Web as Corpus : osservazioni preliminari e prospettive di un mega-corpus italiano

Autor(en): **Gerstenberg, Annette**

Objekttyp: **Article**

Zeitschrift: **Revue de linguistique romane**

Band (Jahr): **74 (2010)**

Heft 295-296

PDF erstellt am: **08.08.2024**

Persistenter Link: https://doi.org/10.5169/seals-781700

Web as Corpus: osservazioni preliminari e prospettive di un mega-corpus italiano *

Il web come risorsa linguistica

L'interesse della linguistica per le risorse del web come immenso repertorio di dati linguistici è motivato principalmente da due fattori: da una parte il web sembra essere uno specchio autentico degli sviluppi anche più recenti della lingua contemporanea e delle sue varietà sia speciali sia colloquiali; dall'altra si tratta di materiale apparentemente già preparato per la ricerca linguistica, grazie alla sua natura digitalizzata.

La rete può fornire ricchi materiali per una molteplicità di problemi linguistici: p.es. l'osservazione degli effetti del contatto linguistico[1], l'integrazione di neologismi e prestiti[2], la concorrenza tra grafie diverse[3], l'emergere di nuove forme o strutture grammaticali[4] e la rilevanza pragmatica delle nuove condizioni per la nascita di nuovi stili di comunicazione. Queste prospettive sono un buon motivo per portare avanti i progetti di esplorazione scientifica

---

* Ringrazio Narno Pinotti per la sua attenta revisione non solo linguistica dell'articolo.

1 Cf. gli esempi di code-switching in Berruto (2005, 150).

2 Per il contributo delle ricerche su corpora alla lessicografia cf. Heid (2008). Nella documentazione di esempi e di collocazioni dei dizionari italiani attuali non si integrano attestazioni da corpora (Schafroth 2009, 81).

3 Cf., per una discussione critica di una ricerca condotta con Google per stimare l'accettazione delle recenti riforme ortografiche del francese, del rumeno e del tedesco, Ernst (2009).

4 Cf. per questioni di morfologia francese: «The Web has thus been successfully used to discover words previously considered to be unlikely or even theoretically impossible to coin. For example, one can find the prefix *anti-* attached to simple (non constructed) adjectives, such as in *anti-triste* 'anti-sad' or *anti-obèse* 'anti-obese', and even to adjectives following the *V-able* scheme, such as *anti-inflammable* 'non-flammable'. This latter case was previously declared as theoretically impossible» (Hathout / Montermini / Tanguy 2008, 68).

del linguaggio della rete. A questo scopo, si devono affrontare problemi di preparazione dei dati tratti da internet, in modo che l'esplorazione corrisponda a standard scientifici.

L'uso di motori di ricerca come Google permette un primo orientamento riguardo la frequenza di forme nuove; ma ne sono ben noti i suoi difetti, dovuti alle restrizioni della possibile formulazione di richieste e alle cifre stesse, in quanto viene conteggiata solo un'occorrenza per pagina[5]. Un altro problema, più fondamentale, concerne la questione, ancora aperta, su quali tipi di uso caratterizzino la massa di fatti linguistici nella rete.

In internet si trovano molti documenti che duplicano testi pubblicati su carta o che ne presentano una trasformazione più o meno marcata. In più, internet è molto dinamico nella produzione di nuovi tipi di testi, parallelamente allo sviluppo sempre in corso di nuove tecnologie. Nella discussione di una tipologia testuale adatta alle condizioni di internet è stato introdotto il termine di *web genre* o *cybergenre*[6]. Esistono vari approcci al problema della classificazione dei *web genres* (che comportano anche per il ricevente l'esigenza di un orientamento per ciascun tipo di testo[7]), secondo aspetti pragmatici e linguistico-testuali[8] ovvero più secondo la prospettiva della linguistica computazionale[9]. Le tipologie di *web genres* propongono inventari molto diversi: in una discussione su un sito linguistico si è creato tramite un lavoro collettivo un elenco di 87 *web genre classes* (WebGenreWiki, versione 26.03.2009). Altri studi fondamentali propongono una distinzione tra quattro tipi di base[10]. Si tratta quindi prevalentemente di criteri para- o extra-linguistici, che non trattano dei criteri interni che possono contraddistinguere i vari *web genres*; i tratti propriamente linguistici sono considerati solo in modo molto generale[11].

---

5 Bergh / Zanchetta (2008, 316); Kilgariff (2007) critica il lavoro con i risultati di motori di ricerca, riferendosi alla sua scarsa trasparenza e scientificità con il motto polemico *Googleology is bad Science.*

6 Cf. la terminologia di Sheperd / Watters (1998).

7 Cf. Storrer (2001, 45) che descrive lo stabilirsi di rapporti di coerenza in siti web.

8 Questo aspetto fa parte della recente discussione su nuovi generi testuali del web o sulla trasformazione di generi già esistenti; Jakobs (2003); i contributi in Döring / Osthus / Polzin-Haumann (2004); Giltrow / Stein (2009).

9 Mehler / Gleim (2006) analizzano la strutturazione dei documenti che interpretano in corrispondenza con la funzione del web genre rispettivo; Santini (2007) propone un modello di classificazione automatica sulla base della frequenza e della distribuzione p.es. di function words, POS-tags e HTML-tags.

10 Rehm *et al.* (2008) propongono: *conference website, personal academic, project, city website.*

11 «It comes as no surprise that studies of the linguistic markers of Internet genres are few and far between. It is possible that no one would argue, as is certainly not

Ciononostante ci pare importante non perdere di vista la possibilità di stabilire relazioni tra i tipi di testo presenti in internet, la loro combinazione in una pagina web[12], e le loro particolarità linguistiche, dal punto di vista della qualità e quantità dei fenomeni da evidenziare.

Una dinamica linguistica specifica di internet è oggetto degli studi sulla *Computer Mediated Communication* (CMC)[13]. L'evoluzione di nuovi mezzi linguistici, non solo grafici, ma anche morfologici, semantico-lessicali, sintattici e pragmatico-testuali è fortemente influenzata dallo sviluppo dei modi di comunicazione in internet. I tipi di siti web o spazi internet che presentano l'oggetto di ricerca preferito degli studi sulla CMC (o CMO in francese) sono quelli di chat, liste e forum di discussione, blog e piattaforme di giochi, posta elettronica; anche i messaggi SMS ne fanno parte[14]. Le ricerche romanistiche sulla CMC si dedicano molto spesso alla sua caratterizzazione ibrida tra scritto e parlato (imitato)[15]. Nel contesto della lingua del web si parla anche di *scrittura secondaria*[16]. Questo termine allude al programma di ricerca sulla lingua scritta e parlata, e pertanto a una (possibile) regolarità degli usi innovativi e non-standard della scrittura in rete. Elementi caratteristici di una tale scrittura in rete sono la tolleranza verso errori tipografici e brachilogie o verso l'uso di

---

true for traditional genres, that Internet genres could be ‹recognized› by their linguistic characteristics. Features like emoticons :) and turn-taking strategies in chat may spread across many internet genres and simply indicate ‹internet› rather than anything specific enough to infer genre, – if the genre identity were not already established by pre-signals» (Giltrow / Stein 2009, 11).

12 «[A] web page can be considered as a sort of container of multiple texts» (Santini 2007, 6).

13 ‹‹Computer-Mediated Communication (CMC)› is a research field that explores the social, communicative and linguistic impact of communication technologies, which have continually evolved in connection with the use of computer networks (esp. the Internet)» (Beißwenger / Storrer 2008, 292).

14 Cf. p.es. Gadet (2008, 513); oggetto della descrizione del *parlar spedito* di Pistolesi (2004) sono le chat, gli sms e le e-mail. Per la chat italiana cf. Gerstenberg (2004) e per *le français tchaté* Pierozak (2003).

15 Cf. Gadet (2008, 513) che discute i fenomeni della CMO (non solo) francese: «la communication médiée par ordinateur (désormais CMO), comme une source de questionnements pour la discipline, prenant place parmi les effets de nouvelles méthodes liées à la possibilité de faire reposer la réflexion sur de grands corpus. [...] [s]ouvent qualifiée d'oral dans l'écrit (*écrit contaminé par l'oral*, *oralité par écrit*, *parler lisible*, *hybridation oral/écrit*)».

16 Cf. l'argomentazione di Pistolesi (2004, 31) «è proprio il dominio dell'oralità secondaria a deformare il codice scritto in direzione della voce e a ispirare le strategie che mirano a reintrodurre la fisicità dell'atto linguistico nel testo scritto», che ricalca l'espressione di Walter Ong (*secondary orality*). Schmitz (2009) parla di una *sekundäre Schriftlichkeit*.

abbreviazioni vere e proprie. In molti aspetti la ricerca sulla CMC riprende da vicino gli approcci al linguaggio giovanile, soprattutto per l'aspetto innovativo e ludico (Pierozak 2000), perché il mezzo di comunicazione permette la dialogicità e l'uso dei tratti tipici della CMC, tra cui i tratti con valore pragmatico come i fatismi[17]. Viceversa, i grafismi tipici della scrittura elettronica si ritrovano in temi di studenti[18].

Contrariamente ai testi dei giornali, nel trattamento computazionale internet rappresenta una risorsa linguistica problematica. Solo parzialmente possono essere utilizzate le esperienze della linguistica dei corpora, che si occupa da più di due decenni del problema di sistemare dati testuali di origine anche eterogenea in modo coerente e predisposto per trovare risposte a ricerche scientifiche[19]. Il corpus della lingua italiana di maggiori dimensioni è attualmente rappresentato dal corpus *la Repubblica* (di seguito: REPUBBLICA), con 380 milioni di parole grafiche (token)[20].

