I NUNC-ES: strumenti nuovi per la linguistica dei corpora in spagnolo

Manuel BARBERA

Dip. di Scienze letterarie e filologiche
Università di Torino
b.manuel@inrete.it

RIASSUNTO
Dopo aver in breve presentato NUNC, suite muiltilingue di corpora basati su testi di UseNet e liberamente interrogabili online *(http://www.bmanuel.org/projects/ng-HOME.html),* se ne evidenziano le caratteristiche innovative; l'articolo si sofferma quindi sui corpora in lingua spagnola (NUNC-ES), descrivendo sommariamente le gerarchie di UseNet in spagnolo ed i corpora attualmente posti online, da cui è presentato qualche esempio di query. Si delineano infine gli sviluppi futuri, anticipando soprattutto la pubblicazione di un nuovo tagset.

Parole chiave: linguistica dei corpora, newsgroup, spagnolo, corpora, tagset

NUNC-ES: New Tools for Corpus Linguistics

ABSTRACT
After a short presentation of NUNC, a freely available multilingual suite of corpora based on newsgroups texts *(querable online at http://www.bmanuel.org/projects/ng-HOME.html),* this paper intends to investigate the Spanish subset of data collected in NUNC-ES. A brief description of the Spanish hierarchies is given, and some examples of corpus queries are suggested. The third part of the work presents an outline of future developments, especially the release of a new tagset for Spanish.

Key Words: corpus linguistics, newsgroup, Spanish, corpora, tagset

0. INTRODUZIONE

I NUNC (*Newsgroups UseNet Corpora*) sono una innovativa collezione multilingue[1] di corpora[2] di lingua contemporanea, tanto generici quanto specialistici[3] basati sui messaggi dei newsgroup[4], e liberamente interrogabili online (homepage: *http://www.bmanuel.org/projects/ng-HOME.html*) tramite un'unica interfaccia, che si appoggia all'architettura del Corpus Query Workbench (CWB[v]), con il potente motore di ricerca CQP (sul quale cfr. Christ - Schulze 1996 e Heid 2007), sviluppata dall'Institut für maschinelle Sprachverarbeitung di Stuttgart[6].

Molto in breve, «un newsgroup è un forum telematico a libero accesso, gratuito, disponibile su Internet, che si manifesta nella forma di testi scritti, ed il cui funzionamento è assai semplice: ogni utente scrive un messaggio, il post, e lo invia ad una specie di "bacheca elettronica" mantenuta presso una rete di server (i newsserver che costituiscono UseNet), dai quali gli altri utenti del gruppo possono scaricarlo, leggerlo e rispondervi, costruendo anche articolate catene (thread) di botte e risposte. La facilità d'uso garantisce la grande diffusione dello strumento tra le categorie più diverse di utenti e giustifica la grande quantità di traffico esistente su UseNet. Queste "bacheche elettroniche" che sono i newsgroup sono poi articolate in una tassonomia precisa, ossia in un sistema di cornici argomentative che si chiamano "gerarchie", a base geografico-nazionale e/o tematica; anche queste gerarchie, peraltro, nascono dal basso in base alla iniziativa degli utenti» (Barbera 2007b).

I NUNC-ES attualmente online sono 4, un generico di circa 30+ milioni di parole e tre specialistici, di dimensioni varie, presentati nella Tav. 1 con le rispettive cifre di token[7] e type[8]:

---

[1] Le lingue coperte dal progetto sono per ora danese, estone, finnico, francese, italiano, inglese (britannico ed australiano), portoghese (europeo e brasiliano), spagnolo, tedesco (non austriaco e svizzero) ed ungherese.

[2] Sulla decisione di considerare "corpus" (pl. "corpora"), ed analogamente "newsgroup", "post" ecc. (pl. invariabili), come prestiti a tutti gli effetti (e pertanto di rappresentarli in tondo anziché in corsivo), cfr. Barbera - Marello (2003 *i.s.*) e Barbera 2007a.

[3] I settori specialistici su cui abbiamo per ora sperimentato sono quelli dell'alimentazione, della fotografia e dei motori; ma ovviamente in futuro se ne potranno studiare altri ancora.

[4] Indovinai l'utilità di questa fonte testuale ed iniziai i primi download sperimentali di testi nell'inverno 2001; il primo prototipo di corpus, di lingua italiana, fu allestito nel 2002. Forte di queste esperienze, proposi (Barbera 2004) quindi UseNet come principale fonte dei corpora del progetto FIRB (cfr. Barbera 2007b).

[5] Cfr. *http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/*

[6] *http://www.ims.uni-stuttgart.de/ims-home.html.es*

[7] Per una definizione formale di "token" cfr. Barbera - Corino - Onesti 2007a, specie § 1.3; in termini spiccioli il token è l'unità minima di cui è composto un corpus, perlopiù individuata da due spazi bianchi; la sua coincidenza con la nozione tradizionale di "parola grafica" è (almeno nelle principali lingue europee) relativamente buona, ma solo parziale: in italiano, ad es. "l'oro" costituisce un solo blocco grafico, ma contiene due token; tanto in italiano come in spagnolo i segni di interpunzione formano un unico elemento grafico con quanto li precede, ma pure costituiscono token distinti, ecc.

[8] Per una definizione formale di "type" cfr. Barbera - Corino - Onesti 2007a, specie § 1.3; molto alla buona, in prima approssimazione, si può intendere come la classe di un gruppo di token uguali (le "forme", come talvolta si dice).

	n. token	n. type
NUNC-ES *Generic*	31.240.227	809.977
NUNC-ES *Cooking*	2.098.489	118.250
NUNC-ES *Photo*	725.389	30.956
NUNC-ES *Motor*	13.415.613	487.228

Tav. 1: *I NUNC-ES online*

L'articolazione del sottoinsieme NUNC-ES è sostanzialmente la medesima di NUNC-IT, che è un po' il capostipite di tutti i NUNC:

	n. token	n. type
NUNC-IT *Generic* I	127.708.505	1.346.652
NUNC-IT *Generic* II	109.692.794	1.098.829
NUNC-IT *Cooking*	4.161.627	187.544
NUNC-IT *Photo*	8.544.089	374.289
NUNC-IT *Motor*	8.544.089	374.289
NUNC-IT *Photo-uncut*	17.580.298	513.404

Tav. 2: *I NUNC-IT online*

Le uniche differenze sono infatti le diverse dimensioni (per cui cfr. *infra* § 1 punto *j* e § 3.1), e la presenza di un "double" di Photo (per cui cfr. *infra* § 1, ultimo capoverso).

1. CARATTERISTICHE GENERALI DEI NUNC

Oltre a presentare alcuni generali vantaggi (cfr. *infra* j-ij) per un costruttore di corpora, UseNet (cfr. Corino 2007) presenta alcune caratteristiche peculiari (cfr. *infra* a-d), che conferiscono poi specifiche interessanti ai corpora che ne sono tratti, e che sono pertanto comuni (in maggiore o minore misura) a tutta la suite NUNC.

