

## « L'influenza della formulaicità sull'analisi sintattica delle carte altomedievali toscane – l'intertestualità nel contesto diplomatico » (working paper)

**[Slide 1]** Il tema di questo intervento è la formulaicità e le sfide che la formulaicità pone all'analisi sintattica e morfosintattica del latino delle carte altomedievali in un corpus annotato. Si può dire che la questione si riduce al ruolo dell'intertestualità. Nel contesto diplomatico l'intertestualità va intesa come un concetto molto concreto: nella stesura di ciascuna carta si applicavano molte formule di diversa origine che collegano le carte ad altre carte, formando così una rete di interdipendenze tra le carte ed entro il corpus.

Nella mia ricerca di dottorato studio la morfosintassi del latino in un corpus di carte toscane provenienti dall'ottavo e nono secolo. Il corpus contiene quasi duecentomila parole, e le carte sono state pubblicate in tre serie diplomatiche, p. es. nel *Codice diplomatico longobardo* di Luigi Schiaparelli. Mi concentro sugli argomenti centrali del verbo, cioè su soggetti e oggetti, tenendo conto specialmente dell'allineamento morfosintattico. Per annotare la morfosintassi, utilizzo le interfacce online provviste dal progetto Perseus Latin Dependency Treebank. Tutte le parole del mio corpus verranno fornite di un'analisi lemmatica, morfologica e sintattica nella forma di un codice XML. Le richieste sul corpus saranno poi effettuate per mezzo del motore di ricerca Annis, accompagnato probabilmente da altre applicazioni.

Se le mie carte fossero in latino standard, tutto sarebbe più semplice. Le carte toscane altomedievali presentano comunque un latino fortemente substandard, il che pone severe esigenze all'annotazione, dato che le linee guida di annotazione del Perseus Latin Dependency Treebank, come anche quelle del LASLA, sono state disegnate per il latino standard. Ma come si sa, alcuni testi agiografici tardoantichi e medievali sono già stati annotati presso il LASLA, quindi le sfide relative alla morfologia substandard sono già familiari.

È comunque la natura formulaica delle carte che provoca difficoltà profonde sia per l'analisi morfosintattica che per il trattamento numerico dei dati in generale. 1) In primo luogo, il testo di una carta si basa (in principio) su una formula convenzionale, ma in pratica ciascun atto giuridico richiede allo scriba una certa misura di improvvisazione, perché i dettagli rilevati all'atto possano essere trasmessi per iscritto. I dettagli di questo tipo sono per esempio la descrizione della proprietà venduta, condizioni del contratto, ecc. 2) In secondo luogo, gli scribi toscani riproducevano le formule delle carte anzitutto a memoria, ma di quando in quando avranno anche utilizzato carte già esistenti come modello. In ogni caso, non ci restano tracce di formulari di nessun tipo. Quindi sarà chiaro che ciascuna carta contiene materiale testuale vario che deriva da premesse altrettanto varie e che pertanto non è possibile trattare in un modo uniforme.

Nelle parti principalmente formulaiche, la complessità e l'oscurità generale della lingua cancelleresca hanno indotto gli scribi a produrre errori contaminazionali, errori che si derivano da malintesi o da una completa mancanza di comprensione di certi passi. Le parti più libere rifletteranno, al contrario, più strettamente la lingua parlata e i problemi rilevati alla sua trascrizione per iscritto. Nell'epoca in cui sono state scritte le mie carte, il latino parlato si era certamente allontanato considerevolmente dal latino classico o standard che, ciò nonostante, si utilizzava sempre come il modello per la lingua scritta. Insomma, i problemi della struttura interna delle carte come anche quelli dell'intertestualità tra le carte sono cospicui.

Alcune ricerche quantitative hanno dimostrato che la morfologia e la sintassi si differiscono tra loro sostanzialmente nelle parti formulaiche e nelle parti più libere delle carte. Non è comunque facile dire con sicurezza quali parti delle carte longobarde o caroline siano formulaiche e quali improvvisate. Inoltre anche l'improvvisazione poteva seguire certi modelli, dato che le compravendite, le donazioni, i cambi, ecc. in generale si riproducevano seguendo procedure abbastanza convenzionali.