La metodologia con cui esplorare sistematicamente il web e i suoi dati particolari costituisce il programma di ricerca del Web as Corpus. In questo contesto si possono distinguere differenti approcci, basati su concezioni diverse di corpus. L'uno segue una definizione molto ampia di corpus come «a collection of texts when considered as an object of language or literary study» (Kilgarrif / Grefenstette 2003, 2). Conseguentemente, oggetto di studio viene considerato l'insieme dei fatti linguistici della rete, che si cerca di

---

17 «Rendendo dialogica la scrittura, le nuove tecnologie permettono lo scambio di ruoli fra emittente e ricevente e supportano un feedback quasi sincrono. Alla dialogicità si possono ricondurre: i fatismi, che verificano l'apertura del canale comunicativo; i segnali discorsivi, che surrogano la mimica e i tratti soprasegmentali, oltre a modulare l'enunciato e a funzionare da congiunzioni testuali; i costrutti marcati, connessi alla gestione del topic nelle sequenze domanda-risposta; l'implicitezza, propria della turnazione, presente in diversa misura in tutti sistemi considerati» (Pistolesi 2004, 29).

18 Lorenzetti / Schirru (2006, 97) ipotizzano: «non è impossibile che questi rappresentino l'avamposto di un futuro arricchimento delle norme grafiche della nostra lingua».

19 Cf. per considerazioni generali Habert / Nazarenko / Salem (1997); Tognini-Bonelli (2001) e i contributi in Wynne (2005); per l'italiano cf. Chiari (2005); per una veduta d'insieme romanistica Grands Corpus (1999) et Corpus (2007); il corpus di riferimento della lingua parlata (francese, italiana, spagnola) è presentato in Cresti / Moneglia (2005); per i corpora in rete cf. i contributi presentati in Barbera / Corino / Onesti (2007); lo state of the art delle sottodiscipline romanistiche è presentato nei congressi *Corpora Romanica* di Friburgo; per aspetti metodologici Gerstenberg (2009).

20 Cf. per la descrizione del corpus, le procedure della sua elaborazione e per dettagli tecnici Baroni *et al.* (2004); Aston / Piccioni (2004).

esplorare nel modo più ampio possibile, creando motori di ricerca sempre più efficaci. Per rendere sistematici gli usi attuali del web come corpus sono stati individuati diversi tipi di corpora, a seconda che si siano indagati una parte del web o l'insieme dei dati disponibili, e a seconda del modo in cui i dati sono trattati: in forma di corpus vero e proprio o in forma di concordanze (ricerche KWiC). Un approccio letteralmente più conservatore – perché fornisce anche un deposito di dati attuali, possibile materiale per future ricerche sulla diacronia del linguaggio web – consiste nello scegliere dati da internet e nel costruire con essi un «mega-corpus»[21].

Anche questo approccio si integra nel programma Web as Corpus, che si potrebbe definire con più precisione Web for Corpus (Bergh / Zanchetta 2008, 315). In questa prospettiva, la ricerca è orientata alla raccolta di archivi in internet per stabilire un corpus chiuso, quasi un'istantanea del flusso sempre in movimento dei dati linguistici nel web. È questo l'approccio presentato qui di seguito, sulla base dell'esempio italiano di questo tipo di corpus, costruito a Bologna. Il corpus in questione è itWaC (*italian Web as Corpus*) e presenta una raccolta molto grande di dati linguistici derivati da internet. Si tratta del più grande corpus italiano di linguaggio contemporaneo (in quanto usato in internet). È stato stabilito parallelamente alla costruzione di corpora francese, inglese e tedesco, di cui segue gli stessi principi[22]. L'insieme di questi corpora web apre dunque interessanti opportunità di analisi contrastive.

Al fine di presentare un tipo di corpus ancora poco conosciuto dalla comunità scientifica e di discuterne vantaggi e problemi aperti, la prima parte di questo articolo sarà dedicata al problema del concetto di corpus, in modo da caratterizzare il corpus preso in considerazione secondo i suoi singoli aspetti definitori. Questa parte si baserà sulla descrizione del corpus da parte dei suoi autori[23], discussa alla luce di contributi alla definizione del concetto di corpus in lavori recenti e classici della linguistica dei corpora. Per vari aspetti della trattazione dei dati, si confronteranno in questo primo capitolo itWaC e repubblica: se parliamo in questo contesto di prospettive della utilizzazione di dati internet, intendiamo anche indicare quali sono le difficoltà che questo tipo di corpus comporta. itWaC è senza dubbio un corpus di grande

[21] Secondo questi criteri, Bernardini / Baroni / Evert (2006, 10) distinguono: (1) The Web as a corpus surrogate, (2) The Web as a corpus shop, (3), The Web as corpus proper, (4) The mega-corpus/mini-Web. Una discussione critica della concezione di Web as Corpus si trova in Barbera / Onesti / Corino (2007, 44-45).

[22] La presentazione di tutti questi progetti si trova in internet sul sito WaCky.

[23] Seguendo soprattutto gli articoli di Bernardini / Baroni / Evert (2006), Baroni (2008) e di Baroni / Bernardini / Ferraresi / Zanchetta (2009).

importanza, perché offre una preziosa base empirica per future analisi linguistiche, ma nello stesso tempo evidenzia desiderata metodologici e teorici.

Nel secondo capitolo saranno esemplificati, per illustrare il valore empirico di itWaC, analisi condotte su una vasta porzione del corpus, che chiamiamo itWaC'. Partiremo da una discussione più dettagliata della qualità del pos-tagging. In questo modo si riveleranno da un lato problemi tecnici di natura generale, cioè non specifici della lingua scritta nel web; dall'altro lato, analizzando la qualità del pos-tagging, si discuteranno alcune caratteristiche della *Computer Mediated Commucation* (cmc)[24]. Sulla base di esempi frequenti presenteremo alcune di queste caratteristiche. Esamineremo in questo contesto le frequenze dei fenomeni trattati nell'insieme di itWaC'. Si darà un altro esempio di analisi su un tratto tipico di registri anche più formali: il gerundio modale. Questo esempio ci porterà a un'altra specificità dei dati linguistici di itWaC, cioè il trattamento di sintagmi generati automaticamente.

Nella terza sezione dell'articolo approfondiremo il problema dei tratti della cmc per quanto riguarda la loro distribuzione. Mentre nella maggioranza degli studi sul linguaggio di chat, e-mail ed sms si discutono questi fenomeni soprattutto da un punto di vista qualitativo e in uno solo di questi ambiti, il vasto corpus di itWaC' ci offre nuove possibilità di delineare anche i tratti della loro co-occorrenza e distribuzione quantitativa in pagine web diverse. Questa discussione della distribuzione di tratti scelti (abbreviazioni, varianti grafiche, faticismi) e ben noti nella letteratura linguistica sulla cmc, ci pare un passo importante per approssimarci a una differenziazione interna della lingua in rete, specchio di realtà linguistiche diversissime.

Sulla base di queste osservazioni, nella quarta parte saranno formulate, in forma di riassunto e di prospettive, osservazioni e proposte. Si tratta di trovare mezzi adeguati che non solo rendano accessibile la lingua standard in forma di grandi corpora, ma con cui possa anche essere modellato in modo più chiaro l'affascinante continuum, tipico della rete, di comunicazione tra lingua speciale e grafie innovative. Se internet è stato chiamato un «kontextarmes Medium» (Debatin 1998), cioè un mezzo di comunicazione in cui mancano le informazioni sui parametri esterni, pare tanto più importante sapere di più delle sue strutture linguistiche interne[25].

---

[24] Per Lorenzetti / Schirru (2006, 76) sigle e abbreviazioni, brachilogie e anglicismi sono i caratteri della cmc «che sono propri, con tutta evidenza, non della scrittura elettronica in generale, ma piuttosto di quei mezzi di scrittura elettronica che servono a dialogare o, meglio, allorché sono usati soprattutto per dialogare: e-mail, chat, instant messaging, nonché, uscendo dall'ambito della scrittura al computer, gli sms».

[25] Cf. Wenz (1998, 2), riferendosi a Halliday: «Elektronischer Kommunikation läßt sich kein solches semiotisch einheitliches Feld zuordnen. Der Text selber wird zu einem semiotischen Feld».

La linguistica dei corpora e ITWAC

I tratti definitori di un corpus specificati da Sinclair[26] sono la forma del corpus, digitalizzato, e la sua rappresentatività, che si basa sulla scelta dei testi secondo criteri esterni espliciti. Qui si marca una chiara differenza con una banca di dati testuali, la cui composizione segue criteri pratici o impliciti, come nell'esempio del canone letterario di una certa epoca[27]. Discutiamo in seguito questi aspetti fondamentali per capire come un corpus di testi tratti dalla rete possa corrispondere soprattutto alla concezione di testo che prevale nella linguistica dei corpora.

Rappresentatività

Il nome di un corpus esprime in quale senso e per quale tipo di lingua è rappresentativo: un corpus può essere rappresentativo di una lingua nazionale (BNC), di generi testuali (p.es. giornalistici, REPUBBLICA) o del codice parlato (C-ORAL-ROM, Cresti / Moneglia 2005). Per stabilire un campione rappresentativo più o meno ampio, è di importanza fondamentale riferirsi a una classificazione teorica approfondita del complesso dei tipi di testo di cui consiste una lingua nazionale, la lingua dei giornali o la lingua parlata. La scelta finale nell'insieme di questi testi deve essere aleatoria (*random sampling*).