I vantaggi generali sono presto detti. In primo luogo (j) vi è l'abbondanza che di solito si ha di materiale testuale: in alcune gerarchie il traffico è assai elevato e bastano poche settimane per accumulare grandi quantità di testi; altre gerarchie, però, sono assai meno frequentate e può essere necessario raccogliere più annate di post (e vedremo che questo è in parte il caso per alcune gerarchie spagnole). Il secondo punto (ij) concerne la libertà da copyright, ossia la verosimile disponibilità legale dei materiali presenti nei newsgroup: UseNet, infatti, per definizione e tradizione è il regno del pubblico dominio, ed è rimasta molto più stretta del *World Wide*

Web alle sue radici anarchico-liberali (ammesso che una tale qualificazione abbia alcun senso), e quindi ciò sembrerebbe una ovvia assunzione[9].

Altre caratteristiche sono forse meno ovvie. In primo luogo, (a) la lingua dei newsgroup è una sorta di *Umgangssprache*[10] molto variegata nei registri (dalla chiacchierata informale al saggio, alla novella, od al parere tecnico) e nei temi, ma sempre assolutamente contemporanea e reale. In ciò gli appelli alla "datità" e "genuinità" dei materiali caratteristici della linguistica empirica (cfr. fra tutti Sampson 2004: 1 ed in generale Sampson 2001) trovano perfetto riscontro: i post dei newsgroup sono senz'altro «naturally-occurring language texts» nel senso di Sinclair (1991: 171).

È assai interessante, inoltre, (b) che i newsgroup siano organizzati in una struttura gerarchica[11] classificata (anche) tematicamente, ed il fatto che questa tassonomia stessa sia nata dal basso, in base all'iniziativa degli utenti medesimi (ad ulteriore conferma di quella "genuinità naturale" di cui si diceva poc'anzi). Se le gerarchie di base (le "radici") sono nate, infatti, per iniziativa spesso non ascrivibile all'utente singolo, le gerarchie terminali (cioè i vari newsgroup) sono istituite da un utente o da gruppi di utenti. Altrettanto interessante (c) è che accanto alle storiche gerarchie[12] contenutistiche geograficamente non connotate (ma di emanazione statunitense, e comunque in lingua inglese), vi siano delle gerarchie nazionali (ed in lingue diverse), variamente articolate e frequentate. Queste gerarchie nazionali costituiscono una relativa garanzia di uniformità diacorica: è infatti statisticamente probabile che uno che scriva su *uk.comp.security* sia britannico, perché sennò avrebbe più facilmente scritto sull'internazionale *comp.security*.

Quanto fin qui detto giustifica l'idea (d) che UseNet sia una sorta di "enciclopedia popolare", organizzata secondo una "folk taxonomy"[13] (non poi così diversa da

---

[9] «In realtà – con le parole di in Barbera 2007b – se lo si dovesse sostenere legalmente, le cose potrebbero non essere così pacifiche (talvolta si è ricorso ad un cosiddetto "diritto implicito"), ma dato che il comune sentire sostiene comunque la nostra *bonam fidem*, e che non vi sono ad ogni conto interessi rilevanti lesi, è certo assai improbabile che contestazioni significative possano essere sollevate. In effetti sono anni che Google mantiene commercialmente archivi di newsgroup senza che ciò sia avvenuto».

[10] La nozione – come spiego anche in Barbera 2007b – è vetusta, legata soprattutto alle problematiche sorte intorno al cosiddetto "latino volgare" tra i grandi *patres* della romanistica (Löfsted, Mohrman, Hofmann, Spitzer, ...), ma è stata riproposta anche recentemente (Kiesler 2006). L'analogia sembra abbastanza buona, in quanto si tratta, molto in soldoni, di una lingua comune, usuale e media, non tematicamente o sociologicamente delimitabile, più vicina al parlato ma di fatto scritta, e per la quale, in realtà la dicotomia scritto-parlato non è veramente pertinente. Per il fatto che qui (e nelle CMC in genere) «the existing dichotomy speech vs. writing [...] is considered illusory and ineffective» (Allora 2005), cfr. Allora 2005 e Corino 2007: § 1.1.

[11] Avremo, ad esempio, nella gerarchia radice *es*.* (Spagna), le gerarchie tematiche .charla (corrispondente nelle *Big 8* - cfr. n. 12 – a *talk*.*) o *.ciencia* (corrispondente a *sci*.*), con poi ulteriori sottogerarchie, che possono a loro volta essere ulteriormente ramificate; il frammento di tassonomia così esemplificata comprende pertanto i newsgroup *es.charla.actualidad, es.charla.educacion, es.charla.misc, es.charla.educacion.ciencia, es.ciencia, es.ciencia.fisica,* ecc.

[12] UseNet, invero si è sviluppata intorno alle cosiddette *Big 8 hierarchies (comp*.*, misc*.*, news*.*, rec*.*, sci*.*, soc*.*, talk*.* + *humanities*.*),* istituite nel 1987 (ma *humanities*.* è stata aggiunta nel 1995), ed alla loro risposta anarchica ed incontrollata *alt*.*.

[13] Il concetto, che risale probabilmente a Durkheim 1912, è oggi ben studiato soprattutto dal punto di vista biologico ed antropologico (cfr. ad esempio Berlin - Breedlove - Raven 1973 e Healey 1993, con bibliografia).

quelle che si studiano ad es. in linguistica antropologica): donde il forte interesse lessicografico, antropologico e sociologico.

Un ultimo punto (e) interessante, questa volta, testualmente, è il fenomeno del quoting[14], ossia della citazione di (parti di) post, cioè messaggi, precedenti cui si fa riferimento, organizzabili anche in lunghe e ramificate catene (thread).

Se queste caratteristiche possono rendere i newsgroup irresistibili per un costruttore di corpora, non mancano tuttavia anche aspetti negativi, che, al contempo, rendono questi materiali una considerevole sfida: alcune peculiarità linguistiche mediate dal mezzo (abbreviazioni idiosincratiche, gergo informatico, emoticons) sono più o meno agevolmente circoscrivibili (e quindi marcabili e neutralizzabili con script appositi[15]), ma le sporcature interne al testo, dovute a battitura veloce od a problemi di transcodifiche, relative alle classi di testi (crossposting[16]) o pertinenti alla gerarchia tematica (spam, OT "Out of Topic") e linguistica (post in lingue straniere non previste), sono meno facilmente risolvibili, ed impongono la creazione di numerosi script di pulizia[17].

Computazionalmente, però, il problema più rilevante è la presenza di molto testo ripetuto, originato parte dal quoting (quindi inerentemente "buono") e parte dal crossposting ("cattivo"), testo ripetuto che va contenuto entro soglie le più basse possibili, pena la irrilevanza statistica del corpus a fini lessicografici. Per ora nei NUNC l'abbattimento del testo ripetuto è stato di default conseguito solo a caro prezzo, a scàpito dell'integrità dei thread[18]: si è, ossia, scelto solo un messaggio per thread. Abbiamo però talvolta (per ora sperimentalmente solo per l'italiano) approntato anche dei doppioni dei corpora normalmente ridotti, che presentassero invece i thread completi. Tali "doubles", che sono inservibili per ricerche lessicografiche, sono invece assai utili per ricerche testuali, consentendo efficacemente lo studio del quoting e dei problemi ad esso relati (come ad esempio Marello 2007).

