FRANCESCO ROVAI

Mutamento ed Entropia. Un approccio informazionale ai processi di grammaticalizzazione

1. *Introduzione*[*]

Il presente contributo intende proporre un modello di analisi linguistica che consenta di tradurre in un valore numerico quantificabile categorie qualitative e intrinsecamente scalari come quelle di grammaticalizzazione e lessicalizzazione. La funzione dell'entropia, mutuata dalla Teoria dell'Informazione, appare in grado di assolvere a tale compito, mediante l'istituzione di un nesso esplicito tra una nozione intuitiva come l'*informazione* e un dato empiricamente rilevabile come la *frequenza*.

La prima parte del lavoro (§§ 2-4) cercherà dunque di mostrare che esistono fondati presupposti per l'applicazione di un modello informazionale allo studio delle lingue naturali. Dopo aver illustrato i concetti basilari di Teoria dell'Informazione e i principi che governano qualsiasi sistema di comunicazione (§ 2), attraverso essi saranno interpretati sia alcune sistematiche tendenze e regolarità sincroniche (§ 3), che alcuni processi diacronici di mutamento linguistico (§ 4).

Dopo aver così definito le basi teoriche del modello proposto, la seconda parte del lavoro (§§ 5-7) vuole tentare una prima e parziale verifica della validità delle sue previsioni, attraverso una rilettura in termini informazionali di un ben noto fenomeno di mutamento linguistico, l'ausiliarizzazione di *habere* nel passaggio dal latino all'italiano e la conseguente grammaticalizzazione del costrutto perfettivo *avere + participio passato*.

2. *Elementi di Teoria dell'Informazione*

2.1. *Entropia puntuale*

Nella Teoria dell'Informazione (Shannon, 1948), i messaggi codificati da un sistema di comunicazione sono rappresentati come eventi mutuamente esclusivi generati da un sistema aleatorio, costituito da un insieme finito di esiti possibili $\{e_1, e_2, \in, e_n\}$ e da un insieme di probabilità assegnate alla realizzazione di

* Desidero ringraziare il Dott. Alessandro Lenci, la Prof.ssa Giovanna Marotta e la Dott.ssa Domenica Romagno, i quali hanno seguito lo sviluppo di questo lavoro contribuendo a migliorarlo con critiche ed osservazioni.

Ricevuto: ottobre 2007
Accettato: giugno 2008

ciascuno di essi $\{p_{(e1)}, p_{(e2)}, \in, p_{(en)}\}$. Sulla base di queste premesse, l'*informazione puntuale*, ossia la quantità di informazione apportata da ciascun evento, viene stimata in funzione della riduzione dell'incertezza che la sua realizzazione opera all'interno dei possibili esiti del sistema. Conseguentemente, più un evento è frequente e prevedibile, minore sarà l'informazione che esso ci fornisce, più esso è raro e imprevedibile, maggiore sarà il suo contenuto informativo[1].

La funzione dell'entropia puntuale *h(e)*, così come formulata da Shannon (1), esplicita il rapporto inverso che intercorre tra la probabilità ed il valore informativo di un messaggio, quantificando al contempo il numero minimo di *bits* necessari trasmetterlo[2].

$$(1)\ h(e) = - \log_2 p(e)$$

A livello della codifica del messaggio, la prima conseguenza della proporzionalità inversa tra entropia e probabilità, è che i messaggi più frequenti e meno informativi possono essere codificati con meno *bits* rispetto ai messaggi più rari e più informativi.

2.2. *Entropia del sistema*

Quanto visto sino ad ora a proposito dell'entropia puntuale di ciascun evento, si riflette inoltre sull'organizzazione e l'efficienza del sistema di comunicazione, la cui entropia globale *H(E)* è definita dal valore medio delle entropie puntuali dei singoli messaggi. L'equazione individuata da Shannon (1948) per definire tale funzione è la seguente (2):

$$(2)\ H(E) = - \Sigma\, p_i \log_2 p_i$$

[1] Nel seguito del lavoro, sarà tenuta presente una definizione frequentista di probabilità, secondo cui la probabilità *p(e)* di un evento *e* può essere stimata sulla base della sua frequenza relativa *f(e)* all'interno di un campione significativo di esperimenti.

[2] Il segno "-" esprime la proporzionalità inversa tra l'entropia di un messaggio *h(e)* e la sua probabilità *p(e)*: la prima diminuisce all'aumentare della seconda. Il ricorso al logaritmo in base 2 è invece motivato dall'unità di misura dell'entropia informazionale, il *bit*. Si tenga infatti presente che la Teoria dell'Informazione nasce e si sviluppa nell'ambito dei sistemi informatici, i cui messaggi sono codificati attraverso un sistema binario. Dal momento che ogni *bit* (cifra binaria) può assumere solo due valori (0 e 1), il numero (*w*) delle possibili combinazioni delle due cifre binarie 0 e 1 è dato dalla funzione esponenziale $w = 2^n$, in cui *n* è il numero di *bits* impiegati: 1 *bit* ammette perciò solo 2 combinazioni (0; 1), 2 *bits* ne ammettono 4 (00; 01; 10; 11), 3 *bits* 8 (000; 001; 010; 011; 100; 101; 110; 111), e così via. Mediante il ricorso al logaritmo, funzione inversa dell'esponenziale, l'entropia esprime un'operazione inversa rispetto alla precedente ($n = log_2 w$), stimando quanti *n bits* sono necessari per poter ottenere un numero *w* di combinazioni.

Occorre innanzi tutto rilevare che, poiché l'entropia *H(E)* di un sistema di comunicazione è innanzi tutto una misura dell'incertezza media di un sistema aleatorio, il suo valore sarà massimo nel caso in cui tutti gli eventi siano equiprobabili, ossia totalmente imprevedibili: in questo caso il sistema si definisce caotico. Al contrario, data una classe di eventi non equiprobabili, in cui gli elementi più frequenti sono codificati con meno *bits* rispetto a quelli più rari, il valore dell'entropia del sistema diminuisce. Ciò significa che l'esistenza di elementi molto probabili, frequenti e ripetitivi (definiti, in termini informazionali, *ridondanti*)[3], tra i quali si instaurano relazioni e corrispondenze non casuali, comporta una riduzione dell'incertezza media del sistema, ossia un aumento della prevedibilità del suo comportamento: il sistema, in questo caso, si definisce organizzato.

È dunque chiaro che, più sono i vincoli che impongono delle restrizioni sui possibili esiti del sistema, più aumenta la sua organizzazione, più il suo comportamento diventa prevedibile e quindi l'incertezza, l'entropia globale, diminuisce. Da questo punto di vista, *H(E)* costituisce un indice inverso della quantità di organizzazione e di struttura presente in un sistema.

In considerazione di quanto appena rilevato, è opportuno sottolineare che a fianco dell'informazione *puntuale* contenuta nei singoli messaggi, esiste un tipo di informazione che può essere definita *strutturale*, legata alla quantità di organizzazione del sistema. L'esistenza di vincoli e schemi regolari e ripetitivi, infatti, apporta anch'essa informazione in quanto determina una riduzione dell'incertezza sui possibili esiti del sistema: la natura organizzata dell'informazione è esattamente ciò che la distingue dalla casualità del rumore.

3. *Principi informazionali in sincronia*

Il presente paragrafo intende mettere in evidenza come i principi informazionali precedentemente delineati operino anche all'interno delle lingue naturali, qui considerate in quanto sistemi di comunicazione. Ciò che interessa mostrare, attraverso l'analisi di un testo reale (il *Decameron*), è che anche all'interno di esse è ravvisabile una relazione tra la frequenza di un elemento, il suo contenuto informativo (§ 3.1), e la sua codifica (§ 3.2).

3.1. *Frequenza e contenuto informativo*

Così come nei sistemi di comunicazione l'informazione contenuta in un

[3] Nel seguito del lavoro, i termini *ridondante* e *ridondanza* sono da intendersi unicamente in questa accezione.

messaggio è inversamente proporzionale alla sua probabilità, anche nelle lingue naturali sembra sussistere una proporzionalità inversa fra il contenuto informativo di tipo semantico-lessicale e la frequenza di un elemento. Secondo quanto rilevato da Dahl (2003: 159), con esplicito riferimento alla Teoria dell'Informazione, «[t]here is a direct link between frequency and informational value – indeed, in information theory they are simply two sides of the same coin».

