patavii (1864-1887) 1940 del Forcellini, il *Dictionnaire latin-français des auteurs chrétiens*, Turnhout 1962 di A. Blaise.

Abbiamo così avverbi derivati da usi aggettivali di participi che vengono lemmatizzati sotto il rispettivo verbo. Quindi *commutatim* (ἐναλλάξ) è registrato sotto il lemma *commuto, adiacenter (praedico)* (προσκυτηγορέω) sotto *adiaceo, contraiacenter* (ἀντικείμενος) sotto *contraiaceo*.

Avverbi derivati da aggettivi che vengono registrati sotto l'aggettivo. Per cui *comprehensive* (συλλήβδην) si trova sotto il lemma *comprehensivus; argumentative* (ἐλεγκτικῶς) sotto *argumentativus; aliene* (ἀλλοτρίως) sotto *alienus; cyclopice* (κυκλωπικῶς) sotto *cyclopicus*.

Participi presenti e passati riportati sotto il verbo all'indicativo: *coordinatus* (σύστοιχος) sotto *coordino; comperegrinans* (συναπόδημος) sotto *comperegrinor; conchoreans* (θιασωτής) sotto *conchoreo; contraaffectans* (ἀντεραστής) sotto *contraaffecto*.

Per quanto riguarda le numerose traslitterazioni e gli adattamenti operati sul greco con mutevoli convenzioni grafiche, è parso opportuno collocarli sotto la grafia più aderente alla voce greca, sia nel caso che essa trovi riscontro nei testi, sia nel caso contrario. In quest'ultima eventualità tuttavia tale lemma sarà indicato tra parentesi quadre finché non se ne rinvenga una attestazione.

Abbiamo dunque: *aorti* (ἀορτή) sotto *aorta; agricus* (ἀγροῖκος) sotto *agroicus; apedepsia* (ἀπαιδευσία) sotto *apaedeusia; crimatisticus* (χρηματιστικός) sotto *chrematisticus; acrivodikeos* (ἀκριβοδίκαιος) sotto [*acribodicaios*]; *agonoterus* (ἀγονώτερος) sotto [*agonos*]; *antifilisis* (ἀντιφίλησις) sotto [*antiphilesis*]; *cytharistica* (κιθαριστική) sotto [*citharisticus*].

L'esemplificazione di varianti grafiche nelle traslitterazioni dal greco ci permette di osservare che il dibattito intorno alla lemmatizzazione trova coinvolto un settore ad essa strettamente collegato, quello dell'ortografia. Anche qui, di fronte all'esigenza affermatasi fin dai primi del '900 nell'ambito del *Thesaurus Linguae Latinae* di testimoniare anche le grafie cosiddette secondarie – esigenza rafforzatasi con il sempre maggiore sviluppo delle edizioni critiche e paleografiche dei testi, dove anche lo studio delle varianti grafiche è fondamentale – si è imposta la ricerca di un criterio di unificazione che non appiattisse la molteplicità della realtà lessicale. Ancora una volta il trattamento elettronico delle varianti grafiche, che non privilegi come nel

passato la sola grafia del periodo classico, si presenta come uno strumento in grado di offrire un sistema di raccordo di un materiale altrimenti disperso[1].

1. *Cf.* V. Lomanto, *Grafia del tardo latino nell'elaborazione elettronica dei testi*, in *La cultura in Italia fra tardo antico e alto medioevo*, Roma 1981, pp. 373-395. *Cf.* anche D.G. Brearly (*Texts and studies in latin orthography to 1977*, in "The classical world", LXXII, 1979, pp. 384-392) che oltre ad offrire una panoramica dei principali studi moderni sull'ortografia latina, ne sottolinea i legami con la linguistica, la lessicografia, la fonetica e la morfologia. Auspicando la costituzione di un *corpus orthographicum* esaustivo, non manca di accennare ai progetti a carattere informatico che si muovono in tale direzione.
Nel settembre 1984 il LIE ha organizzato un Seminario internazionale sul tema *Grafia e interpunzione del latino nel medioevo*, cui hanno partecipato studiosi del settore tra cui G. Polara, F. Bertini, M.B. Parkes.