2. I NUNC-ES

2.1. Le gerarchie di UseNet di lingua spagnola

Le gerarchie di UseNet che presentano, in misura alquanto variabile, materiali di lingua spagnola sono solo[19] sette (cfr. Tav. 3):

---

[14] Per una prima idea linguistica del fenomeno cfr. Corino 2007 e Marello 2007; significativo, tra l'altro, il proliferare di guide pratiche (stile vecchio manuale di galateo della zia!) che pullulano sul web, e che poche ricerche su Google bastano a scoprire.

[15] Ossia con relativamente semplici listati scritti in un comodo linguaggio interpretato come il Perl.

[16] Il cosiddetto "crossposting" consiste nell'invio del medesimo messaggio a più newsgroup contemporaneamente: e, naturalmente, se il corpus pesca testi da tutti tali newsgroup, avrà anche forti probabilità di prendere più di una volta lo stesso messaggio.

[17] Cosa che per i NUNC si è ovviamente fatta, con risultati statisticamente accettabili, anche se ancora suscettibili di miglioramenti.

[18] Sono però in studio sistemi più raffinati.

[19] Almeno nei grandi newserver a pagamento (Giganews, Newsreader, Supernews, Newshosting, Activenews e Newsfeeds) che si sono usati per fare i download dei NUNC.

ar.*	*NG dell'Argentina*
es.*	*NG di Spagna*
esp.*	*NG in spagnolo*
chile.*	*NG del Cile*
mex.*	*NG messicani*
mx.*	*NG del Messico*
peru.*	*NG del Perù*

Tav. 3: *Le gerarchie di lingua spagnola in UseNet*

Non tutte, però, queste gerarchie sono effettivamente funzionanti: in particolare *peru*.* e *mx*.* risultano sostanzialmente vuote e popolate solo di spam (prevalentemente, per sovramercato, anglofono); e, sempre tra lo spam predominante, poco più traffico registrano anche *ar.*, mex*.* e la gerarchia internazionale *esp*.*;* l'unica gerarchia latino-americana pienamente funzionante sembra essere la cilena, ed a fronte di questa la sola gerarchia con ricca tassonomia e grandi volumi di traffico è quella di Spagna.

Se avevamo argomentato che la presenza di gerarchie nazionali è in genere una garanzia della relativa oggettività e tracciabilità diacorica dei post, in una situazione come la spagnola in cui l'America Latina (Cile escluso) non sembra avere gerarchie vitali, ciò risulta considerevolmente indebolito[20].

2.2. I NUNC-ES attualmente disponibili

Sic stantibus rebus la scelta di trattare solo gruppi di *es*.* (riservandosi semmai di fare un corpus autonomo di *chile*.* quando si fosse raggiunta una quantità di scarichi sufficiente) era inevitabile. Il lavoro è stato portato avanti anche grazie a due tesi di laurea (anno accademico 2003-4: Stefania Morra e Valeria Carretto) ed ha messo capo ai corpora le cui specifiche abbiamo fornito in Tav. 1. Dal novembre 2002 al luglio 2003 la gerarchia *es*.* è stata scaricata al completo da sei newserver (cfr. n. 19) su una macchina dedicata nella sede di Via Piazzi; i newsgroup da cui sono stati selezionati i testi per i corpora sono rappresentati sinotticamente nella Tav. 7 in appendice (i testi del corpus generico sono stati attinti da tutti i newsgroup, quelli degli specialistici solo dai newsgroup attinenti al settore desiderato).

Questi corpora sono stati preparati in formato CQP come gli altri corpora della suite NUNC e sùbito (fin dal 2004) messi a disposizione online; a differenza degli altri NUNC, però, non sono stati immediatamente lemmatizzati e annotati per parte

[20] Difatti alcuni utilizzatori del NUNC-ES Generic non hanno mancato di riscontrare la presenza di sporadici americanismi.

del discorso, cioè POS-taggati, perché il tagger (il Tree Tagger[21] dell'IMS Stuttgart: cfr. Schmid 1994) non disponeva ancora di un *parameter file* per lo spagnolo.

2.3. Alcuni esempi di utilizzo

Anche in questa loro prima versione, abbastanza spartana e certo perfettibile, i NUNC-ES hanno già mostrato la loro utilità, e sono ormai stati usati molte volte.

Volevo qui riportare pochi esempi di ricerche caratteristiche, anche se probabilmente abbastanza banali, più che altro per evidenziare le potenzialità d'uso degli attuali NUNC-ES.

La presenza di tre corpora specialistici accanto al generico consente, ad esempio, ricerche di interesse fraseologico e terminologico, incrociando i risultati tra i diversi corpora.

Un pattern caratteristico è quello che si può esemplificare con la parola *abertura*, che si riscontra con accezioni diverse nei tre specialistici e nel generico, che così si integrano ricomponendo un ideale lemma abbastanza ricco. I dati più significativi sono riportati nella Tav. 4[22]:

NUNC-ES Cooking	
697547	compasión , se las abre por el centro . Esta <abertura> dejará paso a un picadillo bien sazonado , cuyos detalles
926429	. Pensaba que las licuadoras eran las que tenían una <abertura> pequeña para ir metiendo trozos de frutas y hacía zumo
1086285	durante tres meses , mientras dura la fermentación por la <abertura> irá saliendo una espuma espesa que se recogerá en un
1321493	200 1 El pollo que este[23] vaciado , por la <abertura> del vientre le metes un diente de ajo y media

Tav. 4: *Query <"abertura">*

[21] *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html*.

[22] Pur non modificando in nulla i risultati delle query (onde più potessero emergere, oltre che i pregi, anche i difetti residui presenti in questa prima versione dei corpora), per esclusive ragioni di spazio, ho riprodotto solo una parte dei risultati, cercando comunque di dare una campionatura rappresentativa (eliminando, ad es., dal generico tipi che non fossero già presenti negli specialistici); le proporzioni selezione/totale sono le seguenti: 6/6 Cooking, 1/2 Photo, 11/13 Motor e 10/43 Generic.

Il numero di riga, riportato per facilitare gli interventi sulle nuove versioni dei corpora, non compare di default nei risultati delle query online (anche se ne può essere richiesta la visualizzazione con appositi comandi).

[23] L'ortografia corrente nei newsgroup è ingenere più rilassata di quella che lo standar vorrebbe. In particolare, sono largamente tollerate grafie senzo accento (cfr. 1321493 este] esté, 2952860 habra… sera] habrá… será, 4752246 demas] demás, 5232721 deberian] deberían, 9942337 paso] pasó, 15153271 hacian … hacia] hacían … hacía, 1725012 el] él) e non ci si formalizza troppo per gli errori di batti: tuva (cfr. 8364384 dene] debe, 466035 hirbas] hierbas; 111730 elije] elige sarà invece più probabilmente una *spelling pronunciation*.