Tali affermazioni possono trovare una conferma immediata nell'analisi statistica di un testo. La Tabella 1 riporta, in ordine decrescente (*rango*), i 30 tipi lessicali (*types*) più frequenti all'interno del *Decameron* e il relativo numero di attestazioni (*tokens*)[4]:

rango	*type*	*tokens*	rango	*type*	*tokens*	rango	*type*	*tokens*
1.	e	12953	11.	con	2340	21.	egli	1502
2.	che	11156	12.	l(o)	2235	22.	gli	1498
3.	di	6326	13.	si	2073	23.	ciò	1477
4.	la	5651	14.	se	1935	24.	d(i)	1400
5.	a	5236	15.	come	1893	25.	era	1358
6.	il	5024	16.	da	1767	26.	una	1357
7.	per	4139	17.	le	1766	27.	ma	1329
8.	non	4039	18.	quale	1763	28.	del	1312
9.	in	3222	19.	più	1658	29.	disse	1273
10.	io	2727	20.	un	1523	30.	della	1257

Tabella 1

È facile osservare che si tratta unicamente di parole funzionali e semanticamente vuote: congiunzioni (*e*, *che*, €), preposizioni (*di*, *a*, *in*, *da*), pronomi (tonici: *io*, *egli*; clitici: *ci*, *gli*), avverbi (*non*, *più*), articoli (*la*, *il*, €). Uno dei due verbi è un ausiliare (*era*), l'altro, in un testo narrativo come il *Decameron*, è funzionale all'introduzione dei discorsi diretti (*disse*)[5]. Occorre inoltre notare che sebbene essi rappresentino appena lo 0,17% del vocabolario del testo (30 tipi lessicali su 17548), il totale delle loro occorrenze (93189 *tokens*) copre da

[4] Lo spoglio è stato effettuato su un testo del *Decameron* in formato elettronico ".txt" attraverso un programma di estrazione delle concordanze (*Simple Concordance Program*, scaricabile alla pagina *www.textworld.com/scp*) utilizzato anche per le analisi presentate nel seguito del lavoro.

[5] Trattandosi di un semplice rilevamento delle frequenze su un testo solo tokenizzato e non annotato a livello morfosintattico, la forma *che* assomma le frequenze della congiunzione e del pronome. Una distinzione peraltro secondaria ai fini di quanto si intende mostrare.

solo un terzo (33,96%) dei 274382 *tokens* che costituiscono l'intero testo! Le parole lessicalmente piene sono invece in larga misura eventi rari e poco prevedibili, e la maggior parte del vocabolario del testo (87,19%) è costituito da parole che ricorrono al massimo 10 volte e coprono appena il 12,53% dei *tokens*. Si tratta della tipica distribuzione zipfiana (Zipf, 1929) a cui si uniformano tutti i testi reali: poche parole con frequenze elevate, molte parole rare, moltissimi *hapax*.

Tutto questo non significa comunque che gli elementi funzionali non apportino alcun tipo di informazione[6]: piuttosto, riprendendo la distinzione formulata alla fine del § 2.2., essi veicolano un'informazione di tipo non *puntuale* ma *strutturale*. Congiunzioni, preposizioni, articoli, ecc€ costituiscono infatti un insieme di elementi frequenti e ripetitivi che permettono di individuare le strutture presenti in un testo, la sua organizzazione. È ovvio che in questa analisi esemplificativa, condotta su un testo tokenizzato ma non annotato morfosintatticamente, è possibile cogliere solo un aspetto superficiale di tale organizzazione, ossia quello relativo all'articolazione del periodo (dato dalle congiunzioni) e alla struttura dei sintagmi nominali (dato da preposizioni e articoli).

Gran parte dell'informazione *strutturale* è invece veicolata, almeno in una lingua flessiva come l'italiano, dalla morfologia di accordo, anch'essa facente parte di quegli elementi informazionalmente ridondanti che concorrono a definire l'organizzazione di un testo. Si veda ad esempio una frase come la seguente (3):

(3) molte strade sono state distrutte

In essa l'informazione "genere: femminile; numero: plurale", marcata su tutti gli elementi nominali, è certamente ridondante per quanto riguarda l'apporto di nuovo contenuto informativo *puntuale*. Questo dispendio di codice ha però una duplice funzione. Da un lato, segnala che tutti gli elementi che veicolano tale informazione appartengono ad un medesimo costituente sintattico, dall'altro, ricopre un ruolo fondamentale nelle dinamiche funzionali della lingua intesa come sistema di comunicazione. Un destinatario che, a causa di un canale rumoroso, ricevesse questo messaggio come (4):

(4) molte stradX sono statX distrutte

non avrebbe alcuna difficoltà a ricostruire il messaggio originario reintegrando l'informazione mancante sulla base di una elevata prevedibilità data dal contesto.

[6] Questo lascerebbe intendere, ad esempio, la seguente affermazione di DAHL (2001: 477): «Grammatical markers tend to carry little or no information that is relevant to the message».

3.2. *Frequenza ed estensione del segnale*

Come sottolineato al § 2.1, all'interno di un sistema di comunicazione, la frequenza di un messaggio è rilevante non solo per stimarne il contenuto informativo, ma anche per determinare la quantità di segnale necessaria alla sua codifica. Anche questo secondo aspetto trova riscontro all'interno delle lingue naturali, in cui le parole più frequenti (e perciò meno informative) tendono ad essere codificate con meno caratteri rispetto a quelle meno frequenti (e perciò più informative).

Tutto ciò è stato verificato attraverso una ulteriore analisi condotta sul *Decameron* e sintetizzata dal grafico di Figura 1, che rappresenta l'andamento del rapporto tra la frequenza di una parola (riportata sull'asse *y*) e la sua lunghezza in caratteri (riportata sull'asse *x*), e da cui emerge con chiarezza la proporzionalità inversa che intercorre tra i due parametri.

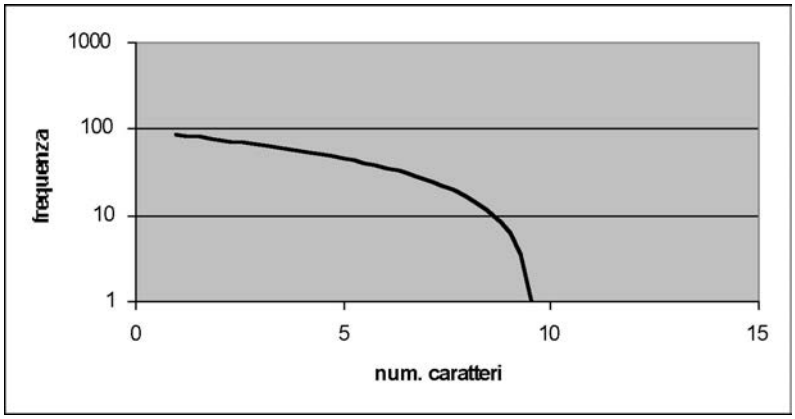

Figura 1

3.3. *Quantità di informazione e quantità di segnale: a proposito del principio di iconismo*

In chiusura di questo paragrafo, in cui si è cercato di mostrare come le lingue naturali rispondano a principi di tipo informazionale, appare necessario soffermarsi sulla corrispondenza fra contenuto informativo ed estensione del segnale, già rilevata in Zipf (1929) e recentemente ridiscussa in Fenk-Oczlon (2001).

Traendo le conclusioni di quanto emerso dai §§ 3.1-3.2, risulta infatti che gli elementi con un contenuto semantico-lessicale ridotto tendono ad avere una codifica più breve rispetto a quelli semanticamente pieni. Tutto ciò, almeno in

apparenza, sembra confermare le previsioni di un principio iconico, secondo cui è rilevabile una proporzionalità diretta fra complessità del significato e complessità del significante[7]. In realtà, le motivazioni di tale rapporto, che certamente esiste, non sono da ricercarsi in una relazione di causa-effetto fra i due piani del significato e del significante. Questa corrispondenza è piuttosto un prodotto epifenomenico dei principi informazionali che governano il linguaggio in quanto sistema di comunicazione. Tanto un maggiore o minore contenuto informativo quanto una maggiore o minore quantità di segnale, sono entrambi conseguenze, rispettivamente, di una minore o maggiore probabilità di un certo elemento.