Questo postulato teorico viene seguito nella costruzione di ITWAC soprattutto per quanto riguarda la rappresentatività e il carattere esplicito dei criteri di scelta[28].

Come il nome di ITWAC esprime, questo corpus dovrebbe essere rappresentativo della lingua del web. Ma che cosa si sa del complesso dei testi del web? Come stabilire uno standard per scegliere testi rappresentativi per il complesso della lingua del web? Il vasto insieme dei materiali linguistici in internet presenta una massa di testi che si può descrivere solo approssimativamente. Anche le domande più basilari sono ancora aperte: quali tipi di testi

---

[26] «A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research» (Sinclair 2005, 12).

[27] «Nous employons le mot corpus dans une acception restreinte empruntée à J. Sinclair [...] À cette aune, nombre de ressources textuelles perdent cette dénomination. Il s'agit souvent de collections ou de rassemblements de textes électroniques plutôt de corpus à proprement parler» (Habert / Nazarenko / Salem 1997, 11).

[28] «Moreover, this article provides the most detailed and complete report we are aware of describing random sample of web pages representative of the language of the web» (Baroni / Bernardini / Ferraresi / Zanchetta 2009, 4).

sono più importanti in internet? Quali sono i loro tratti caratteristici? Pur senza discutere approfonditamente tali questioni, ITWAC presenta una risposta implicita, in quanto è orientato verso l'italiano comune presente in internet. Il metodo scelto dagli autori si potrebbe descrivere così: per rappresentare l'insieme dei testi internet usano fonti della lingua standard (giornali) e fonti lessicografiche. Da queste fonti sono state estratte liste di frequenza di singole parole e di coppie di parole (*bi-grams*)[29]. Gli inventari lessicali chiamati *seed* (*words*, *pairs*) derivano da REPUBBLICA, con l'aggiunta di liste di frequenza da dizionari di base. Un motore di ricerca trova in internet siti il cui lessico corrisponda a queste bi-grams e ne salva gli URL[30]. Dall'elenco di URL che ne risulta (*seed URL*) vengono estratti gli URL della top level domain *.it, dopodiché si eliminano i doppioni. Secondo parametri precisi, che definiscono la dimensione minima dei siti da comporre e il tipo di documento (né *.doc né *.pdf), un programma *crawler* usa questi URL per raccogliere in modo randomizzato documenti html da cui, in seguito, viene composto il corpus.

Questa procedura si fonda sull'idea che in un grande corpus di uno dei più grandi giornali italiani si possa trovare una soddisfacente rappresentatività, espressa nelle liste dei *seed*. Se questo approccio può essere messo in dubbio per quanto riguarda la lingua comune (un corpus di giornali può davvero fornire uno standard valido?), il dubbio pesa ancora di più se si prende in considerazione la lingua del web. Attualmente, non si può adottare una tipologia testuale completa e adatta a internet, ma proprio questo aspetto sembra essere di importanza fondamentale nel programma di Web as Corpus.

Identificazione dei testi

I testi che costituiscono un corpus linguistico devono essere identificati tramite l'aggiunta di informazioni metatestuali (di norma integrate fisicamente nei singoli testi nella forma di un markup) che documentano il genere del testo e la sua origine[31].

---

[29] La documentazione completa di queste liste di parole (*seed words*) e di coppie di parole (*seed pairs*) si trova in internet sul sito WaCky (2008-2010).

[30] «Seeds for Italian and German are randomly selected among mid-frequency words in two newspaper text collections *la Repubblica*-, and *Süddeutsche Zeitung*), as well as from basic vocabulary lists, from which function words and particles are removed [...] For Italian, a total of 1.000 pairs are constructed by randomly mixing words from the newspaper and the basic vocabulary list» (Baroni / Bernardini / Ferraresi / Zanchetta 2009, 5).

[31] «Common criteria include: (1) the mode of the text; whether the language originates in speech or writing, or perhaps nowadays in electronic mode; (2) the type of text; for example if written, whether a book, a journal, a notice or a letter; (3) the domain of the text; for example whether academic or popular; (4) the language or languages

Il riferimento a una tipologia testuale contingente all'ambito per cui un corpus deve essere rappresentativo è alla base della sua progettazione. Proprio per il suo ruolo fondamentale, la tipologia testuale di un corpus deve essere molto esplicita, anche per poterlo valutare. Il problema di una tipologia testuale come struttura di base di un corpus linguistico è che deve rispettare nello stesso tempo parametri comunicativi e aspetti tematici (*domains*), e quindi ridurre la complessità del continuo testuale in una griglia adoperabile[32]. In REPUBBLICA i singoli testi sono classificati secondo i tipi news-report vs. comment e ciascun testo è classificato tematicamente (*church*, *culture*, *economics*, *education*, *news*, *politics*, *science*, *society*, *sport*, *weather*). In più, il markup rende visibile anche la struttura testuale interna (*title*, *subtitle*, *summary*, *text*). Per completare i dati metatestuali in REPUBBLICA sono stati aggiunti il nome dell'autore e la datazione di ciascun testo.

In un corpus composto da testi web non possono essere fornite nemmeno le informazioni più basilari: l'origine dei testi non è certa o non è completamente documentata[33]; mancano le informazioni sull'origine geografica del testo, sull'autore, sulla datazione. In più sono da considerare i già citati problemi di tipologia testuale.

Per i testi di ITWAC non possono essere fornite informazioni metatestuali. La sola informazione sulla datazione si riferisce al periodo di composizione del corpus e non dice quindi nulla sulla datazione del testo. Il markup metatestuale in ITWAC è meno che minimale: a ogni testo corrisponde una ID che contiene la URL. In alcuni casi, questa URL contiene parole come *forum*, che quindi dicono qualcosa sulla natura del testo, ma queste indicazioni non sono affidabili, visto quanto può essere complessa la struttura di un sito. Infine, molte delle pagine indicate nella text ID non sono più disponibili.

---

or language varieties of the corpus; (5) the location of the texts; for example (the English of) UK or Australia; (6) the date of the texts» (Sinclair 2005, 2).

[32] La tipologia del BNC, spesso discussa, prevede p.es. in primo luogo una suddivisione dei testi in scritti e parlati e un'ulteriore suddivisione in domains e, per il parlato, in contesti e classi sociali; cf. Aston (2001); Lee (2001) con una discussione critica.

[33] Cf. Lüdeling / Evert / Baroni (2007, 14s.): «For the web as corpus, it is reasonable to assume that all categories of written language are represented to some extent. However, there are no explicit meta-data, at least not of the kind required for linguistic research. The only possibilities for categorizing (or filtring) search results are by (-) language: Google's automatic classifier currently distinguishes between 35 languages; (-) domain name: this has sometimes been used to approximate geographic location (national domains) or even dialect [...] (-) file format [...]: this has presumably little linguistic relevance, except for highly specialized studies; and (-) date: whether a web page has been updated within the last 3, 6 or 12 months» (14s.).

Testualità: contesto comunicativo e coesione

Un altro elemento della definizione di un corpus è la testualità dei suoi elementi, che dovrebbero provenire da situazioni comunicative autentiche e perciò qualificabili come linguaggio naturale (Bergenholtz / Mugdan 1989, 141; EAGLES 7). Rispettando un altro criterio di testualità, quello della coesione[34], liste ed elenchi dovrebbero essere quindi esclusi. Ovviamente REPUBBLICA corrisponde a questi due criteri in quanto è più facile escludere dal corpus elementi non testuali (impressum, enumerazioni redazionali). Al contrario, in un conglomerato di testi web questi criteri di testualità sono più difficili da garantire. Per assicurare che in ITWAC siano inclusi solamente testi nel senso citato, la procedura di preparazione informatica prevede quattro filtri prima che una pagina web sia integrata nel corpus (Baroni / Bernardini / Ferraresi / Zanchetta 2009):

- *boiler-plate stripping*: secondo il criterio del linguaggio naturale, la distinzione tra testi autentici e quelli generati automaticamente, identici in un gran numero di siti web, è molto importante per la costruzione di corpora web. In ITWAC si è scelto un indicatore formale per escludere testi generati automaticamente o brani di questo tipo inclusi in testi autentici, in quanto i primi contengono una grande parte di code HTML. Di conseguenza, sono esclusi testi o paragrafi con una maggiore densità di code HTML, così come i rimanenti tag HTML stessi. Il problema di questo approccio è che nel corso di questa procedura si perdono anche le informazioni strutturali, cioè se una porzione testuale corrisponde a un titolo, un link, all'elemento di una lista o a un paragrafo standard di testo[35].
- *function word filtering*: la presenza o assenza di *function words* è il criterio del secondo filtro. Data una lista di 411 elementi come preposizioni, articoli ecc., si scelgono solamente pagine web che contengono almeno 10 type e 30 token di detta lista per pagina; inoltre, un quarto di tutti i token deve corrispondere a questa categoria. Un effetto collaterale positivo di questo tipo di scelta è che pagine scritte per la maggior parte in lingue straniere vengono eliminate da questo filtro.

---

[34] Cf. de Beaugrande / Dressler (1981, 5). Si potrebbe continuare la discussione con gli altri criteri citati dagli autori: coesione, coerenza, intenzionalità, accettabilità, informatività, situazionalità, intertestualità (ib., 3-14).