NUNC-ES Cooking	
1497866	de dentículos , o dientes , que salen de la <abertura> de la boca y raen las rocas y las hojas
1803498	las depositamos en una manga pastelera con una boquilla de <abertura> grande (no rizada , vamos) . Esto es
NUNC-ES Photo	
111730	elije una velocidad baja , para que el tiempo de <abertura> del obturador sea superior al de refresco de la imagen
NUNC-ES Motor	
2952860	de freno no habra que lubricarla , sera en la <abertura> al pulsar el freno) , ni que se le
4096258	moto pero no he visto que hubiese ningún mecanismo de <abertura> . No lo tengo claro Saludos . 12039 <h_From>nospam@hotmail.com (BadMan!) ansi
4752246	pistion en su sitio y los platinos con la justa <abertura> .. por lo demas .... para manuales mira en :
5232721	deberian de ser capaces de cambiar también el ciclo de <abertura> de válvulas de dos de los pistones . 1 Eso
6454552	al punto de partida . ¿ Hacia dónde dejaron la <abertura> del casco en esta maniobra ? ¿ se remolcó con
8062917	pueda lucirse en Monaco . 1 Tienen problemas con la <abertura> de las bancadas de cilindros , actualmente usan una V
8062993	ultra-compacto , pero por lo visto con dicho ángulo de <abertura> las bancadas de cilindros se descompensan y producen demasiadas vibraciones
9089600	reloj (Lotus)que llevo en la mano derecha ejercio una leve <abertura> en el gas y zasssssssssssss,lo pase...jejejeje,a[24] luego vas y lo
9449586	giraran un poco el árbol de levas (controla la <abertura> y cierre de válvulas) y que después el motor
10985964	hace un tiempo , el Golf no me detecta la <abertura> de la puerta del conductor con lo que , al
12181320	\ \ \ Bueno , lógicamente a menos grado de <abertura> . jejeje . El asci-art no es lo mio .
NUNC-ES Generic	
4682553	general de la empresa . - Marketing- estudios sensibilidad - <abertura> nuevos nichos mercado . PLASTIX CHILENA (1989-199) “
6904812	de programa (P automático) A (prioridad de <abertura>) , S (prioridad de velocidad) y M

Tav. 4: *Query <”abertura”>*

[24] Cfr. nota 25]27.

NUNC-ES Generic	
8364069	opuesto (también abierto) pero de menor tamaño su <abertura> , de forma que quede perfectamente ajustado al tamaño de
8364384	porque el corte de la caja mayor dene tener una <abertura> exactamente del tamaño de la pantalla o Tv creo que
9942337	se me paso por la mente que esa era una <abertura> pa cartridge .... On Thu , 26 Feb 2004 09:28:15
13131297	abiertas se tira al suelo y ..... en la primera <abertura> de patas en el piso tipo flashdance , quedo pegada
13835812	la mariposa no estaba absolutamente cerrada (quedaba con una <abertura> de unos 2 mm) .... ¿ es normal ?
15153271	potencia . Pero las mediciones se hacian a angulo de <abertura> de la mariposa constante lo que hacia que el consumo
19651856	al suelo y ..... ¡ ZUK ! en la primera <abertura> de patas en el piso tipo flashdance a poto pelà
20893042	verlo ! . Su pene se había escapado por la <abertura> del bóxer y estaba en todo su esplendor , alto

Tav. 4: *Query <"abertura">*

Caso diverso, ma ugualmente paradigmatico, è ad esempio quello della locuzione *manojo de* 'una manciata di', molto usata in un particolare àmbito, ma non in altri (e che infatti è ben attestata in NUNC-ES Cooking, ma è assente da Photo e Motor); per questo si può mettere a confronto il corpus generico e quello specialistico di cucina, ricavandone una mappa abbastanza chiara dell'uso proprio e delle accezioni traslate. I dati più significativi sono riportati nella Tav. 5[25]:

NUNC-ES Cooking	
5683	. 2 Ensalada a la naranja 1 1 lechuga 1 <manojo de> berros 2 naranjas grandes 1 cebolla El zumo de media
54121	, esta es la " mía " : 1 Un <manojo de> trigueros Un manojo de ajetes Una loncha de jamón serrano
54125	" mía " : 1 Un manojo de trigueros Un <manojo de> ajetes Una loncha de jamón serrano gruesa y cortada a

Tav. 5: *Query <"manojo" "de">*

[25] Valgono i medesimi criteri illustrati nella nota precedente; la proporzioni selezione/totale sono le seguenti: 20/26 Cooking e 8/17 Generic.

NUNC-ES Cooking	
112442	-o mejor , captaron ellas deliberadamente mi atención- a un <manojo de> pencas de berza , supongo que gallegas , rizadas ,
251190	dore removiendo bien . Echar el vino tinto y el <manojo de> hierbas en una cazuela ; calentar despacito a fuego lento
251222	empiece a hervir el vino , echarlo junto con el <manojo de> hierbas sobre las anguilas ; sazonar ; se tapa la
319225	: 4 pimientos rojos grandes 4 berenjenas alargadas . 1 <manojo de> cebolletas 2 dientes de ajo Aceite de oliva virgen Vinagre
319254	uno a uno los pimientos , las berenjenas y el <manojo de> las cebolletas en papel de aluminio . Colocar los paquetes
463164	valga la redundancia ;-) Dos cucharadas de azúcar . 1 <manojo de> menta fresca . 1 cucharada de agua hirviendo . 1/2
466035	, 2 zanahorias , 1 tallo de apio , 1 <manojo de> hirbas aromáticas picadas , 1/2 vaso de leche , 1/2
554105	gr. de queso Idiazabal fresco . 1 lechuga . 1 <manojo de> rabanitos . 1 manojo de berros . Aceite de oliva
554110	. 1 lechuga . 1 manojo de rabanitos . 1 <manojo de> berros . Aceite de oliva refinado . Vinagre de sidra
949686	has de tomar una caldera llena de agua con un <manojo de> pajas de hordio o de heno al suelo de la
1167491	coliflor 200grs de judias verdes Una berenjena Un calabacín Un <manojo de> ajos tiernos 1 cebolla mediana 2 alcachofas 100 grs de
1595703	" baby " . 4 tomates tipo pera . 1 <manojo de> los llamados espárragos trigueros . Unas cuantas judías verdes del
1651538	la Mallorquina Ingredientes(6-8 pers.) 2 manojos de espinacas. 1 __UNDEF__ <manojo de> perejil 4 tomates 1 manojo de cebolletas tiernas "sofrits" 2
1651877	2 latitas de anchoas en aceite ajos, comino en polvo, <manojo de> ajos tiernos aceite, sal y pimienta Preparación Se pican muy
1725012	Ya que uno de ellos trajo a casa un enorme <manojo de> espárragos (según el recogidos en el campo , según
1787940	picados 1 tomate bien picado Pimienta Comino Orégano 1 buen <manojo de> cilantro finamente picado Tacos - tortillas Preparación : Caliente la
1804755	picadito 1 ajete fresco (ajo tierno supongo) 1 <manojo de> canónigos (uf esto me va a costar trabajo encontrarlo

