Seminari "Liceo matematico"
Roma – Sapienza, 19 febbraio 2021

Lingue, parole e numeri: fenomeni linguistici calcolabili e non calcolabili

Luca Lorenzetti, Università della Tuscia – VT

l.lorenzetti@unitus.it

qualche preliminare

"calcolabile" probabilmente non è il termine giusto: forse meglio "trattabili con metodi quantitativi con risultati migliori di quelli che si ottengono con metodi qualitativi"

far collaborare linguistica E matematica nei licei: il problema è la linguistica, non la matematica

vedremo alcuni settori della linguistica nei quali il possibile riferimento a considerazioni quantitative è entrato più o meno nella dottrina

non faremo riferimento quasi per nulla a espressioni matematiche formali: il punto saranno possibili convergenze generali di interessi, non la correttezza o efficacia dei trattamenti tecnici.

che cosa si può contare/calcolare delle lingue?

1. la quantità di lingue esistenti
2. il loro grado di diversità comunitario
3. il loro grado di diversità strutturale
4. gli elementi dei livelli della grammatica e la loro distribuzione: suoni, forme, parole, significati, costrutti.

NON toccheremo, se non tangenzialmente, temi di linguistica computazionale, cioè di trattamento automatico computerizzato delle lingue e dei testi.

quante lingue ci sono al mondo?

Ethnologue 2020: 7117 lingue viventi

(nel 2015 erano 7097, nel 2014 7102, nel 2013 7547, nel 2009 6909) (www.ethnologue.com/world, edito a cura del *Summer Institute of Linguistics*, "a global, faith-based nonprofit that works with local communities around the world to develop language solutions that expand possibilities for a better life".

circa 7000 lingue per circa 230 stati =
non meno di 27 lingue per ogni stato

gamma popolazione	lingue viventi		
	lingue[1]	%[2]	% cumulative[3]
più di 100 milioni di parlanti	8	0,1	0,1
da 10 milioni a 99.999.999	75	1,1	1,2
da un milione a 9.999.999	304	4,4	5,6
da 100.000 a 999.999	895	13	18,6
da 10.000 a 99.999	1824	26,4	45
da 1.000 a 9.999	2014	29,2	74,1
da 100 a 999	1038	15	89,2
da10 a 99	339	4.9	94.1
da 1 a 9	133	1,9	96
Ignota	277	4	100
Totali	**6909**	**100**	

	numero di parlanti		
gamma popolazione	conteggio	%	% cumulative
più di 100 milioni di parlanti	2.308.548.848	38,73	38,73
da 10 milioni a 99.999.999	2.346.900.757	39,38	78,12
da un milione a 9.999.999	951.916.458	15,97	94,09
da 100.000 a 999.999	283.116.716	4,75	98,84
da 10.000 a 99.999	60.780.797	1,02	99,86
da 1.000 a 9.999	7.773.810	0,13	99,99204
da 100 a 999	461.250	0,00774	99,99978
da10 a 99	12.560	0,00021	99,99999
da 1 a 9	521	0,00001	100
Ignota			
Totali	**5.959.511.717**	**100**	

limiti dei conteggi delle lingue

Ethnologue, *Italy*: 33 lingue censite, tra cui "napoletano-calabrese", lombardo, piemontese, siciliano, veneziano; sono classificati come dialetti, e quindi non conteggiati nel calcolo della diversità linguistica, pugliese, abruzzese e molisano

→ repertorio basato su fonti difformi

→ distinzioni prive di fondamento obiettivo:
a) teorico (che cosa definiamo "lingua"?)
b) empirico (su che base aggiorniamo il repertorio?)

grado di diversità delle comunità linguistiche

l'indice di diversità linguistica (Greenberg 1956) misura la probabilità totale di scegliere due persone della stessa lingua prelevandole a caso all'interno della popolazione di una data area geografica.