**[Slide 2]** Tra i passi più "liberi" si trovano per esempio elenchi di inventario, come quello del testo a), o descrizioni dei domini posseduti, come nel testo b). **[Slide 2b]** Tra i passi più formulaici sono invece per esempio l'*invocatio* e la *datatio*, come quelle del testo c), o la *sanctio* come nel testo d). Il mio metodo di tener conto di questa varietà è annotare (a mano) le sezioni libere con un tag riservato a queste parti. È vero che la formulaicità è un concetto relativo, cioè un continuo, ma raggruppare singole frasi o parole delle carte in sottoclassi secondo la loro relativa formulaicità non è possibile per i motivi ergonomici.

**[Slide 3]** Per esempio nella carta *ChLA 799* ho intenzione di taggare solo questi tre passi rossi con il tag <seg type="free">, anche se ci sono altre parole o sintagmi abbastanza o del tutto liberi. **[Slide 4]** Queste parole isolate sono comunque solo di rado importanti per un'analisi morfosintattica e perciò le posso trascurare.

**[Slide 5]** Il primo di questi passi taggati è il nocciolo della *dispositio*. In effetti, la *dispositio* è quella parte che naturalmente contiene il massimo dell'improvvisazione, in quanto luogo in cui descrivere i beni trasmessi nella compravendita ecc. Gli altri due passi sono solo riproduzioni della stessa enunciazione, ripetute quasi meccanicamente nella parte confirmatoria. Probabilmente finisco per annotare anche le sottoscrizioni con un tag separato, in caso che sono autografe, cioè scritte dai testimoni stessi e non dallo scriba. Dunque, con l'aiuto di queste annotazioni le richieste morfosintattiche possono essere delimitate (per mezzo del programma Annis) entro certe parti diplomatico-strutturali.

La questione dell'intertestualità diventa attuale quando uno comincia a definire quali passi sono davvero liberi e quali formulaici. Ho intenzione di farlo combinando i risultati di varie applicazioni dirette all'individuazione di similarità e di disparità tra testi. Voglio qui sottolineare che le mie conoscenze sulle diverse tecniche relative al trattamento numerico del testo sono deboli. Perciò sarei molto grato se voleste commentare i miei progetti o magari propormi metodi e applicazioni che non conosco ancora.

Finora ho utilizzato solo tre applicazioni. Per il momento, il programma più utile sembra il Janus Intertextuality Search Engine dell'*Electronic Manipulus Florum Project*. Janus è un programma basato sui n-grammi applicati a sequenze di caratteri, non a sequenze di parole. Anche se le richieste a n-grammi funzionano meglio con le lingue di ordine delle parole fisso (contrariamente al latino), ho usato il Janus per rintracciare tutte le varianti di una formula dentro il mio corpus. In conseguenza, l'ampiezza minima e massima della formula in questione possono essere definite e le loro forme basilari ricostruite.

**[Slide 6]** In questo esempio investigo se la frase *sicut inter nobis bono animo in placitum conuenit* (la vediamo qui, ripetuta sotto ciascuna occorrenza) – se questa frase è un'espressione formulaica. Ho fatto una richiesta sul Janus e il risultato è che la frase davvero sembra essere formulaica. Ho regolato il parametro *n*, cioè il numero dei caratteri nell'n-gramma, a 6, e il motore di ricerca mi ha dato 457 occorrenze. Comunque, solo quindici di queste sono corrette. Fortunatamente, le occorrenze corrette si trovano per la maggior parte all'inizio dell'elenco. Qui vediamo le sette prime occorrenze e sono tutte corrette, anche se alcune varianti mancano il sintagma *in placitum* e/o cominciano con un pronome relativo come *quod* o *quas*. (L'evidenziazione in giallo indica i caratteri che sono comuni alla frase richiesta e all'occorrenza.)