[35] Bernardini / Baroni / Evert (2006, 20s.) sottolineano che la conservazione della strutturazione logica dei siti è importante per la loro categorizzazione. Mentre ovviamente gli argomenti del trattamento computazionale prevalgono, l'annotazione linguistica anche degli elementi strutturanti è un desideratum: «Logical structure and hyperlink information might also be useful for purposes of document categorization. However, structural markup and links will constitute noise for the purposes of further linguistic processing (tokenization, POS-TAGGING, etc.) [...] Optimally, a Web-based corpus should satisfy both needs by providing access to the original, unprocessed HTML documents as well as to a linguistically annotated version that had code and boilerplate removed».

- *pornography filtering*: l'esclusione di siti di tipo pornografico si basa su un argomento più linguistico che morale, perché questi siti contengono regolarmente larghi brani di testo generato automaticamente, ottimizzato per i grandi motori di ricerca. Questo filtro elimina pagine web che contengono almeno 3 type o 10 token di una lista con parole chiave del settore.
- *near-duplicate filtering*: il quarto filtro consiste in un confronto automatico, tramite liste di n-grams, per evitare l'inclusione di doppioni; due pagine web non devono avere più di due 5-grams in comune (sulla base di 25 5-grams estratti da ciascun testo).

Misura del corpus

La grandezza di un corpus non garantisce la sua rappresentatività, ma ne è una condizione importante. A partire da una grandezza di 100 milioni di token si parla di un corpus «molto grande» (Chiari 2007, 45). Corpora di grandi dimensioni sono molto spesso formati da articoli di giornali digitalizzati dalla casa editrice, come il già citato corpus REPUBBLICA. Per renderli meglio comparabili, i singoli testi del corpus possono essere normalizzati (p.es. 2000 token / testo). La scelta dei testi dovrebbe essere randomizzata[36].

Senza dubbio ITWAC è, con i suoi 1.585.620.279 token, un corpus «molto grande», anzi enorme, dell'italiano. Il metodo di *crawling* corrisponde al criterio del random sampling. Come si vedrà, la mancata normalizzazione della lunghezza dei testi non facilita la loro valutazione. Si tratta di testi di lunghezze molto diverse.

Markup linguistico

Se si parla di un corpus linguistico, si intende che è stato preparato per ricerche che richiedono l'aggiunta di informazioni linguistiche a vari livelli, il più elementare dei quali è quello della morfosintassi[37]. Con gli strumenti di part-of-speech-tagging (POS con il TreeTagger e lemmatizzato con MORPH-IT!; Zanchetta / Baroni 2005) si aggiungono le indicazioni sulla parte del discorso a cui appartengono le unità testuali e sul genere, numero, tempo e modo di nomi, aggettivi e verbi, completate dall'indicazione del lemma (fig. 1).

---

[36] Cf. per la discussione di questi aspetti Bergenholtz / Mugdan (1989, 148); Biber (1994b, 179).

[37] Non approfondiamo il problema del tokenizing, discusso p.es. in Barbero / Onesti / Corino (2007).

```
<text id = "http://.comune.genova.it/portal/page/categoryItem?contentId=79482">
<s>
Cliccando          [VER:geru]      cliccare
sul                [ARTPRE]        sul
nome               [NOUN]          nome
di                 [PRE]           di
ciascun            [DET:indef]     ciascun
ospedale           [NOUN]          ospedale
,                  [PUN]           ,
[...]
</s>
</text>
```

Fig. 1 - Markup in ITWAC: text-id, inizio e fine del testo e di frasi, part-of-speech-tagging e lemmatizzazione

La segmentazione della scrittura pone a sua volta problemi di segmentazione delle unità grafiche in unità lessicali. La lista di frequenza dei lemmata identificati dal POS (sul già citato sito WaCky 2008-2010) evidenzia anche i problemi della segmentazione su cui si basa: inclusi nella lista dei lemmata si trovano interpunzioni ripetute, frammenti di URL di siti web, frammenti di tecnicismi e falsi composti da nomi propri.

Accessibilità

I corpora linguistici sono destinati a un uso il più aperto possibile, e non solo da parte di singoli ricercatori della disciplina. Accanto a corpora venduti con pubblicazioni cartacee dalle case editrici (cf. C-ORAL-ROM) o accessibili solamente ad abbonati (cf. frantext.fr) esistono molti corpora liberamente consultabili in rete, perlopiù forniti di una maschera di ricerca. ITWAC può essere scaricato gratuitamente, dopo una semplice registrazione, ma non è possibile condurre su di esso interrogazioni online.

Ricchezza lessicale: ITWAC vs. REPUBBLICA

Per stimare l'apporto scientifico di ITWAC, i suoi autori l'hanno confrontato con REPUBBLICA secondo due criteri di frequenza lessicale: da una parte il *coverage*, cioè l'analisi comparativa del numero di token che ricorrono più di 20 volte; dall'altra l'*enrichment*, cioè il numero di token che ricorrono più

di 20 volte in uno solo dei due corpora. Ne risultava che ovviamente itWaC contiene una parte molto grande del linguaggio contemporaneo. In più, confrontando le liste di frequenza dei due corpora, si constatava che il 94-96% di repubblica è presente in itWaC, mentre secondo le parti del discorso in repubblica è inclusa una parte di itWaC ridotta, che varia dal 21.2% (verbi) al 25.6% (aggettivi) e al 42% (sostantivi) (Baroni / Bernardini / Ferraresi / Zanchetta 2009, 218s.). L'*enrichment* di repubblica è molto basso, con 0.1% (sostantivi), 0.8% (aggettivi e verbi), mentre quello di itWaC è di 70.7% per i sostantivi, 72.6% per gli aggettivi e 75.2% per i verbi. Ovviamente, per quanto riguarda la ricchezza lessicale, itWaC comporta un avanzamento per gli inventari linguistici.

Qualità del part-of-speech-tagging

I tratti linguistici della lingua presente in itWaC pongono gravi problemi ai programmi di pos-tagging. Secondo le analisi degli autori, per più del 20% dei nomi e per un terzo degli aggettivi e dei verbi non sono stati trovati i lemmata corretti. Gli autori chiamano *noise* l'insieme delle forme non correttamente annotate. Secondo la loro definizione[38], la quantità elevata del noise, o rumore statistico si spiega con il grande numero di errori ortografici, di elementi che non fanno parte del lessico e di parole non italiane. Questo alto numero d'elementi non riconosciuti dal programma è unico per il corpus itWaC; i valori corrispondenti del BNC o di repubblica sono tra l'8 e il 4% (Baroni / Bernardini / Ferraresi / Zanchetta 2009, 219).

Nell'ambito della statistica, si definisce rumore la quantità di dati non rilevanti che limitano il valore scientifico dei risultati. Originariamente, si tratta di un termine delle radiocomunicazioni: se non si trova la frequenza esatta, si sentono rumori. Se si parla di *noise* o di rumore a proposito di dati lessicali, si potrebbe dire, pur semplificando, che elementi lessicali e morfologici taggati correttamente dànno un'immagine acustica definita, mentre il resto disturba la purezza del suono.

A questo punto, ci sembra importante una riflessione linguistica sulla natura di suono (elemento interessante) e di rumore (elemento che disturba). Per esemplificare il concetto di noise, gli autori di itWaC citano errori e *garble* che potrebbero però indicare anche forme innovative ed elementi di altre lingue. Questi fenomeni sono davvero un rumore che disturba? Considerando

[38] «For the purposes of this task we define noise as typos, garble or words from foreign language texts – e.g. the verb *woudl*» (Baroni / Bernardini / Ferraresi / Zanchetta 2009, 218).

che si tratta di un corpus di dati linguistici del web, essi sono davvero da eliminare?

La distinzione tra parole «buone» – riconoscibili con i mezzi della linguistica computazionale – e noise è di natura pratica, in quanto si riferisce al trattamento completamente automatico della massa dei dati (assolutamente necessario, data la misura del corpus). Ma nonostante questo, e nonostante gli argomenti tecnico-metodologici, è necessario riflettere sul concetto di norma linguistica che si esprime con questa terminologia. Un approccio alla lingua standard tramite la linguistica computazionale si armonizza difficilmente con realtà linguistiche non conformi allo standard. Nel corso del sec. XX, per esempio nella *Grammaire des fautes* de Henri Frei (2007/1929), si è messa in discussione la dicotomia tra linguaggio corretto e linguaggio erroneo. Frei passava da una concezione normativa a una prospettiva funzionale, che riconosceva l'espressività o più in generale la forza comunicativa anche di forme che non facevano parte della lingua standard[39]. Non occorre a questo punto delineare il processo per cui la ricerca sulla lingua parlata ha contribuito a riconoscere che le forme linguistiche non conformi allo standard non seguono un puro caso. Anzi, nella loro distribuzione si riconoscono regolarità d'uso, che, nell'ambito della ricerca sulla lingua parlata, sono state descritte sia a livello universale sia nelle loro particolarità storico-idiomatiche (Koch / Oesterreicher 2007). Continuando in questa direzione, anche gli studi sulla lingua del web riprendono l'idea che le apparenti irregolarità della lingua possano formare un insieme coerente e un oggetto appropriato della ricerca linguistica.

Nel capitolo che segue cercheremo di dare un quadro preciso della natura dei dati linguistici che compongono itWaC, usando la sua preparazione con metodi della linguistica computazionale, ma nello stesso tempo indagando anche i tratti non standard, anche a livello grafico.