Tav. 5: *Query <"manojo" "de">*

NUNC-ES Generic	
17237094	con escenas de “ El Resplandor “ y con un <manojo de> escenas vuelve un Replicante a Deckard ... Re : The
20670413	voz casi inaudible- Todo el rostro del hacendado era un <manojo de> tics y contracciones incontrolables , a duras penas controló su
20795004	que estaban sueltas , dejándose en la silla un buen <manojo de> pelos , , esto la impulsó a depilarse por completo
21562047	razones de la distancia , y abrir con su metálico <manojo de> dientes el arcón repleto de olvido . Ohé la muerte
22300713	Recuerdan ? Hace a ? os cuando ? ramos un <manojo de> posteadores asiduos, y todos -cual m ? s cual
28114522	. Bastantes también solamente postal , o posiblemente pequeño un <manojo de> 2-3 rosas con el interflora , un spold de 10-15
29804823	los autos , las casas y las calles eran un <manojo de> escombros . La guerra estaba ahí , en su forma
30716870	. De lo demás comentado por aquí respecto a ese <manojo de> células inútiles , nada más queda reseñar . Otro que

Tav. 5: *Query <”manojo” “de”>*

Oltre che per escussioni lessicografiche[26], terminologiche o di fraseologia, i NUNC-ES si sono dimostrati molto utili anche per la didattica; e non sono mancate neppure ricerche su aspetti morfologici o sintattici (cfr. ad es. Bermejo (2007) sulla subordinazione retta da *aconsejar* e Guil - Borreguero Zuloaga (2007) sulla comparazione).

3. RECENTI E FUTURI SVILUPPI DEI NUNC-ES

Nonostante la dimostrata utilità di avere online, liberamente interrogabili, 47.479.718 parole di spagnolo tratto dai newsgroup della gerarchia *es.**, i NUNC-ES sono ancora solo una prima versione, relativamente provvisoria, sottodimensionata e più limitata degli altri NUNC. Alcune, prioritarie, zone di intervento sono già state individuate, ed il lavoro alle nuove versioni è già molto avanzato, sicché non sarà prematuro preannunciarne alcune delle caratteristiche salienti.

[26] Nel progetto FIRB al cui interno questi corpora furono prodotti, d’altra parte, la ricerca lessicografica e terminologica era uno degli obiettivi primari.

3.1. Nuove versioni

Da un punto di vista quantitativo, la dimensione dei corpora sarà aumentata, portando le annate di post scaricate a quattro: dato che la dimensione globale di traffico della gerarchia *es*.* è inferiore a quelle, ad esempio, di *it*.* o *de*.*, per avere corpora comparabili interlinguisticamente, bisognerà per forza rinunciare od alla completa coincidenza dei periodi di scarico, od alla loro comparabilità dimensionale; nella convinzione che non saranno pochi mesi ad alterare significativamente lo status sincronico di una lingua, si è scelto il primo compromesso.

Qualitativamente, saranno corretti gli errori che si sono individuati nelle procedure di tokenizzazione (cfr. ad es. qui la riga 9089600 di *Motor*[27]), di markuppatura[28] (cfr. ad es. l' "UNDEF"[29] alla riga 1651538 di *Cooking*), di selezione del testo ripetuto, e di *language-detecting*. Quest'ultimo argomento merita forse qualche amplificazione, dato che, nonostante un interessante intervento di pochi anni orsono (Grefenstette - Nioche 2000), non è spesso trattato nella bibliografia generale di linguistica dei corpora. Quando si suppone in una collezione di testi, che devono essere solo in una determinata lingua, la presenza di "intrusi" in altre lingue (come ad esempio quando si raccolgono automaticamente materiali dal Web), bisogna predisporre degli appositi filtri (la cui architettura Grefenstette - Nioche (2000) appunto descrivono) che sono lingua-specifici. Filtri che nel nostro caso, per eliminare i post in lingue diverse dallo spagnolo sporadicamente presenti in *es*.*, erano stati predisposti pensando soprattutto all'inglese, che è la lingua internazionale dello spam; è risultato inaspettatamente, però, che esisteva anche qualche post in portoghese, di cui un paio si sono anche infiltrati nel corpus: dovremo così predisporre anche un filtro tarato sul portoghese.

Purtroppo, infine, sull'altra grande questione, quella della minore sicurezza diacorica di *es*.* rispetto, ad esempio, ad *uk*.* o *de*.*, non si può invece fare molto, essendo dovuta ad una asimmetria *in re* delle gerarchie di lingua spagnola, in cui le gerarchie latinoamericane manifestano minore vitalità, ed i loro potenziali utenti finiscono talvolta per usare la gerarchia di Spagna. Siamo però in procinto di preparare[28] un corpus dell'unica gerarchia latinoamericana significativa, *chile*.*, da usare come corpus di confronto (oltre che come interessante oggetto di studio in sé).

---

[27] Dove la stringa <y zasssssssssssss,lo pase...jejejeje,a luego> non è stata correttamente segmentata, certo a causa della anomala struttura della onomatopea "zasssssssssssss".

[28] Molto alla buona, «per markup – secondo scrivevo in Barbera - Corino - Onesti 2007a: § 1 – si intendono tutte le informazioni di carattere in qualche modo "sovrasegmentale" rispetto alla pura successione lineare dei caratteri del testo ed alla loro articolazione in token». Per una caratterizzazione più accurata, cfr. *ibidem*, § 1.4.

[29] Traccia del mancato riconoscimento, da parte della procedura di *encoding* del CQP, di un elemento di markup generato in modo anomalo.

[30] Una tesi di laurea, di Eleonora Bodda, è all'uopo in corso.

3.2. Il tagging

Al di là di quanto detto nel § 3.1, il principale problema dei NUNC-ES è stata la mancanza di tagging, laddove tutti gli altri corpora della suite NUNC sono POS-taggati e lemmatizzati. Ed infatti è su questo punto che abbiamo concentrato i nostri sforzi, con risultati, attualmente sotto testing, che credo significativi: un innovativo tagset adatto per la annotazione stocastica e per il confronto interlinguistico, un nuovo file di parametri per il Tree Tagger, e nuovi corpora annotati. Presentare nei dettagli questo lavoro, frutto di molte sinergie[31], sarà còmpito di un contributo separato, cui stanno attendendo M. Barbera, M. Borreguero Zuloaga e M. Tomatis, ma se ne possono già anticipare qui le caratteristiche principali.

Innanzitutto bisogna premettere che uno degli scopi che il gruppo di ricerca di Torino (segnatamente per il progetto FIRB) si riprometteva di conseguire era quello di produrre uno schema di annotazione, costruito secondo gerarchie tipate[32] in base alle raccomandazioni EAGLES[33], che fosse facilmente rimappabile e che consentisse di effettuare query su più lingue usando una analoga maschera di ricerca. I corpora più direttamente nell'obbiettivo erano ovviamente i NUNC, suite multilingue per eccellenza, ma gli studi per raggiungere questo ideale erano, al solito, partiti da quella che è la nostra palestra sperimentale per eccellenza (cfr. Barbera 2007b: § 2.2.1), il CT (*Corpus Taurinense*) di italiano antico, con risultati presentati al Convegno SILFI 2000 (ora Barbera 2007c). Nell'attesa di perfezionare questo schema interlinguistico (e di preparare i vari file di parametri per il tagger a questo deputati), avevamo però iniziato a produrre corpora (come ad esempio i NUNC attualmente in rete) annotati con gli schemi al momento già disponibili per il TreeTagger. La assenza di tali strumenti per lo spagnolo indicava chiaramente come quella fosse la lingua su cui concentrare le energie per iniziare ad esportare lo "schema-CT".