**[Slide 7]** In questo esempio ho cercato le possibili varianti di una frase più lunga, cioè *medietatem de casa meas infra ciuitatem cum gronda sua liuera, tam solamentum quam ligname fine grondas*. La frase fa parte della *dispositio*, quindi probabilmente è abbastanza libera. E infatti è così: non è formulaica. Solo i sintagmi come *medietatem de casa mea* e *infra ciuitatem* si ripetono altrove, ma non in rapporto tra di loro. La prima occorrenza con completa copertura è naturalmente la frase di richiesta stessa, e la seconda occorrenza, che in parte ripete la frase, deriva dalla stessa carta 23, come vediamo qui. Tutte le seguenti 603 occorrenze sono invece incorrette, il che diventa evidente dal fatto che le sequenze evidenziate in giallo sono più brevi e sparse qua e là nel testo. (Come vediamo qui.) Quindi, le richieste a n-grammi sembrano funzionare in modo soddisfacente anche per quanto riguarda il materiale formulaico di questo tipo.

Come vedete, l'ortografia dei miei testi è fortemente substandard. Per facilitare la ricerca che si concentra su caratteri o intere parole, sarebbe utile normalizzare l'ortografia dei testi prima di sottoporli all'indagine. In effetti, nel mio caso non dovrebbe essere troppo difficile appena tutte le parole del corpus sono state fornite delle rispettive analisi lemmatiche e morfologiche. Così le parole corrispondenti con ortografia e morfologia classiche possono essere ricavate dalla base dati Perseus Dynamic Lexicon.

Sarebbe comunque bello individuare le formule automaticamente. Potrebbe essere possibile con il Janus ma richiederebbe tanto lavoro che non sono capace a fare da me stesso. Uno strumento più avanzato in questo rispetto sarà la tecnica n-merge sviluppato dal progetto Multi-Version Documents (MVD). Il sistema MVD riesce a scoprire le sequenze divergenti e identiche tra due o più testi a condizione che i testi presentino al minimo un 20 % di similitudine. Comunque in molti casi, il requisito del 20 % impedisce di paragonare tra loro le intere carte; perciò le carte devono essere paragonate al livello di sottosezioni come per esempio il protocollo, la *dispositio*, l'escatocollo, ecc. Quindi anche qui devo entrare per la porta stretta invece di potere affidare al computer tutto il compito.

**[Slide 8]** Probabilmente conoscete questi esempi che ho copiato dal sito Testbed di Multi-Version Documents. Il programma rende possibile facilmente comparare le parole e caratteri differenti tra due manoscritti degli Oracoli sibillini. Le parole e caratteri rossi e blu rappresentano le letture che sono esclusive a questi due manoscritti rispettivamente. Ma purtroppo la tecnica MVD non è ancora molto accessibile, ma ci sono speranze di potere nel futuro adattare il metodo anche all'individuazione di similarità e di disparità tra due o più rappresentanti di una formula.

**[Slide 9]** Poi, ho usato anche dei diffusion map. Si tratta di un procedimento computazionale per misurare le distanze entro dataset multidimensionali con l'aiuto delle mappe diffusione che il mio collega Tuomo Sipola cerca di adattare all'uso di linguistica. Per il momento, Sipola sta processando un sottoinsieme del mio corpus per discernere eventuali tendenze che potrebbero distinguere i passi formulaici da quelli più liberi attraverso circa 100 tratti morfologici, sintattici e strutturali. **[Slide 9b]** I risultati preliminari mi sembrano permettere di raggruppare le carte nelle sottoclassi a seconda i loro pattern formulaici. Con l'aiuto di questo tipo di risultati ho potuto scoprire la parentela tra singole carte e le loro formule. **[Slide 9c]** Per esempio queste due carte, i numeri 23 e 46 del *Codice diplomatico longobardo*, sono state scritte dallo stesso scriba Ansof ma in un intervallo di dieci anni.

Dato che il mio corpus sarà completamente annotato sintatticamente, sarebbe peccato non utilizzare anche l'annotazione per esaminare le relazioni delle formule. **[Slide 10a]** Per mezzo del motore di ricerca Annis, si possono individuare strutture somiglianti all'interno di qualsiasi treebank. Questo è particolarmente utile in caso che due varianti di una formula hanno rappresentazioni lessicali diverse. **[Slide 10b]** Ad esempio la frase *sicut inter nobis bono animus in placetum conuinet* sembra essere imparentata

alla frase *quomodo inter nos libenter conuenit*, anche se con le applicazioni basate sulle sequenze di parole o di caratteri la connessione sarà rimasta inosservata. Su questa diapositiva ho marcato in rosso le strutture e le relazioni sintattiche comuni ad ambedue le frasi. Infatti, un sistema per individuare allusioni in base alle loro strutture sintattiche è ormai stato sviluppato anche presso il Perseus Latin Dependency Treebank. Purtroppo non ho avuto ancora l'occasione di provarlo.