Come primo passo, per dare un'impressione generale del corpus discuteremo la frequenza di forme concorrenti in questo corpus e nelle indicazioni date da Google. Riprenderemo poi il problema citato del noise: passeremo in esame i problemi del pos-tagging e altre caratteristiche di itWaC, cercando di spiegare tali fenomeni problematici in un modo che vada oltre la qualificazione semplice e automatica di tutte le forme non riconoscibili come elementi

[39] «La distinction du correct et de l'incorrect est une des premières difficultés auxquelles s'achoppe le grammairien qui étudie un état de langue. Qu'appelle-t-on un fait de langage ‹correct› et, lorsqu'on parle d'une ‹faute›, que veut-on dire par là ?»; «Une autre conception, que nous appellerons la conception fonctionnelle, fait dépendre la correction ou l'incorrection des faits de langage de leur degré de conformité à une fonction donnée qu'ils ont à remplir» (Frei 1929/2007, 17; 18).

disturbanti. Più precisamente, saranno esaminate le categorie particolarmente problematiche del POS, la presenza di fenomeni tipici del plurilinguismo della lingua del web e alcuni fenomeni che sono frequenti nella CMC: anche se, come si è detto nella presentazione di ITWAC, la procedura di crawling preferisce siti internet in cui la frequenza lessicale sia simile a testi della lingua standard (p.es. dei giornali), nella lingua di ITWAC si trovano alcuni fenomeni, non solo grafici, che caratterizzano la comunicazione dialogica nel web. Questi fenomeni saranno esaminati con riguardo alla loro frequenza e alla loro distribuzione, nonostante il loro numero sia, in termini percentuali del corpus, di non grande peso; se in un corpus vasto come quello presentato le cifre dei fenomeni scelti sono relativamente modeste, in termini assoluti si ha comunque ragione di discutere la loro distribuzione e i rapporti di co-occorrenza tra singoli fenomeni.

Analisi di tratti linguistici di ITWAC'

Preparazione dei dati

Per valutare più in dettaglio la solidità dei dati in ITWAC, e quindi il valore del corpus, abbiamo effettuato analisi linguistiche esemplari che si basano su una parte del vasto corpus, rispettando i limiti della memoria di lavoro del computer utilizzato. Di questo corpus parziale si parla in seguito con la sigla ITWAC'.

La misura del corpus parziale ITWAC' si approssima alla misura di REPUBBLICA, in quanto comprende 388.392.159 token. Il corpus parziale di ITWAC' include quindi circa un quarto dell'intero corpus ITWAC. Consiste in 455.222 testi, cioè pagine web computate secondo la text ID inclusa nei metadati. Allo stato attuale, non si può utilizzare un motore concepito per la ricerca in ITWAC. Conseguentemente, gli archivi sono stati trasformati e riorganizzati per la programmazione delle nostre analisi, che saranno descritte di seguito[40].

Frequenze di Google e ITWAC' a confronto

Per paragonare i risultati di ricerche su Google e nel corpus preso in considerazione sono stati contate in ITWAC le occorrenze nello stesso modo del motore di ricerca Google, computando cioè non i token ma le pagine web in cui ricorrono i type.

---

[40] Tutti gli algoritmi di ricerca descritti in seguito sono stati programmati in TUSTEP. Gli archivi XML sono stati – per ragioni pratiche (memoria di lavoro) – divisi e trasformati in archivi del sistema TUSTEP. I grafici e le computazioni sono stati programmati ed effettuati in R.

Abbiamo scelto tre fenomeni a tre livelli diversi. Nel primo caso si tratta dell'anglicismo *cliccare*, frequente all'imperativo, non solo nel sintagma *clicca qui*. Il secondo caso, la ‹d› eufonica, è una marca di stile elevato ancora molto usata, come mostrano le frequenze (tab. 1). Il terzo esempio, l'uso obbligatorio del congiuntivo dopo *credere che*, esprime il rispetto della norma linguistica da parte del parlante.

	Google (sito:.it, 30.03.2010, pagine in italiano)		itWaC	
click / clicca	12600000/16900000	0.75	2593/5308	0.49
a un altro / ad un altro	16100000/22000000	0.73	1388/2840	0.49
credo che sia / credo che è	8470000/10300000	0.82	2741/68	40.31

Tab. 1 - Frequenze Google e itWaC a confronto

Le possibilità di trarre conclusioni da queste cifre sono molto limitate a causa dei problemi del motore di ricerca di Google (Kilgarriff 2007). In più è difficile interpretare questi risultati perché non sappiamo quasi niente sulla composizione dell'insieme dei dati linguistici presenti in internet. Quale dei due metodi porta a risultati più rappresentativi? È impossibile trovare una risposta a questa domanda, perché il complesso dei dati da cui si forma un campione 'rappresentativo' è sconosciuto.

Comunque, nel caso dei primi due esempi si osserva una tendenza di itWaC a un linguaggio leggermente più conservativo. Nel terzo caso invece si osservano non solo un rapporto inverso, ma anche la scarsissima frequenza di *credo che è* in itWaC. In questo caso pare lecito constatare che in itWaC sia rappresentato, nell'insieme, un linguaggio più orientato verso la lingua standard. Questa osservazione corrisponde al metodo, sopra descritto, di cercare le pagine web tramite liste di frequenza che non includono registri informali.

Tagset

L'analisi del tagset di itWaC e della sua applicazione aiuta a precisare la prima impressione di lingua standard in itWaC. Come si è spiegato sopra, per l'insieme del corpus sono da tollerare molti errori dei tag pos. Per primo studiamo il caso paradossale in cui un tipo di identificazione consiste nell'evitare la categorizzazione. Come si osserva nell'analisi dei dati di itWaC', esiste una

categoria di avanzi, cioè di elementi non categorizzabili automaticamente. Questa categoria è integrata nel tagset sotto il nome di NOCAT.

Proprio questo paradosso sembra essere di particolare importanza, data la considerevole quantità di token inclusi (tab. 2): l'1,7% di tutti i token in ITWAC' fa parte di questa classe. La lista dei token più frequenti nella categoria NOCAT dimostra che più della metà di tutte le occorrenze di questa categoria in ITWAC' è nei ranghi 1 e 2. Sulla base della lista si possono fare due osservazioni. La prima potrebbe essere valida per altri tipi di testi, l'altra è più specificamente orientata alla lingua del web.

Token	occorrenze	rango	% in NOCAT
%	260159	1	41.14%
Ciao	68491	2	10.83%
Ah	17584	3	2.78%
???	17532	4	2.77%
??	16874	5	2.67%
beh	16843	6	2.66%
eh	12976	7	2.05%
viva	12806	8	2.03%
oh	10734	9	1.70%
etc	10072	10	1.59%
via	7099	11	1.12%
D	5749	12	0.91%
????	5590	13	0.88%
prego	5279	14	0.83%
mah	4978	15	0.79%
XIII	4953	16	0.78%
XVI	4356	17	0.69%
salve	3773	18	0.60%
Go	3714	19	0.59%

Tab. 2 - ITWAC': i più frequenti token categorizzati in NOCAT

La prima osservazione concerne il segno di percentuale ‹%› e le cifre romane, che ricorrono in percentuali considerevoli. Questo tipo di segni grafici non dovrebbe risultare difficile da trattare, dato che è conosciuto anche in testi non derivati dal web, come i giornali[41]. Siccome le cifre romane sono abbastanza frequenti, non solo in contesti storiografici, ma anche in testi turistici, dovrebbe essere conveniente integrare la categoria nel tagset o trattare le cifre romane allo stesso modo delle cifre arabe. Un altro problema del POS-tagging, che si osserva in un'analisi dettagliata dei risultati, sono i nomi propri. Con l'introduzione di dizionari elettronici di nomi propri dovrebbe essere possibile diminuire il numero di POS-tags erronei. A volte gli errori sono molto creativi dal punto di vista morfologico, in quanto gli algoritmi dànno come risultati un lemma *marlire* e un lemma *brare* dal nome del famoso attore Marlon Brando. Altri aspetti difficili sono i segni di interpunzione all'interno di una parola o di abbreviazioni (*sig.ra*, *P.S.*, *etc.*, *D.M.*). Saranno solamente menzionati, senza discuterli approfonditamente, altri noti problemi del POS-tagging, come la distinzione tra congiunzione e avverbio interrogativo.

La seconda osservazione concerne elementi del parlato così come ricorrono nei contesti dialogici della CMC, p.es. la ripetizione dei segni d'interpunzione[42]. Un altro elemento importante sono le interiezioni, che compaiono nei primi ranghi ma che non sono state riconosciute dal programma POS. Nonostante il tagset del TREETAGGER (citato accanto a MORPH-IT! dagli autori) includa un tag POS per le interiezioni, in ITWAC non sembra essere adoperato. Forse lo sviluppo del tagset riprende le categorie di REPUBBLICA (neanche nel tagset di questo corpus si trova la categoria dell'interiezione). Ovviamente in un corpus di testi presi da un grande giornale, dato che si tratta di una lingua molto vicina allo standard, pare meno naturale prendere in esame una categoria pragmaticamente marcata come le interiezioni. Nonostante questa origine del corpus, pare importante osservare come negli esempi la diversità grafica e i suoi effetti sul POS-tagging presentino un problema fondamentale di natura teorica, anche se le frequenze, in relazione alla vastità del corpus, sono piuttosto basse.