Tutto ciò sembra avvalorare la spiegazione a più riprese proposta da Haspelmath (1999, 2006, 2008), che motiva questi apparenti casi di iconismo come una conseguenza immediata di effetti di frequenza. Occorre però tenere presente che la frequenza si rivela una categoria esplicativa pertinente solo nel momento in cui essa non venga considerata di per sé, ma sia interpretata in funzione della maggiore prevedibilità che essa comporta nella realizzazione di un messaggio.

4. *Principi informazionali in diacronia*

Di seguito cercheremo di fornire un'interpretazione su base informazionale di alcuni processi diacronici: erosione fonetica (§ 4.1), mutamento semantico (§ 4.2) e grammaticalizzazione (§ 4.3). Questa rilettura è resa possibile dallo stretto legame tra entropia e frequenza, un parametro rilevante in numerosi fenomeni di mutamento linguistico. L'adozione di questo quadro interpretativo induce inoltre alcune considerazioni relative alle conseguenze di tali fenomeni sulla lingua in quanto sistema di comunicazione, esposte al § 4.4.

[7] Il principio di iconismo gode di particolare fortuna nell'ambito della Morfologia Naturale, come attestano i seguenti passi: HAIMAN (1980: 528): «increased morphological complexity is an icon of increased semantic complexity. Thus, generally speaking, the positive, comparative, and superlative degrees of adjectives show a gradual increase in the number of phonemes»; MAYERTHALER (1980: 25): «Was semantisch 'mehr' ist, sollte auch konstruktionell 'mehr' sein». Tuttavia, esso è ampiamente condiviso anche in lavori non strettamente collegati a tale paradigma teorico. Si vedano, a questo proposito, LEHMANN (1995 [1982]: 122): «There tends to be a correspondence between the size, or complexity, of the significans and that of the significatum»; LANGACKER (2000 [1992]: 77): «It is worth noting an iconicity between *of*'s phonological value and the meaning ascribed to it»; ed infine, il *Quantity Principle* formulato in GIVÓN (1991: 89): «a larger chunk of information [€] will be given a larger chunk of code».

4.1. *Frequenza e riduzione fonetica*

I primi studi sugli effetti della frequenza in diacronia sono stati condotti soprattutto in relazione ai mutamenti fonetici che ne conseguono: già Schuchardt (1972 [1885]) aveva infatti evidenziato che le parole più frequenti sono quelle più esposte all'erosione fonetica. Tale intuizione è stata poi verificata su una più ampia base di dati da Zipf (1929, 1935, 1938) nell'ambito della *Dynamic Philology*, il primo tentativo di connettere metodi statistici e linguistica: con la formulazione della *Law of Abbreviation* (Zipf, 1935) vengono così rilevate corrispondenze regolari tra frequenza e peso fonetico delle parole, sulla base di un semplice conteggio delle loro frequenze assolute. Più recente e articolata è invece la così detta *Probabilistic Reduction Hypothesis*, proposta avanzata da Jurafsky *et al.* (1998) e Jurafsky *et al.* (2001), secondo cui quanto più una parola è probabile, tanto più la sua forma tenderà ad essere ridotta. Rispetto al modello di Zipf, dunque, la riduzione fonetica di un elemento è messa in relazione non solo a frequenze assolute elevate, ma anche alla sua predicibilità sulla base del contesto. Gli elementi più prevedibili, perché in stretta relazione con elementi vicini, subiscono in misura maggiore fenomeni di riduzione del timbro vocalico e della durata.

Tale ipotesi è confermata anche dallo studio condotto da Marotta (2001) sulla spirantizzazione delle occlusive intervocaliche nel parlato di Pisa. Rispetto all'incidenza media di tale fenomeno che si registra all'interno di elementi lessicali, la lenizione delle occlusive sorde si manifesta infatti in percentuali più elevate nel caso di congiunzioni, preposizioni, pronomi personali clitici, e nel morfema -*Vto* del participio passato[8].

È significativo e coerente con i principi informazionali visti in precedenza che questi processi di riduzione fonetica agiscano in misura maggiore sugli elementi funzionali rispetto a quelli lessicali. È già stato infatti rilevato che all'interno delle lingue naturali, gli elementi grammaticali (in particolare la morfologia flessiva) sono più probabili rispetto a quelli lessicali non solo perché in assoluto più frequenti, ma anche perché inseriti in schemi e *patterns* ricorrenti che consentono di predirli e interpretarli correttamente, anche in presenza di un segnale morfofonetico ridotto (§ 3.1). Viceversa gli elementi lessicali, informativi proprio perché altamente imprevedibili, richiedono un maggiore utilizzo di codice per consentirne la corretta identificazione: in questo caso una riduzione del segnale comporterebbe la perdita dell'informazione stessa[9].

[8] L'indebolimento del morfema verbale di participio passato trova una corrispondenza nell'inglese americano, in cui la desinenza di passato/participio passato -*ed* è frequentemente ridotta o cancellata (BYBEE, 2001).

[9] Si noti come tali risultati sostanzino alcune intuizioni che furono già di MEILLET

4.2. *Frequenza e desemantizzazione*

Gli effetti di frequenza agiscono anche sul contenuto semantico degli elementi lessicali. Il nesso tra frequenza di un costrutto e desemantizzazione è noto almeno a partire da Meillet (1958 [1912])[10], ed è il risultato di quelli che Dahl (2001, 2003) definisce *inflationary effects*, con riferimento ai processi inflazionistici ricorrenti in altri ambiti delle convenzioni culturali umane. L'inflazione è infatti un processo a cui sono esposti tutti quegli elementi che assumono un valore solo sulla base di convenzioni arbitrarie, e si manifesta come un fenomeno di "invisible-hand", ossia come prodotto non intenzionale di azioni intenzionali, risultante dal conflitto fra interessi di singoli agenti a breve termine ed il funzionamento del sistema a lungo termine. È plausibile ritenere che ciò che innesca il processo sia la tendenza del singolo parlante a massimizzare l'effetto retorico del proprio enunciato (Keller, 1994: *Maxims of Action*): se alcuni individui utilizzano espressioni rafforzate, altri parlanti si adegueranno per non rimanere svantaggiati nello scambio retorico. Questo causa l'incremento della frequenza di una certa espressione e, sul lungo periodo, la sua svalutazione semantica, innescando processi di grammaticalizzazione quali il "Ciclo di Jespersen", come nel caso del francese *pas* o dell'italiano *mica*, ormai elementi funzionali privi di ogni rapporto con il significato di "passo" o "briciola". Dahl (2003: 124) trae un significativo esempio di tale processo dal cinese mandarino, in cui l'intensificatore *hěn* "molto", divenuto un modificatore obbligatorio di predicati scalari come *kuài* "veloce" o *dà* "grande", ha perso il valore semantico originario. La seguente frase (5):

(5)	*Zhì*	*suǒ*	*fángzi*	*hěn*	*dà*
	questa	CLASS	casa	molto	grande

significa semplicemente "questa casa è grande". Il corrispettivo di "questa casa è molto grande", richiederebbe l'introduzione di un altro intensificatore, *fēicháng*, il cui significato proprio è "estremamente".

L'interpretazione "inflazionistica" consente di sottrarsi ad un ragionamento a rischio di circolarità ("gli elementi più frequenti hanno una semantica più ampia, poiché in virtù di essa possono ricorrere in un maggior numero di contesti"), istituendo invece un rapporto di causalità fra scelte pragmatiche contin-

(1958 [1912]: 138): «L'histoire des langues montre que, par suite, les mots accessoires ont des traitements phonétiques aberrants [€] leurs éléments constituants, étant abrégés et faiblement articulés, sont exposés à s'affaiblir ou a disparaître dans des cas où les éléments d'un mot principal subsistent intacts».

[10] «A chaque fois qu'un élément linguistique est employé, sa valeur expressive diminuì et la répétition en devient plus aisée» (MEILLET, 1958 [1912]: 135).

genti, aumento della frequenza, e conseguente perdita di contenuto informativo semantico-lessicale (Lehmann, 1985; Haspelmath, 1999; Dahl, 2001).