L'indice di diversità linguistica A per uno stato sarà così uguale a 1 meno la somma dei quadrati delle frazioni di parlanti ogni lingua presente nello stato:

$$A = 1 - \Sigma_i(i^2)$$

grado di diversità delle comunità linguistiche

Se in uno stato sono presenti 3 lingue:
A parlata da 1/8 della popolazione
B parlata dai 3/8
C parlata da 1/2

allora l'indice di diversità linguistica (IDL) sarà uguale a

$[1 - (1/8)^2 + (3/8)^2 + (1/2)^2] =$
$1 - 26/64 =$
$38/64 =$ **0,59**.

problemi e correttivi al calcolo di IDL – 1

IDL conta gli individui come parlanti di una sola lingua, una condizione rara nel mondo

→ si contano i parlanti bilingui come se fossero due persone, i trilingui come se fossero tre persone e così via

→ anziché basarsi sull'assunto che "communication among people speaking different languages can successfully occur only if at least one single language is shared", si sviluppano nuovi indici "that describe the probability that people with different linguistic repertoires can effectively communicate by relying on their receptive competence in multiple languages, or a mix between the two communication strategies" (Gazzola, Templin, McEntee-Atalianis 2019).

→ ma più si raffinano i criteri più diventa difficile raccogliere i dati: non esistono né possono esistere censimenti della popolazione che registrino in maniera attendibile le percentuali di plurilinguismo attivo e passivo, sicché un calcolo comparativo su scala globale diventa di fatto impossibile.

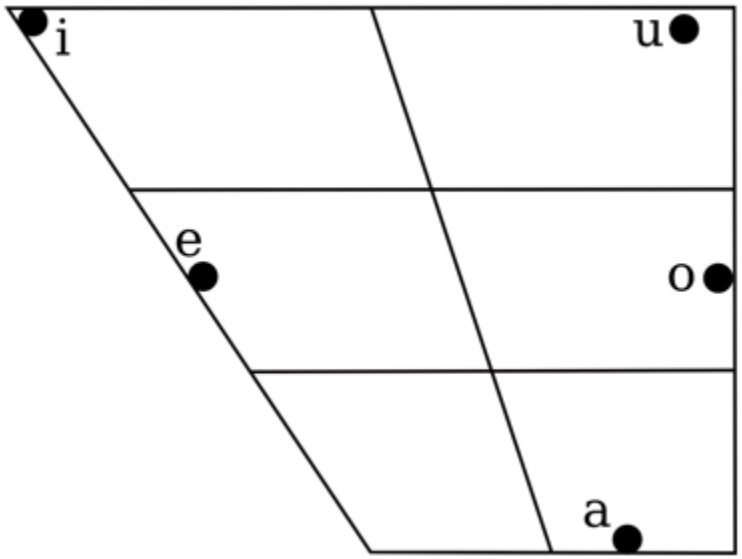
problemi e correttivi al calcolo di IDL – 2

IDL non fa differenza di grado nella distanza tra le lingue considerate: uno spagnolo che parla spagnolo e uno che parla catalano, lingue piuttosto simili, si capiranno di più che uno spagnolo che parla spagnolo e uno che parla basco, lingue totalmente estranee l'una all'altra.

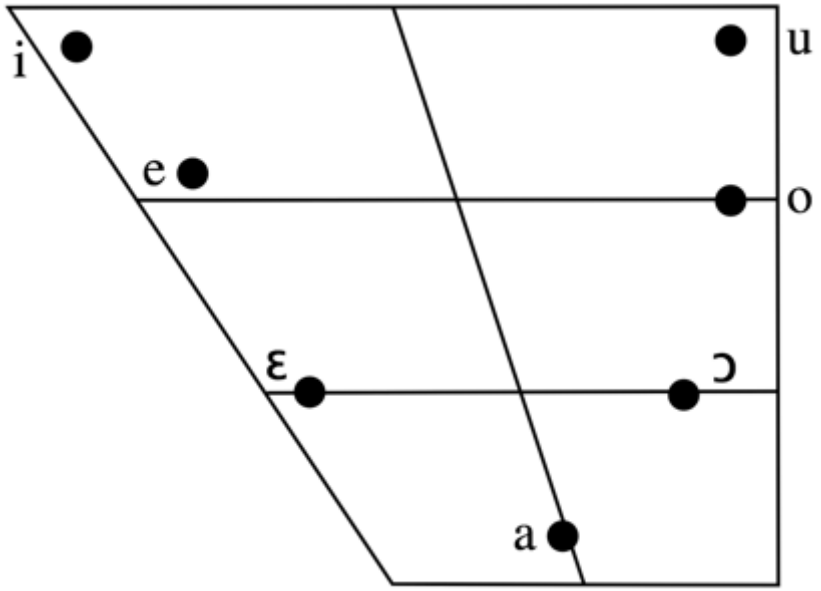
	PADRE	MADRE	SORELLA	ZIO	NONNO
catalano	pare	mare	germana	oncle	avi
spagnolo	padre	madre	hermana	tío	abuelo
basco	aita	ama	ahizpa	osaba	aitatxi