Interiezioni

Quanto al tagset di ITWAC, il caso delle interiezioni dimostra come con un'aggiunta molto semplice si potrebbe diminuire sensibilmente il numero di

---

41 Si tratta della preparazione di base del testo, descritta p.es. in Rehm (2001, 187).

42 Fanno parte della «sekundären Schriftlichkeit» (Schmitz 2009) e sono stati descritti regolarmente nel contesto delle ricerche sulla comunicazione delle chat (Pistolesi 1997; Pistolesi 2004).

errori indicati come NOCAT. Per approfondire la discussione sull'analisi della categoria NOCAT, prendiamo in esame le interiezioni più frequenti dell'italiano parlato, con l'aggiunta di due interiezioni frequenti nelle chat, *sigh* e *smack*[43] (fig. 2): in prima posizione si trova l'interiezione *ciao*; anche *salve* è usato frequentemente nella sua funzione di saluto, ma ricorre con eguale regolarità come aggettivo.

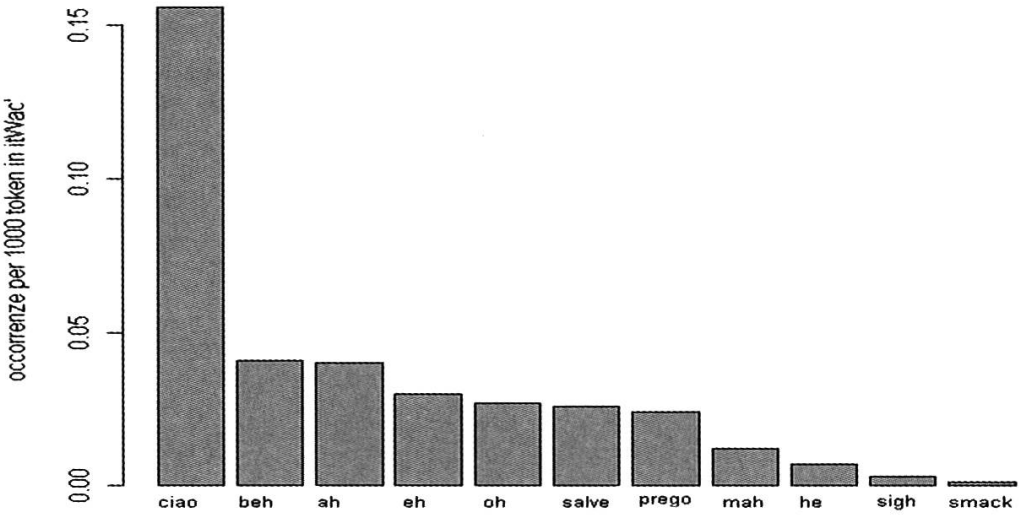

Fig. 2 - interiezioni in ITWAC'

La grande distanza tra questo saluto e le altre interiezioni potrebbe segnalare che esso non comporta automaticamente la compresenza di altri elementi di comunicazione dialogica o elementi del parlato. In più, i tratti di raddoppiamento grafico e gli ideofoni *sigh* e *smack* sono molto più rari di *beh*, *ah* ecc. Queste osservazioni evidenziano quella che è stata chiamata l'ibridizzazione della CMC: alcuni dei suoi tratti tipici si diffondono, altre forme risultano più circoscritte.

È da aggiungere che – come nel caso dei segni d'interpunzione – anche le interiezioni presentano una forte variazione grafica, che contribuisce alle irregolarità del POS-tagging. La categorizzazione del POS-tagging non tratta

[43] Sono incluse nell'analisi le interiezioni più frequenti del corpus italiano di C-ORAL-ROM (Cresti / Moneglia 2005), quelle menzionate di ITWAC' (tab. 2) e descritte per la chat, cf. Pistolesi (2004, 103).

le interiezioni in modo omogeneo. La presenza di una maiuscola conduce a una classificazione come sostantivo [NOUN] o come nome proprio [NPR]: *mAh* [NOUN], *HHH* [NPR], *cIAO* [NOUN], *aH* [NOUN], *eH* [NPR].

Grafie espressive

Aggiungiamo a quest'analisi l'osservazione che in ITWAC' si trovano alcuni tipi di emoticons (fig. 3). In questo caso, come per la ripetizione di segni di interpunzione o vocali[44], si tratta di un altro fenomeno che fa parte dei mezzi della scrittura secondaria, in quanto sembra rivolgersi «all'orecchio più che all'occhio» (Pistolesi 2004, 102). Le ripetizioni, spesso scritte in maiuscolo (soprattutto vocali, ma anche consonanti), imitano l'allungamento di un suono o l'alzarsi della voce. Queste osservazioni concernenti aspetti fatici ed espressivi, con mezzi lessicali e segni grafici, mostrano che la frequenza dei singoli elementi è molto diversa.

Le forme abbreviate come *cmq*, *ke*, *nn* e la sostituzione di *per* con ‹x› sono molto tipiche della CMC (fig. 4). Non si tratta di forme grafiche specifiche della CMC, tanto che ricorrono in molti testi scritti di poca formalità e ormai hanno una loro tradizione[45]. Non sono quindi nuove le forme in sé, ma la loro frequenza e la loro espressività in quanto «allografi connotativi» (Pistolesi 2004, 99) ne fanno un tratto caratteristico di alcuni ambiti della CMC.

La forma grafica *cmq* può anche stare per 'centimetro quadro'. Ma se si intende questo significato, l'abbreviazione è generalmente preceduta da un numero. Questa caratteristica è stata utilizzata per escludere le occorrenze dell'unità di misura. Nel caso di *nn* è possibile distinguere automaticamente tra 'non' e 'numeri', significato quest'ultimo frequentissimo in testi giuridici o amministrativi, perché *nn* 'numeri' è seguito da un punto o da una cifra.

---

[44] Per evitare l'inclusione di cifre romane o di sigle, per le vocali sono state contate serie di quattro o più segni ripetuti.

[45] Per questi aspetti, nella CMC si proseguono le tradizioni grafiche della *mala lingua* e del linguaggio giovanile; cf. Gerstenberg (2004, 317); per il francese cf. Anis (2002): «Aucun des procédés utilisés n'est nouveau, on les trouve dans les abréviations scolaires, les jeux de lettres (entre autres les rébus), la littérature (Queneau et les autres), la chanson (L.N.A.H.O. de Polnareff, par exemple). On peut d'ailleurs considérer qu'il s'agit de compléments à l'écriture latine alphabétique qui évoquent les logographies, les syllabaires, les alphabets consonantiques. Néanmoins, la combinaison, dans le cadre d'un nouveau mode de communication, de ces procédés pour produire des messages brefs et expressifs, est originale».

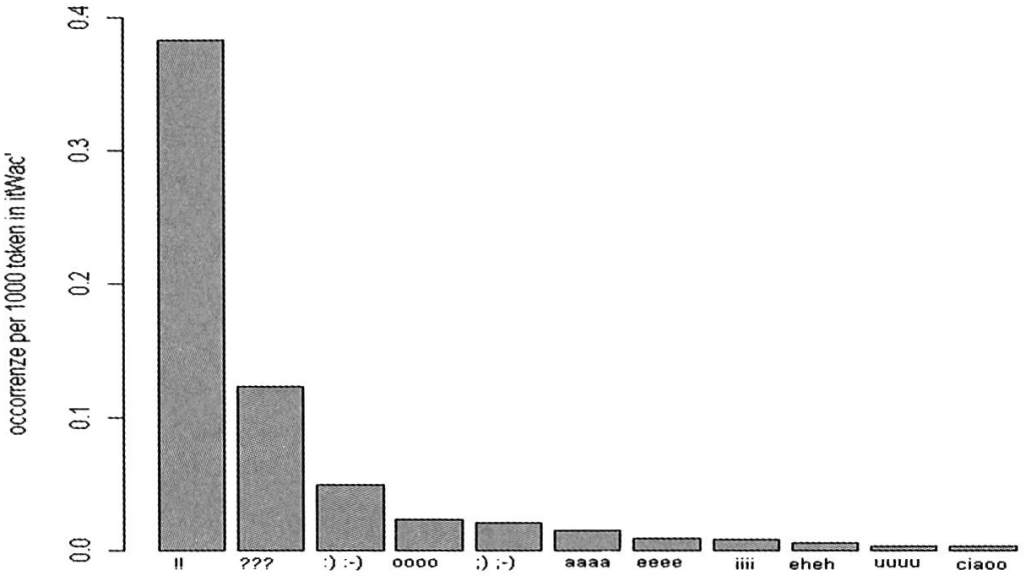

Fig. 3 - Grafie di funzione espressiva (le forme raddoppiate possono essere continuate: *????????*, *ciaooooooo*)

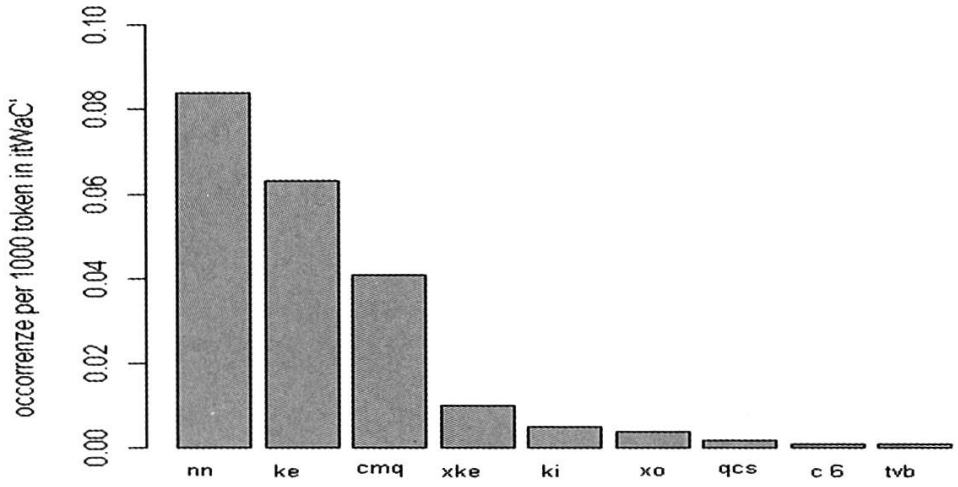

Fig. 4 - Brachilogie (in *(x)ke* sono raggruppati anche *(x)ké*, *(x)kè*, in *xo* anche *xò*)

L'esame del POS-tagging in ITWAC mostra che queste varianti sono parzialmente riconosciute dal POS. È questo il caso di (*nn* 'non', *ke* 'che'). Altre forme non sono riconosciute e di conseguenza non sono state taggate correttamente.