In breve, i lavori, che furono avviati dalla tesi di laurea di Giovanna Brino, produssero un mapping dei tagset più diffusi, CRATER e IULA, articolatissimi (500+ tag!) ed adatti al funzionamento solo con grammatiche di microregole, ed una prima ipotesi del loro riversamento in una struttura "CT-like" numericamente (70- tag) adatta al funzionamento con tagger stocastici. I lavori erano in questa fase quando furono rilasciati (liberi sotto licenza GNU) da Achim Stein un file di parametri ed un training corpus di spagnolo basati su un radicale "disboscamento" del CRATER: materiali che furono utilissimi alla nostra officina. Attraverso un rimappaggio di tutti i sistemi e risorse disponibili, nell'inverno 2005 si giunse così ad un tagset operativo che fu testato su un primo microcorpus sperimentale di 8000 parole.

---

[31] Coordinato da Manuel Barbera, è iniziato con la tesi di laurea di Giovanna Brino, 2004-5, e si è giovato soprattutto della preziosa opera informatica di Marco Tomatis e dell'essenziale supervisione e testing di Margarita Borreguero Zuloaga.

[32] Per la nozione di gerarchia tipata cfr. Barbera 2007c: § 3.

[33] *Expert Advisory Group on Language Engineering Standards* (http://www.ilc.cnr.it/EAGLES96/home.html), iniziativa ora proseguita da ISLE.

A partire da questi risultati, attraverso varie fasi di ricorrezione dei corpora campione, aggiustamento di dettagli del tagset, acquisizione di formari da varie fonti, si è giunti ai risultati che dicevamo, che saranno presto illustrati e resi disponibili, ma di cui qui anticipiamo, in forma riassuntiva e senza commenti, il risultato più cospicuo: il nuovo tagset di 62 tag. La versione 1.1. è presentata nella la tavola qui di seguito:

codice	*tag*	*esempio*	*codice*	*tag*	*esempio*
20	n.c	perro	116	v.m.f.sub.im	quitaran
21	n.p	Juan	119	v.m.f.sub.fu	tuviere
26	adj	bonito	117	v.m.f.cnd.pr	entraría
30	pd.dem	este	118	v.m.f.imp.pr	vete
32	pd.ind	alguno	121	v.m.n.i.pr	emplear
33	pd.pos	mi	123	v.m.n.p.pa	saludado
35	pd.int	qué	124	v.m.n.g.pr	hablando
36	pd.rel	que	211	v.a.f.ind.pr	es
37	pd.per.s.n	yo	212	v.a.f.ind.im	era
38	pd.per.s.o	mí	213	v.a.f.ind.pa	fue
39	pd.per.w	me	214	v.a.f.ind.fu	será
40	pd.exc	qué	215	v.a.f.sub.pr	sean
45	adv	bien	216	v.a.f.sub.im	hubiese
50	con.c	y	219	v.a.f.sub.fu	hubiere
51	con.s	que	217	v.a.f.cnd.pr	estaríamos
56	adp.pre.p	de	218	v.a.f.imp.pr	ten
58	adp.pre.art	del	221	v.a.n.i.pr	ser
60	art.d	el	223	v.a.n.p.pa	sido
61	art.i	uno	224	v.a.n.g.pr	siendo
64	num.c	dos	311	v.d.f.ind.pr	puedo
65	num.o	tercero	312	v.d.f.ind.im	debía
68	intj	ay	313	v.d.f.ind.pa	pudo
70	pun.f	.	314	v.d.f.ind.fu	deberemos
71	pun.n	,	315	v.d.f.sub.pr	pueda
75	r.frg	song	316	v.d.f.sub.im	pudiera
77	r.for	2?2=4	319	v.d.f.sub.fu	pudiere

Tav. 6: *Il nuovo tagset per lo spagnolo*

codice	*tag*	*esempio*	*codice*	*tag*	*esempio*
111	v.m.f.ind.pr	sale	317	v.d.f.cnd.pr	deberían
112	v.m.f.ind.im	estaban	321	v.d.f.imp.pr	debe
113	v.m.f.ind.pa	ocurrió	321	v.d.n.i.pr	poder
114	v.m.f.ind.fu	buscaré	323	v.d.n.p.pa	podido
115	v.m.f.sub.pr	cuezan	324	v.d.n.g.pr	pudiendo

Tav. 6: *Il nuovo tagset per lo spagnolo*

Al di là della menzionata struttura gerarchica dei tag, e della presenza nello schema delle consuete parti del discorso, si noterà la marcatura unitaria di pronomi e determinanti (giusta le considerazioni illustrate in Barbera 2002), la presenza di tag per le parole straniere ed espressioni matematiche, e l'elevata batteria di tag disponibili per il sistema verbale (pure rinunciando alla marca delle forme composte).

Le "etichette" (*labels*), cioè i nomi attribuiti ai tag, sono state formulate in modo da essere analoghe a quelle usate negli altri tagset (italiano, francese, ecc.), all'insegna della massima universalità interlinguistica delle query, anche a costo di scostarsi talvolta dalla tradizione nomenclatoria ispanica; la decodifica dovrebbe essere abbastanza intuitiva, una volta giusto glossate le etichette più sintetiche, come v.m.f "verbo lessicale (*main*) di modo finito", v.d.n.p "verbo modale di modo non finito, participio", e così via[34].

4. APPENDICE: I NEWSGROUP DELLA GERARCHIA es.*

Come precedentemente accennato (§ 2.2) ci pare opportuno riportare nella sua integrità la lista dei newsgroup della gerarchia es.* da cui sono stati tratti i testi per i corpora NUNC-ES, per consentire al lettore di meglio valutare la sua articolazione (tassonomia) e la ricchezza di tematiche ricoperte (e quindi anche la loro potenzialità per la creazione di lessici specialistici). Dal nòvero sono stati esclusi solo i newsgroup vuoti ed i doppioni creati erroneamente su qualche server (come ad es. es.misc.anuncios.trabajos per es.misc.anuncios.trabajo), specie se popolati (come consueto) prevalentemente da spam.

es.comp.sistemas.hp48	es.alt.anuncios.trabajo.demandas
es.alt.anuncios.compra-venta	es.comp.sistemas.misc
es.comp.sistemas.inteligentes	es.alt.anuncios.trabajo.ofertas

Tav. 7: *I newsgroup della gerarchia* es.*

[34] Una ulteriore riflessione sulle labels è attuata in Barbera 2007d, che mette capo ad una nuova versione (1.2), immutata nei tag ma parzialmente rinominata nelle labels, del tagset spagnolo.