4.3. *Frequenza e grammaticalizzazione*

I fenomeni di grammaticalizzazione sono il risultato dell'interazione e della compresenza dei due processi discussi in precedenza[11]. Nel passaggio dal lessico alla grammatica, schematizzato dal seguente *continuum* di grammaticalità (Hopper e Traugott, 1993), un elemento subisce al tempo stesso una riduzione del proprio contenuto semantico ed un processo di erosione fonetica[12]:

content item > *grammatical word* > *clitic* > *inflectional affix*

L'avanzamento lungo tale *continuum*, riletto in termini informazionali rappresenta la progressiva perdita di informazione *puntuale* di tipo semantico-lessicale dovuta ad effetti inflazionistici, a cui si accompagna una riduzione del

[11] I singoli processi non sono infatti di per sé sufficienti a definire la grammaticalizzazione. Da un lato, l'erosione fonetica è infatti onnipresente nei processi di mutamento linguistico, dall'altro, la desemantizzazione può ricorrere al di fuori dei processi di grammaticalizzazione (LEHMANN, 1995 [1982]). Nel passaggio dal latino tardo all'italiano, i vocaboli *adripare*, *caballu*(*m*), *casa*(*m*), hanno perso una semantica specifica, senza per questo divenire elementi funzionali (sebbene in italiano antico *casa* avesse assunto lo stesso valore grammaticale del francese *chéz*; v. LONGOBARDI, 1995). Inoltre, a differenza della grammaticalizzazione, di cui costituisce un presupposto necessario ma non sufficiente, il mutamento semantico può essere bidirezionale: opposto ai precedenti, è il caso di *mulier* che, perso il significato originario di "donna", ha assunto quello specifico di "moglie".

[12] L'andamento parallelo di tali processi è rilevato in tutti i principali studi sui fenomeni di grammaticalizzazione. Si vedano tra gli altri MEILLET (1958 [1912]: 139): «L'affaiblissement du sens et l'affaiblissement de la forme des mots accessoires vont de pair; quand l'un et l'autre sont assez avancés, le mot accessoire peut finir par ne plus être qu'un élément privé de sens propre, joint à un mot principal pour en marquer le rôle grammatical. Le changement d'un mot en élément grammatical est accompli»; LEHMANN (1995 [1982]: 126): «Decrease in the semantic integrity of a sign is desemanticization; decrease in the phonological integrity is phonological attrition. The parallelism between these two processes has been emphasized repeatedly in the literature». Un'ipotesi sulla riduzione parallela del contenuto semantico e del segnale morfofonetico (*Parallel Reduction Hypothesis*) è esplicitamente formulata in BYBEE e PAGLIUCA (1985: 59-60): «[i]t is often observed that grammatical meaning develops out of lexical meaning by a process of generalization or weakening of semantic content [...]. It can be further hypothesized that [€] this semantic change is paralleled over a long period of time by phonetic erosion»; e riproposta in BYBEE *et al.* (1994: 6): «Parallel to semantic reduction, phonological reduction continues to take place throughout the life of a gram».

segnale morfofonetico: entrambi gli aspetti sono conseguenti all'incremento della frequenza, ossia alla maggiore predicibilità, di un elemento[13]. D'altra parte, il legame tra grammaticalizzazione e incremento dei contesti d'uso di un elemento era noto fin da Schlegel (1971 [1818])[14], e rientra nella classica definizione proposta da Kuryłowicz (1975 [1965])[15].

Un modello informazionale consente di esplicitare queste corrispondenze, da tempo note in linguistica, e di tradurle in un indice numerico che consenta una stima del contenuto informativo di un elemento. In particolare, la funzione dell'entropia è in grado di cogliere la duplice dimensione dei processi di grammaticalizzazione (svuotamento semantico ed erosione fonetica), grazie al rapporto inverso istituito fra la probabilità/frequenza del messaggio, il suo contenuto informativo, e l'estensione del segnale. Su queste stesse basi, essa appare inoltre in grado di quantificare numericamente il livello di grammaticalizzazione di un certo elemento, assegnando ad esso un valore su una scala di numeri reali, a partire da una stima della sua probabilità.

4.4. *Grammaticalizzazione e organizzazione del sistema*

In base a quanto visto in § 2.2, si consideri la lingua come sistema di comunicazione non caotico avente un livello medio di entropia che ne definisce sia la quantità di informazione globale che l'organizzazione interna. Si consideri inoltre l'entropia *puntuale* come valore indicante non solo la quantità di informazione di un elemento, ma anche lo sforzo procedurale necessario a produrlo/interpretarlo (Fenk-Oczlon, 2001; Goldsmith, 2001). Un sistema linguistico efficiente risulta dunque dalla combinazione fra elementi funzionali molto frequenti, poco informativi e che contribuiscono all'organizzazione del sistema, ed elementi semantico-lessicali rari e molto informativi. In conseguenza del rapporto inverso istituito tra entropia e frequenza/probabilità, è dunque

[13] Anche in questo caso, il fatto che i morfemi grammaticali siano universalmente più brevi di quelli lessicali è la conseguenza sincronica di queste dinamiche e non la risposta ad un principio di iconismo. Per quanto riguarda la spiegazione in termini frequentisti di vari fenomeni di grammaticalizzazione, si rimanda ai contributi raccolti nel volume di BYBEE e HOPPER (2001).

[14] SCHLEGEL (1971 [1818]: 28): «On dépouille certains mots de leur énergie significative, on ne leur laisse qu'un valeur nominale, pour leur donner un cours plus général et les faire entrer dans la partie élémentaire de la langue».

[15] KURYŁOWICZ (1975 [1965]: 52): «Grammaticalization consists in the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status, e.g. from a derivative formant to an inflectional one».

prevedibile che i primi saranno caratterizzati da valori di entropia bassi, mentre i secondi da valori più elevati. L'efficienza del sistema, a sua volta, risulta dal mantenimento di un livello medio di entropia. Un valore troppo basso indicherebbe infatti un eccesso di ridondanza, con conseguente perdita di informazione puntuale semantico-lessicale. Viceversa, l'aumento dell'entropia porterebbe il sistema verso il caos: il valore sarebbe massimo nel caso limite in cui tutti gli eventi fossero equiprobabili e quindi totalmente imprevedibili.

La realtà storica delle lingue mostra che il sistema, sebbene in continua evoluzione, si presenta come una successione di stati sincronici ordinati, strutturati e non caotici. Proprio la grammaticalizzazione si configura come un processo volto a ridurre l'entropia del sistema: gli esiti sono infatti elementi funzionalmente ridondanti, prevedibili e frequenti, che consentono di individuare regolarità strutturali e quindi di imporre un ordine sul sistema stesso. Riprendendo quanto già sottolineato al § 2.2, sia chiaro che affermare che la grammaticalizzazione produce tali elementi non equivale a dire che il sistema perda progressivamente informazione. Piuttosto, essa viene ridistribuita al suo interno, passando da una informazione di tipo *puntuale* ad una di tipo *strutturale*.

Da questo punto di vista, trova conferma l'affermazione di Haspelmath (2008) secondo cui il mutamento diacronico costituisce il legame fra i *patterns* d'uso di una lingua e la sua grammatica, in quanto le strutture di quest'ultima vengono a configurarsi come prodotto dell'individuazione e dell'astrazione delle regolarità distribuzionali presenti nei testi. Tutto ciò presuppone ovviamente una visione della grammatica come struttura emergente, instabile e continuamente rinegoziabile, secondo le proposte di Hopper (1987, 1988, 1998) e Bybee e Hopper (2001)[16].