È il caso di *cmq* 'comunque', taggato come [ADJ] da un «lemma» ugualmente indicato con *cmq*, ma anche come 'centimetro quadro' [NOUN]. Anche le grafie ormai quasi classiche del linguaggio giovanile[46] sono difficili per l'algoritmo di tagging (*k* < *ch*, *x* < *per*).

Le occorrenze su 1000 token fanno vedere che questo tipo di grafia non è molto diffuso nei testi inclusi in ITWAC'. Anche la forma più frequente, *nn*, non ricorre con regolarità. Relativamente rare risultano le forme *c 6* 'ci sei' e *tvb* 'ti voglio bene'; mancano o sono presenti con singole occorrenze le abbreviazioni tipiche delle chat come *dgt* (*dgti*, *dgto* < *digitare*).

Testi plurilingui

Nella definizione citata di noise sono considerati disturbanti anche elementi lessicali da altre lingue. Quest'approccio pare in contraddizione con l'oggetto stesso della ricerca: *web* sta per il *world wide web*, il cui plurilinguismo è costitutivo. In ITWAC questo plurilinguismo è presente in forma di prestiti, anglicismi e tecnicismi. Cercando i contesti di parole inglesi, non si trovano soltanto singoli lessemi e citazioni ma anche brevi testi paralleli – tradotti come nell'esempio (fig. 5).

<text id = "http://www.fuoriluogo.it/forum/posting.php?mode=quote&p=1075">

<s> = « Moloch «] NARCOTERRORISMO : DEFINIZIONE FUTURA DI UN CRIMINE Tirare su una canna fa di te un terrorista ? </s>

<s> Se Bush ottiene ciò che vuole gli utilizzatori di droghe diventeranno bersagli della guerra al terrorismo. </s>

<s> NARCOTERRORISM : FUTURE CRIME DEFINED Does packing a bowl make you a terrorist ? </s>

<s> If Bush gets his way , drug users will find themselves targets in the War on Terror. </s>

[...]</text>

Fig. 5 - Testi tradotti in un forum di discussione giornalistica (ITWAC')

Accanto a testi tradotti, per esempio in forum di discussione, si trovano testi mistilingui, redatti da autori di contributi che scrivono in più lingue.

Sottolineiamo ancora di più, non per l'aspetto del plurilinguismo ma della variazione intralinguistica, esempi di uso marcato in senso diatopico. Così

[46] E non solo: cf. Arnold (2008) per la storia italiana del grafema ‹k›.

ricorrono l'articolo romanesco *er* (*er tram*, *er Monnezza*, *Er/er mejo*, *er core*, *Er sogno*, *Er faciolo*, *Er papa*, *Er Pecora*, *er vino*, *er monnezza*, *er cervello*, *er Cipolla*, *er piacere*, *er pupone*, *Er Centurione*, *er septembre*) e molte forme aferetiche. Nella lista dei lemmata si trovano p.es. le varianti *'ncoppa*, *'nnammurato*, *'nnaggia*, *'nnucente* ecc.

Esempio di analisi e caveat

Nell'esplorazione del corpus si dimostra di grande valore soprattutto la quantità dei dati linguistici inclusi, perché permette di ritrovare anche attestazioni delle strutture meno frequenti nella lingua contemporanea. Questo aspetto è tanto più importante in quanto, in quello che è stato detto finora, si sono trovati molti indicatori che in itWaC siano largamente presenti varietà diafasiche anche elevate della lingua.

Il valore del corpus per ricerche linguistiche che vanno oltre le caratteristiche della CMC è esemplificato nella tabella seguente sulla base delle collocazioni con gerundio (tab. 3)[47], più precisamente dei *gerundi modali*. Il gerundio sembra essere oggetto del cambio linguistico recente[48] ed è stato descritto come marcato stilisticamente e limitato all'uso burocratico e scientifico[49].

Sembra perciò interessante sapere quali sono le regolarità del suo uso, se p.es. si osservano tendenze di fissazione limitate a collocazioni con alcuni verbi. Includiamo qui quest'analisi perché nel caso del gerundio si svela un problema fondamentale del corpus, che presenta un grave pericolo di falsificazione dei risultati di una ricerca e che tuttavia è radicato nella natura stessa dei dati.

Il tipo di ricerca scelto per il caso del gerundio è possibile solamente in un corpus con markup linguistico (POS)[50]. Si basa sull'osservazione che la posizione canonica del gerundio modale è dopo il verbo principale (Ferrari / Zampese 2006, 50). È stata formulata una ricerca che richiedeva la forma flessa di un verbo, eventualmente seguito da un avverbio e da un verbo al gerundio. Sono stati esclusi gerundi composti e perifrasi verbali con *andare*, *stare*,

---

47 L'ambito di ricerca sulle collocazioni in corpora linguistici è molto importante, ma non può essere discusso in dettaglio qui; per le nozioni generali cf. Evert (2009, 1213).

48 Mentre nel decennio 1990-1999 la percentuale dei verbi all'indicativo è cresciuta, gli altri modi, tranne il gerundio, sono divenuti meno frequenti (Bolasco 2005, 351).

49 Cf. l'analisi di Policarpi / Rombi (1983) e per la sintassi del gerundio italiano Antonini (1974-1975); Lonzi (1991); Zampesi (2004).

50 Questo tipo di ricerca per *patterns grammaticali* viene effettuato anche da motori di ricerca come SketchEngine, descritto in Kilgarriff / Rychly / Smrz / Tugwell (2004).

venire. Il problema morfologico dell'avverbio è stato risolto parzialmente, in quanto tra le due forme verbali sono stati accettati avverbi semplici e avverbi in *-mente* (che sono stati taggati correttamente), combinazioni di avverbi e alcuni sintagmi avverbiali.

Il risultato dell'analisi (tab. 3) dimostra che nell'uso del gerundio modale nella posizione descritta prevalgono ovviamente costrutti che modificano dei verba dicendi. L'osservazione della combinazione sistematica di verbi come *concludere dicendo*, *rispondere dicendo* ecc. presenta un tipo di atto illocutivo complesso.

Lemmata delle forme flesse del verbo	verbo al gerundio	occorrenze (in assoluto)	rango	occorrenze in %
rispondere	citando	3218	1	2.58%
concludere	dicendo	428	2	0.34%
effettuare	utilizzando	246	3	0.20%
permettere	citando	227	4	0.18%
concludere	affermando	210	5	0.17%
rispondere	dicendo	196	6	0.16%

Tab. 3 - Gerundi modali semplici: collocazioni in ITWAC'

Nella ricerca sui verba dicendi, l'uso di *speech act verbs* vale sia in senso pragmatico sia per indicare i verba dicendi (Gansel 2002, 1563). Gli esempi citati possono essere descritti in termini pragmatici come composti di un verbo di locuzione (*dire*) con un verbo illocutivo che indica p.es. la conclusione di un turn (*concludere*, *rispondere*) – quasi una illustrazione della distinzione di Searle (1991) tra l'atto di locuzione e il valore pragmatico di un atto illocutivo.

Queste considerazioni illustrano quali tipi di ricerca si possono effettuare in ITWAC. Ma invece di approfondire questa discussione, in questo contesto pare più importante fare attenzione ai risultati quantitativi della ricerca. Le occorrenze dei primi ranghi dimostrano una forte differenza tra il primo valore e gli altri: le occorrenze del tipo in prima posizione sono quasi otto volte più numerose di quelle in seconda posizione.

Con il sintagma *rispondi citando* – come conferma il controllo delle attestazioni nel loro contesto – si ha a che fare con un elemento fisso che fa parte della cornice prefigurata di una pagina web in un forum di discussione. Clic-

cando su *rispondere citando* (o sul bottone accanto) si include il messaggio precedente nel proprio messaggio. Come descritto sopra, con il filtro di boilerplate, nella costruzione di itWaC si è cercato di escludere tali tipi di linguaggio non naturale, generato automaticamente. L'esempio di *rispondi citando* mostra quanto è difficile realizzare questo proposito. Ricerche come quella presentata sui gerundi, o il controllo ancora più completo di liste di frequenze, possono fornire metodi adatti a individuare tali elementi, che non dovrebbero far parte di un corpus linguistico, o per meglio dire richiedono un trattamento particolare.