es.comp.sistemas.pc	es.news.anuncios
es.alt.anuncios.trabajo	es.charla.enfermedad.ela
es.comp.sistemas.sinclair	es.news.grupos
es.alt.anuncios	es.charla.enfermedad.misc
es.comp.virus	es.news.misc
es.alt.chistes	es.charla.enfermedad
es.compra-venta	es.news.preguntas
es.alt.sexo.relatos	es.charla.enfermeria
es.eunet.spanish-tex	es.rec.aviacion
es.charla.actualidad	es.charla.gastronomia
es.humanidades.anuncios	es.rec.bricolaje
es.charla.conexion.misc[35]	es.charla.medio.ambiente
es.humanidades.arte	es.rec.cine-en-casa
es.charla.conexion.tarifa.plana	es.charla.misc
es.humanidades.derecho	es.rec.cine
es.charla.conexion	es.charla.moteros
es.humanidades.filosofia	es.rec.colecionismo
es.charla.cooperacion	es.charla.motor
es.humanidades.gramatica	es.rec.comics
es.charla.economia.bolsa	es.charla.politica.izquierdaunida
es.humanidades.literatura	es.rec.deportes.atletismo
es.charla.economia.contabilidad	es.charla.politica.misc
es.humanidades.misc	es.rec.deportes.aventura
es.charla.economia.misc	es.charla.religion
es.humanidades.psicologia	es.rec.deportes.baloncesto
es.charla.educacion.ciencia	es.charla.sexo
es.misc.admin	es.rec.deportes.buceo
es.charla.educacion.distancia	es.charla.sindical
es.misc.anuncios.compra-venta	es.rec.deportes.esqui
es.charla.educacion.drogas	es.charla.utopia
es.misc.anuncios.misc	es.rec.deportes.futbol
es.charla.educacion.educ-fisica	es.ciencia.astrofisica.misc
es.misc.anuncios.trabajo.demandas	es.rec.deportes.kayak
es.charla.educacion.misc	es.ciencia.astrofisica.telescopios
es.misc.anuncios.trabajo.misc	es.rec.deportes.misc
es.charla.educacion.trafico	es.ciencia.astrofisica
es.misc.anuncios.trabajo.ofertas	es.rec.deportes.motor
es.charla.educacion	es.ciencia.electronica.micros
es.misc.anuncios.trabajo	es.rec.deportes.mountain-bike
es.charla.enfermedad.anorex-bulimia	es.ciencia.electronica.misc
es.misc.miscs	es.rec.deportes.natacion
es.charla.enfermedad.cancer	es.ciencia.electronica
es.misc.publicidad	es.rec.deportes.nautica
es.charla.enfermedad.diabetes	es.ciencia.enologia

Tav. 7: *I newsgroup della gerarchia* es.*

[35] Per convenzione internazionale i nomi delle gerarchie UseNet sono sempre senza accenti o caratteri speciali (possono, ossia, contenere solo i codici 033-0126 del charget ASCII).

es.rec.deportes.parapente	es.comp.hackers
es.ciencia.fisica	es.rec.mascotas.gatos
es.rec.deportes.pesca.submarina	es.comp.hardware.misc
es.ciencia.marketing	es.rec.mascotas.misc
es.rec.deportes	es.comp.hardware
es.ciencia.martematicas	es.rec.mascotas.peces
es.rec.ficcion.misc	es.comp.infografia
es.ciencia.medicina.depresion	es.rec.mascotas.perros
es.rec.fotografia	es.comp.infosistemas.bbs
es.ciencia.medicina.lab-clinico	es.rec.misc
es.rec.humor	es.comp.infosistemas.internet
es.ciencia.medicina.misc	es.rec.modelismo
es.rec.ilusionismo	es.comp.infosistemas.misc
es.ciencia.medicina	es.rec.motor
es.rec.jardineria.bonsai	es.comp.infosistemaswww.misc
es.ciencia.meteorologia	es.rec.musica.blues
es.rec.juegos.ajedrez	es.comp.infosistemaswww.paginas-web
es.ciencia.misc	es.rec.musica.clasica
es.rec.juegos.comp.arcade	es.comp.infosistemaswww
es.ciencia.quimicas	es.rec.musica.grupos.beatles
es.rec.juegos.comp.misc	es.comp.infosistemas
es.ciencia	es.rec.musica.grupos.misc
es.rec.juegos.comp.simuladores.misc	es.comp.ingenieria.software
es.comp.amiga	es.rec.musica.jazz
es.rec.juegos.comp.simuladores.vuelo	es.comp.lenguagea.c++
es.comp.artes-graficas	es.rec.musica.partituras
es.rec.juegos.estrategia	es.comp.lenguagea.c
es.comp.bd.misc	es.rec.musica.techno
es.rec.juegos.magic	es.comp.lenguages.clipper[36]
es.comp.bd.ms-access	es.rec.musica
es.rec.juegos.misc	es.comp.lenguages.delphi
es.comp.cad.autocad	es.rec.naturismo
es.rec.juegos.pinball	es.comp.lenguages.java
es.comp.cad.misc	es.rec.radio.amateur
es.rec.juegos.rol	es.comp.lenguages.misc
es.comp.cd-rw	es.rec.radio.misc
es.rec.juegos	es.comp.lenguages.php
es.comp.cracks	es.rec.radio.ondacorta
es.rec.labores	es.comp.lenguages.visual-basic
es.comp.demos	es.rec.radio
es.rec.manga	es.comp.macintosh.misc
es.comp.emuladores	es.rec.trenes
es.rec.mascotas.exoticas	es.comp.misc

Tav. 7: *I newsgroup della gerarchia* es.*

[36] I nomi delle gerarchie sono generati automaticamente da ogni newsserver della rete Usenet; può cosí succedere che siano generate grafie scorrette (cfr. lenguagea e lenguages per lenguajes), che talora si "fissano" (ossia gli utenti vi portano) come newsgroup veri e propri.

es.rec.tv.concursos	es.soc.cultura.misc
es.comp.msx	es.comp.os.ms-windows.misc
es.rec.tv.decodificacion	es.soc.cultura.sin-tabaco
es.comp.neuronal	es.comp.os.ms-windows.programacion
es.rec.tv.misc	es.soc.cultura.teatro
es.comp.os.as400	es.comp.os.ms-windows
es.rec.tv.misc	es.soc.misc
es.comp.os.linux.anuncios	es.comp.programas
es.rec.viajes	es.soc.org.policia
es.comp.os.linux.instalacion	es.comp.redes.adsl
es.rec.video.dvd	es.tecnica.arquitectura
es.comp.os.linux.misc	es.comp.redes.miscl
es.rec.video.edition	es.tecnica.automatica
es.comp.os.linux.programacion	es.comp.seguridad.misc
es.rec.video.misc	es.tecnica.ingenieria.teleco
es.comp.os.linux.redes	es.comp.seguridad.pgp
es.soc.consumidor	es.tecnica.redes.telefonia.movil
es.comp.os.linux	es.comp.seguridad.so
es.soc.cultura.agenda	es.tecnica.sonido
es.comp.os.misc	es.viajes

Tav. 7: *I newsgroup della gerarchia* es.*

RIFERIMENTI BIBLIOGRAFICI

ALLORA, A. (2005): *A Tentative Typology of Net Mediated Communication*, comunicazione presentata alla *Corpus Linguistics Conference, Birmingham July 14-17 2005*, disponibile online alla pagina http://www.corpus.bham.ac.uk/ PCLC/

BARBERA, M. (2002): *Pronomi e determinanti nell'annotazione dell'italiano antico. La POS "PD" del Corpus Taurinense*, in BAUER - GOEBL 2002, pp. 35-52.