5. *Per una prima verifica del modello: la grammaticalizzazione di* habere

Sulla base dei presupposti teorici precedentemente esposti, in questa seconda parte del lavoro cercheremo di verificare se effettivamente il valore dell'entropia puntuale di un certo elemento (più basso per gli elementi funzionali, più elevato per quelli lessicali) può costituire un valido indicatore del suo livello di grammaticalizzazione o lessicalizzazione. A tale scopo, come già an-

[16] A conclusioni simili approdano anche modelli del mutamento morfologico basati sul connessionismo lessicale (BYBEE e SLOBIN, 1982; BYBEE e MODER, 1983) secondo cui una regola produttiva si genera attraverso la progressiva estensione di uno schema, da intendersi come l'insieme dei tratti formali comuni ad una categoria di elementi lessicali.

ticipato nell'introduzione, il modello proposto sarà applicato ad un ben noto fenomeno di mutamento linguistico occorso nella transizione dal latino alle lingue romanze, l'ausiliarizzazione di *habere* e la grammaticalizzazione del costrutto perfettivo *avere + participio passato*.

Tale sviluppo si articola in diverse tappe attraverso le quali il verbo *habere* perde progressivamente la propria autonomia semantica e l'originario valore di "possedere" finendo con l'assumere il valore funzionale di marca tempo/aspettuale. Il processo si avvia a partire da costrutti transitivi in cui una forma di participio perfetto concorda con l'oggetto diretto di *habere*: il costrutto perde inizialmente l'interpretazione possessiva, l'accordo tra oggetto diretto e participio perfetto diviene facoltativo, ed infine, divenuta opzionale anche la presenza dell'oggetto diretto, il costrutto si estende dai predicati transitivi a quelli intransitivi[17].

L'aspetto rilevante ai fini dalla nostra analisi è che un originario elemento lessicale che ricorre in costrutti dotati di un proprio valore semantico, in seguito ad un processo diacronico diviene un elemento funzionale all'interno di un costrutto grammaticalizzato. Se la funzione dell'entropia è realmente in grado di stimare il livello di grammaticalizzazione o di lessicalizzazione di un certo elemento o di un costrutto, allora le seguenti previsioni devono essere confermate:

1) in latino, rispetto ad *esse*, che ricopre la funzione di ausiliare, e ad *esse + PP*, già grammaticalizzato come costrutto perfettivo passivo, *habere* e *habere + PP* devono avere valori di entropia più alti;
2) all'interno di un testo italiano (il *Decameron*), l'entropia di *essere* e di *avere*, entrambi ausiliari, e di *essere + PP* ed *avere + PP*, entrambi costrutti grammaticalizzati, devono risultare equivalenti;
3) in prospettiva diacronica, rispetto al costrutto latino *habere + PP*, il costrutto *avere + PP* deve mostrare un significativo spostamento verso valori di entropia propri di elementi grammaticalizzati.

6. *Continuum lessico-grammatica*

Onde poter verificare se un certo elemento o costrutto siano più o meno lessicalizzati o grammaticalizzati, occorre innanzi tutto identificare dei valori prototipici a partire dai quali definire la loro posizione relativa su un *continuum*

[17] Non è questa la sede in cui riepilogare i dettagli che caratterizzano le diverse tappe dell'evoluzione o ridiscutere la altre ristrutturazioni che si accompagnano alla grammaticalizzazione del costrutto. Per una più specifica trattazione di tali argomenti da diverse prospettive teoriche si vedano RAMAT (1984); PINKSTER (1987); LA FAUCI (1988); HEINE (1993, 1997); LOPORCARO (1998).

semantico-funzionale. La struttura prototipico-scalare di quest'ultimo, data per implicita in tutti gli approcci funzionalisti, può infatti essere espressa quantitativamente da un modello informazionale attraverso la funzione dell'entropia. Essa è infatti in grado di assegnare un valore su una scala di numeri reali a qualsiasi elemento linguistico, stimandone la probabilità *p(e)* in base alla sua frequenza relativa *f(e)* all'interno di un *corpus* ed applicando a tale valore la formula vista in (1).

I valori che caratterizzano i due poli opposti del *continuum* sono quindi stati individuati, sia per i testi latini (§ 6.1) che per il *Decameron* (§ 6.2), secondo le modalità di seguito esposte[18].

6.1. *Continuum lessico-grammatica: testi latini*

Coerentemente con quanto previsto da una distribuzione di tipo zipfiano, il polo lessicale è costituito dagli elementi che ricorrono da 1 a 10 volte[19]. Ovviamente, a parità di estensione dei testi esaminati, i valori di entropia degli *hapax* (16,73 *bits*) e degli elementi che ricorrono al massimo 10 volte (13,41 *bits*), risultano identici in tutti i testi esaminati.

Il polo grammaticale è costituito invece dagli elementi più probabili. È stato dunque calcolato il valore dell'entropia puntuale di elementi che in tutti i testi latini esaminati rientrano tra le 20 parole più frequenti. La seguente Tabella 2

[18] Per quanto riguarda il latino, l'analisi è stata condotta su un *corpus* di testi differenti per cronologia e genere letterario, che coprono un arco temporale di circa sette secoli: le commedie di Plauto (ca. 210-184 a.C.), le *Epistulae ad familiares* di Cicerone (62-43 a.C.), l'opera storica di Livio (27/5 a.C.-17 d.C.), l'*Historia Augusta* (390-420 d.C.), e il *Digesto* di Giustiniano (533 d.C.). Il *corpus* di riferimento (tratto dalla *Bibliotheca Teubneriana Latina*) è costituito dai seguenti testi: Plauto: *Amphitruo*, *Asinaria*, *Aulularia*, *Bacchides*, *Captivi*, *Casina*, *Cistellaria*, *Curculio*, *Epidicus*, *Menaechmi*, *Mercator*, *Miles gloriosus*, *Mostellaria*, *Persa* (1-737), per un totale di 108673 *tokens*; Cicerone: *Epistulae ad familiares* (I,1,1-XV,4,6), per un totale di 108642 *tokens*; Livio: *Ab urbe condita* (I,1,1-VII,24,8), per un totale di 108654 *tokens*; *Historia Augusta*: opera completa, per un totale di 108632 *tokens*; *Digesta* (I,1,1-VII,8,22,p.3), per un totale di 108631 *tokens*. Dal momento che l'entropia è una funzione estensiva, quindi sensibile all'ampiezza del *corpus*, si è reso necessario esaminare porzioni di testo equivalenti. Il testo italiano di riferimento è una porzione del *Decameron* di analoga estensione (I,1,1-IV,3,6: 108659 *tokens*). Poiché tali testi non sono né annotati morfologicamente né lemmatizzati, le effettive attestazioni di tutte le forme e di tutti i costrutti esaminati sono state verificate manualmente.

[19] La scelta di questo intervallo è arbitraria: è ovvia, infatti, l'esistenza di parole lessicali attestate ben più di 10 volte. Tuttavia, come già mostrato al § 3.1, gli elementi con frequenze comprese fra 1 e 10 arrivano a costituire quasi il 90% del vocabolario di un testo, e possono perciò essere assunti come altamente rappresentativi di tale categoria.

riporta il numero di attestazioni (*tokens*) e l'entropia puntuale (*h*) di ciascuno di essi all'interno dei singoli testi, ed i rispettivi valori medi[20].

	Plauto		Cicerone		Livio		*Hist.Aug.*		*Digesta*		Media
type	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*h*
ego	4479	4,60	3063	5,15	149		236		327		4,87
tu	3011	5,17	2914	5,22	111		249		195		5,20
et	1162	6,55	2739	5,31	1899	5,84	3986	4,77	2922	5,22	5,54
is/ea/id	1528	6,15	2218	5,61	1737	5,97	1740	5,96	3084	5,14	5,77
in	1012	6,75	2008	5,76	2400	5,50	2142	5,66	2116	5,68	5,87
non	947	6,84	1674	6,02	854	6,99	948	6,84	2288	5,57	6,45
ut	1361	6,32	1596	6,09	980	6,79	1383	6,30	961	6,82	6,46
ad	688	7,30	1105	6,62	1265	6,42	1087	6,64	1456	6,22	6,64
cum	444	7,94	1209	6,49	1118	6,60	1402	6,28	717	7,24	6,91

Tabella 2

Sulla base di questo insieme di dati possiamo dunque affermare che, nel *corpus* latino analizzato, gli elementi lessicali prototipici hanno valori di entropia compresi tra 16,73 e 13,41 *bits*, mentre gli elementi grammaticali assumono prototipicamente valori medi compresi tra 4,87 e 6,91 *bits*. Riportando tali valori su una scala numerica, si ottiene la seguente distribuzione (Figura 2)[21].