→ si moltiplica l'indice per un "fattore di somiglianza" ottenuto calcolando le differenze tra lingue sulla base di una lista di 200 parole indicanti concetti o oggetti di uso universale e quindi presumibilmente presenti in tutte le lingue.

calcolare la diversità strutturale delle lingue



vocali spagnolo



vocali italiano

carico o **rendimento funzionale** delle opposizioni fonologiche: grado di uso di una opposizione per distinguere parole con significati diversi in una lingua data

accétta	accètta
affétto	affètto
annétto, annétti	annètto, annètti
crédo	crèdo
dei	dèi
détte, détti	dètte, dètti
e	è
ésca	èsca
ésse	èsse
féci	fèci
légge, léggi	lègge, lèggi
léssi, lésse	lèssi, lèsse
mésse	mèsse
méssi	mèssi
mésto, mésti, mésta	mèsto, mèsti, mèsta
mézzo, mézzi, etc.	mèzzo, mèzzi, etc.
nei	nèi
péne	pène
pésca	pèsca
péste	pèste
re	re
réne	rène
réso	rèso
te	tè
tési	tèsi
vénti	vènti

/ɛ/ ~ /e/ in italiano

calcolo del rendimento funzionale delle opposizioni fonologiche

1) Hockett (1955, 1966), cit. da Chiari (2005): il rendimento funzionale di una coppia di fonemi è un cambiamento dell'entropia del sistema.

Data un'entropia H, pari alla somma dei prodotti della probabilità di ogni fonema *i* del sistema per il logaritmo della probabilità di *i*:

$$H = -\sum_{i \in N} p_i * log_2(p_i)$$

il rendimento funzionale di una coppia di fonemi {*a*,*b*} è pensabile come una sottrazione tra l'entropia H del sistema quando *a* e *b* sono distinti e l'entropia H* che il sistema avrebbe se i fonemi *a* e *b* non si distinguessero più, a sistema altrimenti immutato: (H–H*)/H

2) Rischel (1962), cit. da Chiari (2005): il rendimento funzionale di una coppia di fonemi è dato dalla somma di tutti i prodotti delle frequenze di occorrenza delle parole in coppia minima.

→ i risultati sono diversi a seconda che si prendano in considerazione le frequenze sul vocabolario o le frequenze nel testo (/mɛddzo/ ~ /mettso/ ≠ /ɛ/ ~ /e/)

statistica lessicale: la stratificazione sociolinguistica del lessico

Lessico fondamentale, 2049 lessemi di altissima frequenza, le cui occorrenze costituiscono circa il 90% delle occorrenze lessicali in tutti i testi scritti e parlati = *a, avere, il, cosa, essere, vedere*

Lessico di alto uso, 2576 vocaboli di alta frequenza, le cui occorrenze sono un altro 6% circa delle occorrenze lessicali di tutti i testi = *impaurire, impianto*

Lessico di alta disponibilità, 1897 vocaboli, relativamente rari nel parlare o scrivere, ma ben noti perché legati ad atti e oggetti di grande rilevanza nella vita quotidiana = *alluce, batuffolo, carrozzeria, dogana* …

fondamentale + alto uso + alta disponibilità formano il vocabolario di base, 6522 parole, 98% del discorso

Lessico comune, circa 50.000 lessemi estranei al VDB ma noti a chiunque abbia un livello di istruzione medio-superiore

Lessico tecnico-specialistico, centinaia di migliaia di lessemi usati prevalentemente o solo in ambito scientifico, tecnologico o professionalmente settoriale

Lessico d'uso non tecnico-scientifico ma non comune: basso uso, obsoleto, solo letterario, dialettale, regionale.

come si ottengono le frequenze d'uso dei lessemi?