Tutto sommato, tutti questi strumenti mi aiutano solo per quanto possono individuare formule e parti libere nei casi incerti. Eppure non trasferiscono i risultati direttamente all'annotazione. Lo devo fare a mano. Quindi per motivi pratici, l'annotazione delle parti libere sarà l'ultima fase della mia ricerca, quindi avete molto tempo per darmi dei consigli. Grazie per la vostra attenzione.

## Bibliografia

Bamman, David – Crane, Gregory. 2008. 'The Logic and Discovery of Textual Allusion', *Proceedings of the Second Workshop on Language Technology for Cultural Heritage Data*, Marrakech (LaTeCH 2008).

Cennamo, Michela. 2009. 'Argument structure and alignment variations and changes in Late Latin', *The role of semantic, pragmatic, and discourse factors in the development of case*. Ed. by Jóhanna Barðdal, Shobhana L. Chelliah. *Studies in language companion series* 108, 307–346.

*Charter Encoding Initiative*. <http://www.cei.uni-muenchen.de/>

Everett, Nick. 2003. *Literacy in Lombard Italy, c. 568–774*. Cambridge Studies in Medieval Life and Thought. Fourth Series 53.

*The Janus Intertextuality Search Engine*. <http://web.wlu.ca/history/cnighman/page13.html>

Kane, Andrew – Tompa, Frank. 2011. 'Janus: the intertextuality search engine for the electronic *Manipulus florum* project', *Literary and Linguistic Computing* 26, doi 10.1093/llc/fqr009.

Larson, Pär. 2000. 'Tra linguistica e fonti diplomatiche: quello che le carte dicono e non dicono', *La preistoria dell'italiano*. Atti della Tavola Rotonda di Linguistica Storica. A cura di József Herman e Anna Marinetti, 151–166.

*Perseus Guidelines*. Guidelines for the Syntactic Annotation of Latin Treebanks.  
<http://nlp.perseus.tufts.edu/syntax/treebank/ldt/1.5/docs/guidelines.pdf>

The Perseus Ancient Greek and Latin Dependency Treebanks. <http://nlp.perseus.tufts.edu/syntax/treebank/>

Rio, Alice. 2009. *Legal Practice and the Written Word in the Early Middle Ages: Frankish Formulae*, c. 500-1000. Cambridge Studies in Medieval Life and Thought. Fourth Series 75.

Rovai, Francesco. 2005. 'L'estensione dell'accusativo in latino tardo e medievale', *Archivio glottologico italiano* 90, 54–89.

Sabatini, Francesco. 1965. 'Esigenze di realismo e dislocazione morfologica in testi preromanzesi', *Rivista di Cultura Classica e Medievale* 7, 972–998.

Sanga, Glauco – Baggio, Serenella. 1995. '*Sul volgare in età longobarda*', Italia settentrionale: crocevia di idiomi romanzesi. A cura di E. Banfi, G. Bonfadini, P. Cordin, M. Iliescu, 247–260.

Schiaparelli, Luigi. 1933. 'Note diplomatiche sulle carte longobarde, II: Tracce di antichi formulari nelle carte longobarde', *Archivio storico italiano* 19, 3–34.

Schmidt, Desmond – Fiormonte, Domenico. 2010. 'Documenti multiversione: una soluzione per gli artefatti testuali del patrimonio culturale / Multi-version documents: a digitisation solution for textual cultural heritage artefacts', *Intelligenza artificiale* 4, 56–61.

Sipola, Tuomo – Juvonen, Antti – Lehtonen, Joel. [forthcoming 2011] 'Anomaly detection from network logs using diffusion maps', *EANN/AIAI 2011*, Part I, IFIP AICT 363, 172–181. Ed. by L. Iliadis, C. Jayne.