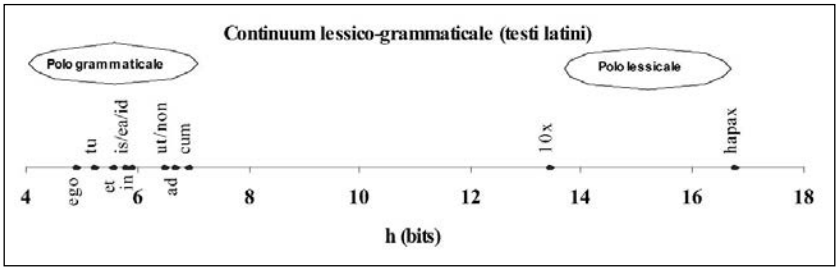

Figura 2

[20] Dei pronomi personali e dei dimostrativi sono state conteggiate tutte le forme flesse. Per quanto riguarda i pronomi personali *ego* e *tu*, la scelta di conteggiare solo le attestazioni presenti in Plauto e Cicerone è dettata da motivazioni legate al genere letterario: rispetto ai testi storiografici e giuridici, testi dialogici ed epistolari riflettono un impiego di tali forme più prossimo a quello reale.

[21] Per alcuni *caveat* metodologici sulla definizione di questo *continuum*, si veda *infra*, § 8.

6.2. *Continuum lessico-grammatica: il Decameron*

Con modalità analoghe è stato quindi definito un *continuum* semantico-funzionale valido per la porzione del *Decameron* esaminata. Anche in questo caso, l'entropia degli elementi lessicali prototipici è compresa tra 16,73 e 13,41 *bits*, dal momento che l'estensione del testo analizzato è uguale a quella di ciascuno dei testi latini. Per quanto riguarda invece i valori che definiscono il polo grammaticale, sono stati calcolati i valori di entropia dei 15 tipi lessicali più frequenti all'interno del testo (Tabella 3).

Decameron		
type	*tokens*	*h*
e	5151	4,40
che	4349	4,64
di	2599	5,39
la	2104	5,69
il	2043	5,73
a	2014	5,75
io/me/mi	1750	5,96
per	1693	6,00
non	1571	6,11
in	1374	6,31
s(i)	1072	6,66
con	880	6,95
l(o)	860	6,98
come	766	7,15
da	754	7,17

Tabella 3

I risultati così ottenuti vengono di seguito riportati sulla scala numerica di Figura 3.

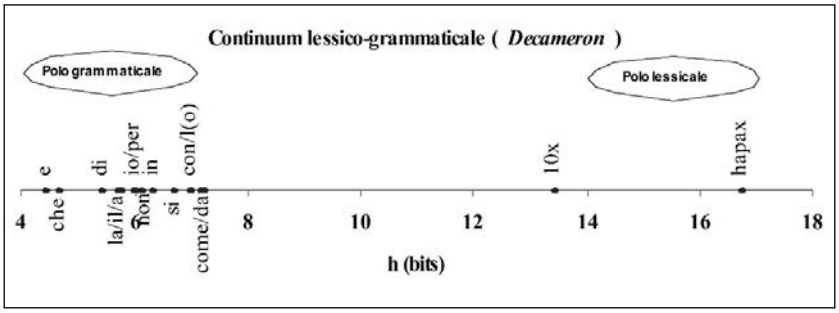

Figura 3

7. *Analisi dei dati*

Dopo aver fissato nel precedente paragrafo alcuni parametri di riferimento, saranno calcolati i valori dell'entropia puntuale di *habere*, *esse*, *avere* ed *essere* (§ 7.1), e dei rispettivi costrutti con participio perfetto (§ 7.2). I risultati di tale operazione verranno quindi raffrontati con i valori che definiscono i due poli opposti nei *continua* semantico-funzionali precedentemente elaborati.

7.1. Esse/habere *vs* essere/avere*: entropia puntuale*

Attraverso il computo delle frequenze di tutte le attestazioni di *habere* e di *esse* nei cinque testi latini presi in esame, sono stati calcolati i rispettivi valori di entropia puntuale, raccolti nella seguente Tabella 4:

	Plauto		Cicerone		Livio		*Hist.Aug.*		*Digesto*		Media
type	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*h*
habere	446	7,95	567	7,58	367	8,22	487	7,80	661	7,36	7,78
esse	4693	4,53	3989	4,77	3300	5,04	3569	4,93	4640	4,55	4,76

Tabella 4

Confrontando tali dati con i valori riportati sul *continuum* di Figura 2, si nota che in tutti i testi analizzati *esse* ha valori di entropia molto bassi e che rientrano pienamente tra quelli degli altri elementi funzionali: d'altra parte *esse* ricopre la funzione di ausiliare non solo nei costrutti perfettivi passivi, ma anche in quelli perifrastici sia attivi che passivi. L'entropia di *habere* ha in effetti valori più elevati, oscillanti tra 7,36 e 8,22 *bits*, che tuttavia collocano anche questo elemento in direzione del polo grammaticale. Questa situazione sembra determinata dal fatto che *habere* presenta comunque una semantica molto ampia i cui significati variano, per citare solo i principali, da "possedere" (Pl., *Cas.* 356: *hariolum hunc habeo domi*; Pl., *Curc.* 530: *ego argentum habeo*) a "stare" (valore primario secondo Heine, 1997; Liv., *aUC* VII,13,7: *utcumque enim se habet res*; Liv., *aUC* VI,35,8: *"bene habet" inquit Sextius*), a "sapere" (Cic., *ad fam.* I,9,20: *Habes de Vatinio. Cognosce de Crasso*). Si noti peraltro che altri predicati semanticamente poco specifici come *facere* hanno valori ancora più bassi ($h = 7{,}15$)[22].

In ogni caso, il confronto con i risultati del *Decameron* rivela una prima

[22] Le frequenze e i rispettivi valori di entropia di *facere* all'interno dei singoli testi sono i seguenti: Plauto: 1290 *tokens*, $h = 6{,}40$ *bits*; Cicerone: 839 *tokens*, $h = 7{,}02$ *bits*; Livio: 522 *tokens*, $h = 7{,}70$ *bits*; *Historia Augusta*: 678 *tokens*, $h = 7{,}32$ *bits*; *Digesta*: 688 *tokens*, $h = 7{,}30$ *bits*.

differenza: nel volgare fiorentino del '300, come prevedibile, *avere* si colloca pienamente all'interno del polo grammaticale, con valori prossimi a quelli di *essere* (Tabella 5). Anche in questo caso il fatto che *essere* abbia comunque valori più bassi rispetto a quelli di *avere* dipende dal fatto che esso è anche ausiliare passivo oltre che perfettivo.

Decameron		
type	*tokens*	*h*
avere	1586	6,10
essere	3240	5,07

Tabella 5

7.2. Esse/habere *+ PP vs* essere/avere *+ PP: entropia condizionata*

L'entropia condizionata consente di ottenere indicazioni più precise, definendo non i valori di un singolo elemento ma quelli di un costrutto. Sono stati dunque calcolati i valori dell'entropia dei costrutti *esse + PP* e *habere + PP*, riportati in Tabella 6.

	Plauto		Cicerone		Livio		*Hist.Aug.*		*Digesto*		Media
type	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*tokens*	*h*	*h*
habere + PP	19	12,48	33	11,69	13	13,03	15	12,82	47	11,17	12,24
esse+ PP	918	6,89	1005	6,76	1073	6,66	1636	6,05	1844	5,88	6,45

Tabella 6

La differenza di valori è in questo caso netta: il costrutto *esse + PP* presenta in tutti i testi un'entropia propria di elementi grammaticali (in media, $h = 6{,}45$), mentre quella di *habere + PP* è fortemente orientata verso il polo lessicale (in media, $h = 12{,}24$). Ad ulteriore conferma, si confrontino i risultati di Tabella 6 con i valori assunti da *avere + PP* ed *essere + PP* all'interno del *Decameron* (Tabella 7).