1) scelta di corpus testuali rappresentativi per quantità e qualità
2) lemmatizzazione: riporta le varie forme con cui le singole parole ricorrono nel testo analizzato alla loro forma di base, ossia al *lessema*, detto in lessicografia *lemma* (ad es., *penso, pensavo, pensarono, pensasse, pensando* ecc. vanno tutti ricondotti all'infinito *pensare*), risolvendo anche i problemi di omografia e di allografia che di volta in volta si possono presentare

[omografia: *ancora capitano danno la le lo leggere perdono scrivano subito venti*; fenomeno estesissimo in italiano: sono state contate (Casadei 2016) 112.344 forme omonime riconducibili a 35.557 lessemi, a partire dal GRADIT, a oggi il più esteso lemmario italiano || allografia: *chimono kimono*]

come si rappresentano le frequenze d'uso dei lessemi?

lessici e liste di frequenza: il *Lessico di frequenza dell'italiano parlato* (1993)

1.	Art	IL	41.459
2.	P	DI	19.915
3.	V	ESSERE	15.716
4.	Art	UNO	12.807
5.	P	A	12.001
6.	Pro	EGLI	10.181

SPARIRE V	1309	2	3	5	5	4	19
sparire		0	0	4	1	2	7
spariscano		0	0	0	1	0	1
sparisce		0	1	1	1	1	4
spariscono		0	0	0	1	0	1
sparite		2	0	0	0	0	2
spariti		0	1	0	0	0	1
sparito		0	1	0	0	1	2
spariva		0	0	0	1	0	1

– liste di frequenza in altre lingue europee: http://www.wordfrequency.info/

– strumenti aggiornati adottati oggi in linguistica computazionale:
http://www.glottoteca.unina.it/computazionale/doku.php ⇒

Le diapositive seguenti, che non ebbi il tempo di presentare a lezione, sono qui corredate di brevi commenti per contestualizzarne la lettura. Si tratta di possibili usi di elementari sinossi quantitative per approfondire le dinamiche di formazione ed evoluzione del lessico italiano.
Alla fine della presentazione ho aggiunto alcune indicazioni in merito ad alcune delle domande emerse dopo la lezione.

La tabelle seguente (tav. 4) mostra la quantità di parole di origine straniera presenti in italiano secondo il lemmario del GRADIT. Colpisce che l'apporto più numeroso sia quello del greco (antico), che supera addirittura l'inglese. Questa apparente curiosità si spiega osservando in tav. 5 che la maggior parte di questi prestiti dal greco all'italiano è attestata per la prima volta nei secc. XIX e XX. Si tratterà quindi non di ingressi diretti dal lessico classico, ma di un altro genere di apporti: un commento si trova alla diapositiva 22.

applicazioni dinamiche della statistica lessicale: greco in italiano

Tav. 4. – Principali apporti esogeni al lessico italiano

lingue di provenienza	*esotismi*	*vocaboli adattati*	*totale*	*di cui* TS *o integrati*
greco	12	8342 [1]	8354	
inglese [2]	4208	1302	5510	2567
francese	1427	2943	4370	910
spagnolo [3]	281	775	1056	152
tedesco	278	326	604	211
arabo	203	266	469	177
provenzale			240	
russo	86	148	234	55
portoghese [3]			168	
giapponese	125	48	173	69
turco	45	88	133	20
longobardo			114	
sanscrito			92	32
ebraico	36	32	68	13
persiano			68	
cinese	18	40	58	12
hindi			54	

[1] Di cui solo 3891 sono etimi diretti di parole italiane e gli altri sono etimi di etimi latini delle diverse fasi e forme della latinità (vedi § 5.1).

[2] Include 135 lessemi originati dall'inglese americano.