BARBERA, M. (2004): *Il progetto FIRB. Stato dei lavori*, documento interno inedito, Ver. 7 aggiornata al febbraio 2004.

BARBERA, M. (2007a): *La resa dei forestierismi in italiano: breve nota ortografica*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa, pp. XV-XVI.

BARBERA, M. (2007b): *Per la storia di un gruppo di ricerca: da bmanuel.org a corpora.unito.it*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa, pp. 3-20.

BARBERA, M. (2007c): *Un tagset per il Corpus Taurinense: italiano antico e linguistica dei corpora*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa.

BARBERA, M. (2007d): *Mapping dei tagset in bmanuel.orgl corpora.unito.it tra guidelines e prolegomeni,* in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa, pp. 373-388.

BARBERA, M. - CORINO, E. - ONESTI, C. (eds) (2007a): *Corpora e linguistica in rete*, Perugia, edizioni Guerra, in corso di stampa, pp. 25-88.

BARBERA, M. - CORINO, E. - ONESTI, C. (eds) (2007b): *Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa, pp. 25-88.

BARBERA, M. - MARELLO, C. (2003 *i.s.*): *Corpo a corpo con l'inglese della* corpus linguistics*, anzi, della linguistica dei corpora*, in *Atti del Convegno Internazionale Lingua italiana e scienze, Firenze, Accademia della Crusca 6-8 febbraio 2003*, in corso di stampa.

BAUER, R. - GOEBL, H. (eds.) (2002): *Parallela IX. Testo - variazione - informatica | Text - Variation - Informatik. Atti del IX Incontro italo-austriaco dei linguisti (Salisburgo, 1-4 novembre 2000) | Akten des IX Österreichisch-italienischen Linguistentreffens (Salzburg, 1.-4. November 2000)*, Wilhelmsfeld, Gottfied Egert, "Pro Lingua" 35.

BERMEJO, F. (2007): Consigliare / aconsejar *e le subordinate esplicite o implicite. Analisi contrastiva nei NUNC generici*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa, pp. 297-308.

BERLIN, B. - BREEDLOVE, D. E. - RAVEN, P. H. (1973) *General principles of classification and nomenclature in folk biology*, in "American Anthropologist", 7, 214-242.

BRINO, G. (2006): *Problemi morfologici nell'etichettatura morfosintattica dello spagnolo. Strategie e procedure,* Tesi di Laurea, Facoltà di lingue e letterature straniere Università di Torino 2004-2005.

CABRÉ, M. T. - MOREL, J. - TORNER, S. - VIVALDI, J, - DE YZAGUIRRE, L. (1998): *El corpus de l'IULA: etiquetaris*, Barcelona, Universitat Pompeu Fabra. Institut Universitari de Lingüistica Aplicada, "Sèrie Informes" 18; disponibile anche online come IULA/INF018/98: http://www. iula.upf.es/ paps1ca.htm

CARRETTO, V. (2005): *Corpora tecnici in lingua spagnola: allestimento di tre corpora specialistici consultabili in rete,* Tesi di Laurea Facoltà di lingue e letterature straniere Università di Torino 2004-2005.

CHRIST, O. - SCHULZE, B. M. (1996): *CWB. Corpus Work Bench, Ein flexibles und modulares Anfragesystem für Textcorpora*, in FELDWEG - HINRICHS: 1996, online a http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/Papers/christ+schulze:tuebingen.94.ps.gz.

CORINO, E. (2007): *NUNC (est disputandum). Questioni metodologiche ed aspetti della testualità*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa, pp. 225-252.

DURKHEIM, É. (1912): *Les formes élémentaires de la vie religieuse: le système totémique en Australie*, Paris, F. Alkan. [riedizione moderna: Paris, PUF, 2003 "Quadrige"].

FELDWEG, H. - HINRICHS, E. W. (eds) (1996): *Lexikon und Text: wiederverwendbare Methoden und Ressourcen zur linguistischen Erschließung des Deutschen*, Tübingen, Max Niemeyer Verlag, "Lexicographica. Series maior" 73.

GREFENSTETTE, G. - NIOCHE, J. (2000): *Estimation of English and non-English Language Use on the WWW*, in *Proceedings of RIAO 2000, 6th Conference: Content-Based Multimedia Information Access, Paris, April 12-14, 2000*, Paris,

Collège de France, pp. 237-246, disponibile online come Arxiv preprint cs.CL/0006032 all'URL http://arxiv.org/ftp/cs/papers/0006/0006032.pdf.

GUIL, P. - BORREGUERO ZULOAGA, M. (2007): *Comparative prototipiche in italiano e spagnolo: I NUNC come base per l'analisi contrastiva*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa.

HEALEY, C. (1993): *Folk Taxonomy and Mythology of Birds of Paradise in the New Guinea Highlands*, in "Ethnology", Vol. XXXII,

HEID, U. (2007): *Il Corpus WorkBench come strumento per la linguistica dei corpora. Princip ed applicazioni*, in BARBERA - CORINO - ONESTI: 2007a, in corso di stampa, pp. 89-108.

KIESLER, R. (2006): *Einführung in die Problematik des Vulgärlateins*, Tübingen, Niemeyer.

MARELLO, C. (2007): «Does Newsgroups "Quoting" Kills or Enhances Other Types of Anaphors?», in KORZEN, I. LUNDQUIST L. (eds.): *Comparing Anaphors Between Sentences, Texts and Languages,* Frederiksberg, Samfundslitteratur Press, *Copenhague Studies in Language*, 34, pp. 145-157.

MORRA, S. (2005): *Corpora tecnici in lingua spagnola: allestimento di un corpus su protocollo web*, Tesi di Laurea Facoltà di lingue e letterature straniere Università di Torino 2004-2005.

SAMPSON, G. (2001): *Empirical Linguistics*, London - New York, Continuum "Open Linguistics"

SAMPSON, G. (2004): *Introduction* to Sampson - McCarthy 2004, pp. 1-8.

SAMPSON, G. - MCCARTHY, D. (eds) (2004): *Corpus Linguistics. Readings in a Widening Discipline*, London - New York, Continuum.

SÁNCHEZ LEÓN, F. (1994): *Spanish tagset for the CRATER Project*, PDF file, Doc. id. CRATER/WP6/FR1, March 7, 1994; disponibile online come Arxiv eprint arXiv:cmp-lg/9406023 v1 alla pagina http://arxiv.org/abs/cmp-lg/ 9406023.

SÁNCHEZ LEÓN, F. - NIETO SERRANO, A. F. (1995): *Development of a Spanish Version of the Xerox Tagger*, PDF file, Doc. id. CRATER/WP6/FR1, May 19, 1995; disponibile online come Arxiv eprint alla pagina http://arxiv.org/ abs/cmp-lg/9505035

SCHMID, H. (1994): *Probabilistic Part-of-Speech Tagging Using Decision Trees*, paper presented at the *International Conference on New Methods in Language Processing, Manchester (UK): 1994;* versione revisionata PS/PDF online sul sito dell'IMS Stuttgart: http://www.ims.uni-stuttgart.de/projekte/corplex/ TreeTagger/DecisionTreeTagger.html.

SINCLAIR, J. MCHARDY (1991): *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.