Decameron		
Type	*tokens*	*h*
avere + PP	990	6,78
essere + PP	1206	6,49

Tabella 7

Se la perifrasi perfettiva *essere + PP* (h = 6,49 *bits*) si sviluppa a partire da un costrutto grammaticalizzato già in latino (*esse + PP*: h = 6,45 *bits*), il perfet-

to del tipo *avere + PP* (h = 6,79 *bits*) è altra cosa rispetto al suo predecessore latino *habere + PP* (h = 12,24 *bits*): il primo è grammaticalizzato e produttivo, il secondo è un composto lessicale[23].

8. *Attuali limiti del modello e* caveat *metodologici*

Prima di concludere, è necessario rilevare alcuni problemi sollevati da un modello di questo tipo ed alcune obiezioni a cui esso inevitabilmente si espone, conseguenze di questioni legate, più in generale, alle stime probabilistiche di dati linguistici. Innanzi tutto, affinché i valori dell'entropia possano consentire inferenze affidabili sul grado di grammaticalizzazione o lessicalizzazione di un certo elemento è infatti necessario fare riferimento a *corpora* sufficientemente ampi e rappresentativi. Tale è senza dubbio il *corpus* latino analizzato, sia per l'estensione (circa 550.000 *tokens* totali) che per la varietà dei generi letterari che esso comprende. Per quanto riguarda la lingua italiana è stato invece ritenuto sufficiente prendere in esame un solo testo: esso infatti assolve all'unico scopo di fornire un termine di confronto a cui rapportare i valori di entropia di *habere* emersi dall'analisi dei testi latini. Per questo motivo la scelta è caduta sul *Decameron*, un testo in cui *avere* ha ormai assunto senza dubbio la funzione di ausiliare, configurando una situazione rimasta fino ad oggi sostanzialmente immutata.

Particolarmente complessa si rivela inoltre una definizione completa e soddisfacente del *continuum* lessico-grammatica. Nell'individuazione dei valori che definiscono il polo funzionale, ad esempio, non è stato possibile tenere conto del ruolo ricoperto dalla morfologia flessiva, in quanto tale *continuum* è stato elaborato a partire dallo spoglio statistico di testi tokenizzati ma non annotati morfologicamente. È assai probabile, comunque, che l'entropia dei morfemi flessivi assuma valori analoghi (se non inferiori) a quella degli elementi funzionali liberi. A titolo indicativo, possiamo rilevare che a seguito di un computo manuale della frequenza della desinenza *-us*, nominativo singolare dei temi in *-o*, l'entropia di tale elemento morfologico all'interno del *corpus* latino esaminato è infatti risultata pari a 4,69 *bits*[24].

[23] Si noti inoltre che i dati di Tabella 6, eccetto nell'ultima parte, non mostrano per il latino (almeno non per quello letterario) una tendenza univoca alla progressiva grammaticalizzazione della perifrasi perfettiva con *habere*, che si conferma un fatto prettamente romanzo.

[24] Le frequenze e i rispettivi valori di entropia del morfema *-us* all'interno dei singoli testi sono invece i seguenti: Plauto: 2799 *tokens*, *h* = 5,28 *bits*; Cicerone: 3335 *tokens*, *h* = 5,03 *bits*; Livio: 4177 *tokens*, *h* = 4,70 *bits*; *Historia Augusta*: 5697 *tokens*, *h* = 4,25 *bits*; *Digesta*: 5958 *tokens*, *h* = 4,19 *bits*.

In secondo luogo, vi è consapevolezza del fatto che, pur avendo individuato i valori propri dei due estremi del *continuum*, sussistono al momento delle difficoltà nel determinare una soglia numerica discriminante fra gli elementi funzionali e quelli semantici, dal momento che al centro del *continuum* le due tipologie sono frammiste e sovrapposte (sebbene mostrino una chiara tendenza ad addensarsi in direzione del rispettivo polo di appartenenza). La soglia a cui situare il passaggio di categoria potrà essere determinata solo in futuro, in seguito a ripetute applicazioni del modello che consentano di disporre di dati empirici sufficienti e di un'esperienza esprimibile in termini di entropia limite. Appare comunque verosimile ipotizzare fin da ora che tale soglia si configurerà non come un confine netto individuato da un valore numerico univoco quanto piuttosto come un confine di tipo *fuzzy* oscillante entro certi limiti.

Esiste infine un limite legato più specificamente al metodo usato in questa sede. Un'analisi basata sul semplice spoglio delle frequenze risulta necessariamente poco raffinata. Senza dubbio essa consente di individuare dei valori prototipici propri degli elementi grammaticali e di quelli lessicali, tuttavia alcuni elementi sfuggono alla maglie larghe di tale analisi: "sebbene" è infatti un elemento grammaticale ma, data la sua scarsa frequenza, è assai probabile che la sua entropia finisca con l'assumere valori analoghi a quello di un elemento lessicale. Quest'ultimo aspetto, a sua volta, si ricollega ad un altro dei problemi legati alla rappresentatività dei *corpora*, ossia la sparsità dei dati, e rivela al contempo la necessità di integrare la definizione frequentista di probabilità attraverso più accurate stime probabilistiche basate su modelli markoviani.

Queste considerazioni ed i rilievi sollevati non inficiano comunque la validità dei dati presentati, sulla base dei quali possiamo correttamente affermare che in latino, elementi senza dubbio funzionali (congiunzioni, preposizioni, pronomi), hanno valori di entropia compresi fra 4,87 e 6,91 *bits*, mentre nella prosa del *Decameron* il valore di elementi analoghi varia da 4,40 a 7,21 *bits*. Per i motivi già esposti (§ 6.1), l'entropia (della maggior parte) degli elementi lessicali è, in entrambi i casi, compresa fra 16,73 e 13,41 *bits*. Inoltre, pur nell'impossibilità di definire una soglia numerica fra usi semantici ed usi funzionali, in questa sede è stato comunque possibile verificare che in termini di entropia esiste una differenza fra *habere* ed *esse* in sincronia, e fra *habere* ed *avere* in diacronia, e soprattutto che, a seguito del processo di ausiliarizzazione, è rilevabile uno spostamento significativo di *avere* e del costrutto *avere + PP* in direzione del polo funzionale rispetto alla posizione occupata in latino da *habere* e *habere + PP*.

9. *Considerazioni conclusive*

Il modello proposto è dunque ancora ampiamente perfettibile nei suoi aspetti tecnici. Tuttavia, pur con tutti gli attuali limiti, o, meglio, nonostante essi, testato su un fenomeno noto come l'ausiliarizzazione di *habere*, esso si è rivelato in grado di cogliere in maniera chiara le differenze che intercorrono tra elementi funzionali e produttivi ed elementi lessicali.

Quanto rilevato nel precedente paragrafo non sminuisce infatti il contributo più significativo che un modello informazionale appare in grado di fornire, ossia la possibilità di esprimere quantitativamente nozioni qualitative e categorie che presentano una intrinseca struttura prototipico-scalare. Gli strumenti di analisi offerti dalla Teoria dell'Informazione, ed in particolare la funzione dell'entropia, consentono infatti una stima numerica del grado di grammaticalizzazione o lessicalizzazione di un certo elemento a partire dalla stima della sua probabilità all'interno di un *corpus* di testi, istituendo così un legame fra categorie qualitative e fatti linguistici numericamente fondati.

L'approccio presentato mostra dunque valide prospettive e si prefigura come uno strumento particolarmente utile se posto al servizio di modelli probabilistici che rimandano ad approcci matematico/computazionali allo studio del linguaggio. L'elaborazione di un modello matematico delle dinamiche linguistiche (sia sincroniche che diacroniche) non può che trarre vantaggio dall'individuazione di una struttura quantitativa e computabile all'interno delle lingue naturali. Più in generale, ogni disciplina che si muova verso gradi maggiori di modellazione scientifica richiede un'unità di misura che permetta di trasformare l'intuizione qualitativa in un indice numerico misurabile e confrontabile: la scienza diventa tale quando qualità apparenti come il caldo e il freddo dei corpi si possono trasformare in affermazioni numeriche sulle temperatura di un corpo.

Bibliografia

BYBEE, J. (2001), *Phonology and Language Use*, Cambridge University Press, Cambridge UK.