[3] Inclusi alcuni prestiti dall'ispanoamericano e dal portoghese brasiliano.

esempi di applicazioni della statistica lessicale: greco in italiano

Tav. 5. – Stratificazione diacronica dei grecismi nel lessico italiano

secoli	*greco*	*dal greco*	*gr. dorico*	*dal gr. tardo*	*gr. bizantino*	*dal gr. bizantino*	*dal gr. mediev.*	*gr. mod.*	*dal gr. mod.*	*Totali*
—	2	1074				5				1076
XII				1						1
XIII		12				5			1	18
XIV		24		2		4				30
XV	1	61		4	1	14				81
XVI	1	202		2		5	2			212
XVII		189		4	1	4				198
XVIII		248		3	1	2	1	1		256
XIX	2	1040	1	16		13	2	2	6	1082
XX	3	884	1	6		27		7	9	937
Totali	*9*	*3734*	*2*	*38*	*3*	*74*	*5*	*10*	*16*	*3891*

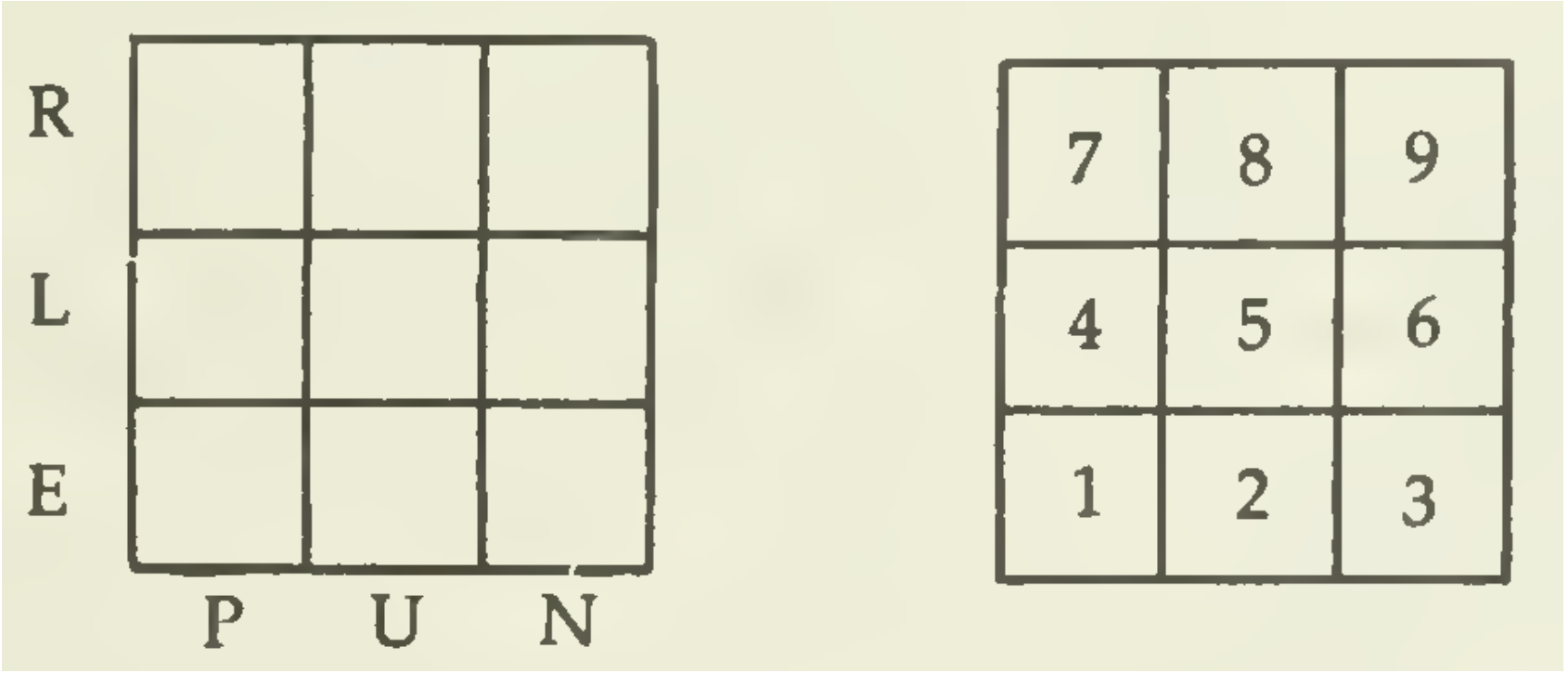
composizione neoclassica: regola di formazione di nuovi lessemi consistente nell'accostare due basi lessicali legate aggiungendo la relativa flessione:

epatopatia, fotografia, ginecologia, glottologia, nevralgia ecc.