BYBEE, J. e HOPPER, P. (2001), *Introduction to frequency and the emergence of linguistic structure*, in BYBEE, J. e HOPPER, P. (2001, *eds.*), *Frequency and the Emergence of Linguistic Structure*, Benjamins, Amsterdam, pp. 1-24.

BYBEE, J. e MODER, C. (1983), *Morphological Classes as Natural Categories*, in «Language», 59, pp. 251-270.

BYBEE, J. e PAGLIUCA, W. (1985), *Cross-linguistic comparison and the development of grammatical meaning*, in FISIAK, J. (1985, *ed.*), *Historical Semantics and Historical Word Formation*, de Gruyter, Berlin, pp. 59-83.

BYBEE, J. e SLOBIN, D. (1982), *Rules and schemas in the development and use of the English past tense*, in «Language», 58, pp. 265-289.

BYBEE, J., PAGLIUCA, W. e PERKINS, R. (1994), *The Evolution of Grammar. Tense, Aspect and Modality in the Languages of the World*, University of Chicago Press, Chicago.

DAHL, Ö. (2001), *Inflationary effects in language and elsewhere*, in BYBEE, J. e HOPPER, P. (2001, *eds.*), *Frequency and the Emergence of Linguistic Structure*, Benjamins, Amsterdam, pp. 471-480.

DAHL, Ö. (2003), *The Growth and Maintenance of Linguistic Complexity*, Benjamins, Amsterdam.

FENK-OCZLON, G. (2001), *Familiarity, information flow, and linguistic form*, in BYBEE, J. e HOPPER, P. (2001, *eds.*), *Frequency and the Emergence of Linguistic Structure*, Benjamins, Amsterdam, pp. 431-448.

GIVÓN, T. (1991), *Isomorphism in the grammatical code: cognitive and biological considerations*, in «Studies in Language», 15, pp. 85-114.

GOLDSMITH, J. (2001), *Probability for Linguists* (*http://humanities.uchicago.edu/faculty/goldsmith/Industrial/Probability.htm*).

HAIMAN, J. (1980), *The iconicity of grammar*, in «Language», 56, pp. 515-40.

HASPELMATH, M. (1999), *Why is grammaticalization irreversible?*, in «Linguistics», 37, 6, pp. 1043-1068 (*http://email.eva.mpg.de/~haspelmt/publist.html*).

HASPELMATH, M. (2006), *Against markedness (and what to replace it with)*, in «Journal of Linguistics», 42, 1, pp. 25-70 (*http://email.eva.mpg.de/~haspelmt/publist.html*).

HASPELMATH, M. (2008), *Creating economical morphosyntactic patterns in language change*, in GOOD, J. (2008, *ed.*), *Language Universals and Language Change*, Oxford University Press, Oxford, pp. 185-214 (*http://email.eva.mpg.de/~haspelmt/publist.html*).

HEINE, B. (1993), *Auxiliaries. Cognitive Forces and Grammaticalization*, Oxford University Press, New York-Oxford.

HEINE, B. (1997), *Possession: Cognitive Sources, Forces, and Grammaticalization*, Cambridge University Press, Cambridge UK.

HOPPER, P.J. (1987), *Emergent grammar*, in «Berkeley Linguistic Society», 13, pp. 139-157.

HOPPER, P.J. (1988), *Emergent Grammar and the A Priori Grammar Postulate*, in TANNEN, D. (1988, *ed.*), *Linguistics in Context. Connecting Observation and Understanding*, Ablex, Norwood NJ, pp. 117-134.

HOPPER, P.J. (1998), *Emergent Grammar*, in TOMASSELLO, M. (1998, *ed.*), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure*, Lawrence Erlbaum Associates, Mahwah-London, pp. 155-175.

HOPPER, P.J. e CLOSS TRAUGOTT, E. (1993), *Grammaticalization*, Cambridge University Press, Cambridge UK.

JURAFSKY, D., BELL, A., FOSLER-LUSSIER, E., GIRAND, C. e RAYMOND, W.D. (1998), *Reduction of English function words in Switchboard*, in «ICSLP-98», 7, pp. 3111-3114.

JURAFSKY, D., BELL, A., GREGORY, M. e RAYMOND, W.D. (2001), *Probabilistic relations between words: Evidence from reduction in lexical production*, in BYBEE, J. e HOPPER, P. (2001, *eds.*), *Frequency and the Emergence of Linguistic Structure*, Benjamins, Amsterdam, pp. 229-254.

KELLER, R. (1994), *Language Change. The Invisible Hand in Language*, Routledge, London.

KURYŁOWICZ, J. (1975 [1965]), *The evolution of grammatical categories*, in «Diogenes», 51, pp. 55-71. Riedito in KURYŁOWICZ, J., *Esquisses linguistiques II*, W. Fink, München, pp. 38-54.

LA FAUCI, N. (1988), *Oggetti e soggetti nella formazione della morfosintassi romanza*, Giardini, Pisa.

LANGACKER, R.W. (2000 [1992]), *The symbolic nature of cognitive grammar: The meaning of of and of of-periphrasis*, in PÜTZ, M. (1992, *ed.*), *Thirty Years of Linguistic Evolution*, Benjamins, Amsterdam, pp. 483-502. Riedito in LANGACKER, R.W. (2000, *ed.*), *Grammar and Conceptualisation*, Mouton de Gruyter, Berlin-New York, pp. 73-90.

LEHMANN, C. (1985), *Grammaticalization: Synchronic Variation and Diachronic Change*, in «Lingua e Stile», 20, 3, pp. 303-318.

LEHMANN, C. (1995 [1982]), *Thoughts on Grammaticalization*, LINCOM EUROPA, München-Newcastle.

LONGOBARDI, G. (1995), *A case of construct state in Romance*, in AJELLO, R. e SANI, S. (1995, *eds.*), *Scritti linguistici e filologici in onore di Tristano Bolelli*, Pacini, Pisa, pp. 293-329.

LOPORCARO, M. (1998), *Sintassi comparata dell'accordo participiale romanzo*, Rosenberg & Sellier, Torino.

MAROTTA, G. (2001), *Non solo spiranti. La 'gorgia toscana' nel parlato di Pisa*, in «Italia dialettale», 66, pp. 27-60.

MAYERTHALER, W. (1980), *Morphologische Natürlichkeit*, Athenaion, Wiesbaden.

MEILLET, A. (1958 [1912]), *L'évolution des formes grammaticales*, in MEILLET, A. (1958, *ed.*) *Linguistique historique et linguistique générale*, Champion, Paris, pp. 130-148.

PINKSTER, H. (1987), *The Strategy and Chronology of the Development of Future and Perfect Tense Auxiliaries in Latin*, in HARRIS, M. e RAMAT, P. (1987, *eds.*), *Historical Development of Auxiliaries*, Mouton de Gruyter, Berlin-New York-Amsterdam, pp. 193-223.

RAMAT, P. (1984), *Un esempio di rianalisi: le forme perifrastiche nel sistema verbale delle lingue romanze*, in RAMAT, P. (1984, *ed.*) *Linguistica tipologica*, Il Mulino, Bologna, pp. 143-164.

SCHLEGEL, A.W. (1971 [1818]), *Observations sur la langue et la littérature provençales*, Libraire greque-latine-allemande, Paris.

SCHUCHARDT, H. (1972 [1885]), *Über die Lautgesetze: Gegen die Junggrammatiker*, Robert Oppenheim, Berlin. Riedito in VENNEMANN, T. e WILBUR, T. (1972, *eds.*), *Schuchardt, the Neogrammarians and the Transformational Theory of Phonological Change*, Athenäum, Frankfurt, pp. 39-72.

SHANNON, C.E. (1948), *A Mathematical Theory of Communication*, in «The Bell System Technical Journal», 27, pp. 379-423; 623-656 (*http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf*).

ZIPF, G.K. (1929), *Relative Frequency as a Determinant of Phonetic Change*, in «Harvard Studies in Classical Philology», 40, pp. 1-95.

ZIPF, G.K. (1935), *The Psycho-Biology of Language: an Introduction to Dynamic Philology*, Houghton Mifflin, Boston.

ZIPF, G.K. (1938), *Phonometry, Phonology, and Dynamic Philology: an Attempted Synthesis*, in «American Speech», 13, 4, pp. 275-285.