Formazioni del tutto nuove per il lessico italiano, con due caratteristiche importanti:
1) non sono parole esistenti in greco, o esistenti in greco con lo stesso significato, bensì parole nate in Europa, perlopiù in francese , a partire dal Settecento. La nascente scienza sperimentale europea, che aveva il suo centro nella Francia illuminista, dovette creare terminologie nuove per esprimere i concetti necessari alle nuove tecniche ed evitare la nascita incontrollata di sinonimi che imbarazzava gli scienziati dell'epoca, specie i chimici. La soluzione fu il ricorso alle terminologie classiche. Iniziò così a circolare in tutte le lingue europee di cultura un gran numero di termini, perlopiù composti, come *bio-grafo*, o a base composta, come (*bio-graf-*)*ico*, formati di elementi la cui origine remota era il più delle volte il greco antico, mentre il veicolo diretto ne era il francese: lessemi composti da due o più elementi non autonomi di origine greca o latina, che però non possono costituire da soli un lessema autonomo: non si può dire ho *l'*epato gonfio* oppure il *mi piacciono le *gineco* (*gineche*?) ecc.
2) l'ordine degli elementi è Determinante+Determinato, come in greco: la *fotografia* è una -*grafia* (disegno) fatta con la *foto-* (luce), la *glottologia* la scienza (*-logia*) della lingua (*glotto-*), al contrario dell'ordine normale per i composti endogeni, che sono Dto+Dte: *capostazione, cassaforte, pescecane* .

Indicazioni e spunti in risposta alle domande emerse durante e dopo la lezione:

il reticolo di Heinz Kloss (1976) permette di rappresentare graficamente l'esistenza o meno in una lingua di testi non letterari di vario livello: elementare (E), liceale (L) e di ricerca (R) e su vari temi: patrii o nazionali (P) e esteri. I temi si suddividono poi in umanistici (U) e temi riguardanti scienze naturali, fisiche e tecnologiche (N). Lo schema funziona così come un indice sintetico del grado di elaborazione funzionale di una lingua data.



indici di vitalità linguistica:
EGIDS: Expanded Graded Intergenerational Disruption Scale 1

Level	Label	Description
0	International	The language is widely used between nations in trade, knowledge exchange, and international policy.
1	National	The language is used in education, work, mass media, and government at the national level.
2	Provincial	The language is used in education, work, mass media, and government within major administrative subdivisions of a nation.
3	Wider Communication	The language is used in work and mass media without official status to transcend language differences across a region.
4	Educational	The language is in vigorous use, with standardization and literature being sustained through a widespread system of institutionally supported education.
5	Developing	The language is in vigorous use, with literature in a standardized form being used by some though this is not yet widespread or sustainable.
6a	Vigorous	The language is used for face-to-face communication by all generations and the situation is sustainable.

indici di vitalità linguistica:
EGIDS: Expanded Graded Intergenerational Disruption Scale 2

6b	Threatened	The language is used for face-to-face communication within all generations, but it is losing users.
7	Shifting	The child-bearing generation can use the language among themselves, but it is not being transmitted to children.
8a	Moribund	The only remaining active users of the language are members of the grandparent generation and older.
8b	Nearly Extinct	The only remaining users of the language are members of the grandparent generation or older who have little opportunity to use the language.
9	Dormant	The language serves as a reminder of heritage identity for an ethnic community, but no one has more than symbolic proficiency.
10	Extinct	The language is no longer used and no one retains a sense of ethnic identity associated with the language.

per approfondire

Chiari I., *Introduzione alla linguistica computazionale*, Laterza 2007.

De Mauro T., Chiari I. (a cura di), *Parole e numeri. Analisi quantitative dei fatti di lingua*, Aracne 2005.

Lenci A., Montemagni S., Pirrelli V., *Testo e computer. Elementi di linguistica computazionale*, Carocci 2005.

su *piuttosto che* e simili innovazioni censurate, si trovano notizie in Lorenzetti, L., *Dictionaries of Language Difficulties*, in F. Lebsanft, F. Tacke (eds.), *Manual of Standardization in the Romance Languages*, de Gruyter 2020, pp. 373-397, a p. 385 sgg. (lavoro sotto copyright, posso inviare un preprint a chi mi scriva).