Argomenti di distribuzione

Nei capitoli precedenti abbiamo analizzato soprattutto frequenze assolute o relative di forme non standard in tutto il corpus di itWaC'. Abbiamo visto che i tratti molto discussi dell'*italiano spedito* (Pistolesi 2004) non sono frequenti nel corpus, il che è da considerare una diretta conseguenza della procedura di preparazione del corpus. Partendo dai risultati ottenuti fin qui – ossia che anche in un corpus web orientato maggiormente verso la lingua standard si trovano tratti non standard caratteristici della CMC dialogica – cerchiamo di fare alcune osservazioni su una possibile ibridizzazione sottostante.

La preparazione dei dati, così come sono inclusi in itWaC, limita le possibilità di analisi statistiche. Il problema consiste nel fatto che i singoli testi sono di lunghezze molto diverse e un *random sampling* non pare lecito per gli altri problemi del tokenizing: allo stato attuale del POS-tagging non si possono conteggiare i token di una pagina in modo ragionato. In più pare molto difficile prendere in considerazione parallelamente anche il POS-tagging, vista la quantità di problemi che comporta.

In questa situazione, per non rinunciare a un abbozzo di distribuzione dei vari fenomeni della CMC di cui si è parlato finora, prendiamo in considerazione solamente le pagine in cui detti fenomeni ricorrono. In armonia con i risultati già citati, in itWaC si trova una percentuale abbastanza bassa di pagine web con tali occorrenze. In seguito abbiamo raggruppato i fenomeni finora discussi e conteggiato le pagine web in cui ricorrono, senza però tener conto della loro frequenza nelle pagine web o della lunghezza dei contesti, per le ragioni citate.

In una prima fase, abbiamo annotato le pagine web in itWaC' in cui ricorrono i fenomeni che nelle analisi precedenti sono risultati i più frequenti. I fenomeni sono stati raggrupati in sei categorie: saluto (*ciao*, escludendo *salve* per le ragioni di ambiguità dette), ripetizione di segni d'interrogazione o

esclamazione, ripetizione delle vocali *a*, *e* o *i* (più di quattro volte, per evitare l'inclusione di cifre romane per la *i*), interiezione (*ah*, *beh*, *eh*, *oh*), smileys (*:)*, *:-)*, *;)*, *;-)*), brachilogie (*cmq*, *ke*, ma escludendo *nn* che può significare 'numeri' o 'non'). Questo tipo di annotazione ha permesso di estrarre per ogni categoria i vettori, consistenti in una serie che con 1 e 0 indicava se in una pagina web inclusa in ITWAC' ricorreva o non ricorreva un fenomeno della categoria in questione.

Per stimare la possibilità di trovare patterns nella distribuzione abbiamo calcolato la correlazione tra i sei vettori. In nessun caso è risultata una correlazione significativa (con valori di r(SPEARMAN) tra 0.1 e 0.14, p>0.01). Più interessante sembra analizzare le pagine web dal punto di vista della densità dei fenomeni osservati. Abbiamo conteggiato le pagine web di ITWAC' in cui ricorrono i fenomeni dei sei gruppi. Risultava che solo nel 14% delle pagine di ITWAC' si trova uno dei fenomeni indicati (tab. 4). Le percentuali molto basse dei tipi più frequenti di combinazione corrispondono al basso valore della loro correlazione.

fenomeni da ... gruppi	% di tutte le pagine web in ITWAC'	combinazione più frequente	% di pagine web con questa combinazione in ITWAC'
0	86.37		
1	7.45	ripetizione di *??*, *!!*	3.20
2	3.20	ripetizione di *??*, *!!*, *ciao*	0.95
3	1.66	ripetizione di *??*, *!!*, *ciao*, interiezioni *ah*, *beh*, *eh*, *oh*	0.31
4	0.84	ripetizione di *??*, *!!*, *ciao*, interiezioni, brachilogie *cmq ke*	0.20
5	0.36	ripetizione di *??*, *!!*, *ciao*, interiezioni, brachilogie *cmq ke*, vocali rip. *a*, *i*, *o*	0.12
6	0.12	ripetizione di *??*, *!!*, *ciao*, interiezioni, brachilogie, vocali rip., smileys :) ;) :-) ;-)	0.12

Tab. 4 - Pagine web e fenomeni tipici della CMC in ITWAC'

Però, sempre a questo livello di frequenza generalmente bassa, si osserva una gerarchia tra i fenomeni indagati: al primo posto la ripetizione dei segni di esclamazione e interrogazione, che ricorre in contesti pubblicitari e in contributi a forum di discussione dove generalmente prevale un uso linguistico orientato verso la lingua standard. A questo punto le cifre possono aiutare a formulare ulteriori ipotesi per descrivere una possibile ibridizzazione del linguaggio in internet secondo la qualità e la densità dei tratti tipici della CMC che ricorrono in certi contesti di blog, forum di discussione e webchat.

```
<text id = "http://blog.mrwebmaster.it/alidifarfalla/4136/">
[...]
<s> 20.5.2005 hey sorellina , ma ci sei ? </s>
<s> sei qui , e mi leggi in tempo reale ? </s>
<s> hai visto che mi spariscono le immagini ?? sono mooooolto tristina , e mooolto
inkakkiata xkè non so il motivo... </s>
<s> UFFIIII !!!!! sto bene , hai conosciuto matto , vero ? </s>
<s> simpatico , è molto carino il suo blog. </s>
<s> manchi solo tu... </s>
<s> domani mare , eh ! </s>
<s> eh ! </s>
<s> e tu , come sei messa con casetta ?? venerdì prox vado dal mio amore.
</s>
<s> e ti dirò che è sempre di più il mio amore grande... </s>
<s> wow !! sono felice !! , spero che il mio cuore stia sempre col suo , vicini... </s>
[...]
</text>
```

Fig 6 - Estratto da un testo con tratti tipici della CMC

L'analisi esemplare dei contesti di pagine web in cui si trovano le combinazioni elencate mostra che questa ibridizzazione non concerne necessariamente l'insieme di una pagina web (fig. 6). In una stessa pagina web si trova una vasta variazione di qualità e densità da un autore di contributi all'altro. Lo dimostra l'esempio di ITWAC': un brano da una delle pagine web in cui ricorrono tutti i fenomeni presi in considerazione.

Non è da escludere che la presenza sporadica di tratti tipici della CMC in ITWAC' corrisponda al metodo con cui esso è costruito, cioè con l'inclusione di pagine web che contengono una determinata quota di lessico della lingua standard. Ma questo esempio evidenzia che, quando si tratta di comprendere

il fenomeno dell'ibridizzazione del linguaggio in rete, si devono distinguere non solo macro-contesti come generi e pagine web, ma anche stili individuali.

Per approfondire questo aspetto si dovrebbe passare ad analisi qualitative, che possono essere facilitate dall'identificazione automatica di contesti secondo la qualità e la densità di fenomeni tipici della CMC. Soprattutto la differenziazione all'interno di uno stesso genere come il blog o un forum di discussione può essere oggetto di future ricerche. Gli esempi di ITWAC' ritrovati con i tipi di combinazione citati sopra dànno prova che sia da un forum all'altro che da un autore all'altro si osserva un continuum di utilizzazione dei tratti tipici della CMC. La loro distribuzione – è la nostra ipotesi – non è aleatoria, e sarebbe da esaminare a quale tipo di sito (forum), di tema, di scambio (diretto o discontinuo), o a quale dinamica sociale osservabile nella comunicazione corrisponda l'uso o il non uso dei tratti tipici della CMC.

Desiderata e prospettive

POS-tagging

Le analisi degli errori del POS-tagging hanno mostrato vari aspetti problematici, alcuni dei quali però non sono specifici della lingua della CMC: tra i problemi generali è da sottolineare che le interiezioni e i nomi dovrebbero essere trattati con più attenzione, tramite l'uso di dizionari preparati per questi tipi di unità lessicali. Un aspetto più specifico della CMC, che qui non possiamo approfondire, concerne la segmentazione dei testi in token. Si trovano molteplici esempi di un tipo digitale di *scriptio continua*[51]. Se in futuri corpora web si prenderanno in considerazione in maggior misura anche testi da chat e da altri contesti meno formali, si dovrebbero mettere a punto dei tools per una loro segmentazione assistita dal computer (cf. p.es. Beaufort *et al.* 2010).

Anche la polimorfia grafica invita a un confronto tra la scrittura secondaria della rete e fenomeni grafici del medioevo, sia per le sue dimensioni fonetiche e morfologiche sia per il sistema di abbreviazioni. Indagare le possibili regolarità e i tratti stabili anche della scrittura informale in rete non è solo un interesse di particolare importanza in sé, ma potrebbe aiutare a sviluppare strumenti adatti per il loro trattamento computazionale[52].

---

[51] Cf. p.es. Pistolesi (2004, 102).

[52] I contributi in Kunstmann / Stein (2007) mostrano quali possono essere le dimensioni del trattamento computazionale di questioni linguistiche e di filologia testuale sulla base del *Nouveau Corpus d'Amsterdam*.

Aspetti pragmatici

Fino ad oggi sono stati costruiti diversi corpora della comunicazione chat. Accanto a vari tipi di corpora creati per ricerche più o meno individuali sono stati costruiti anche corpora di maggiori dimensioni, annotati secondo gli standards attuali (XML). Alcuni di questi corpora (Dortmunder Chat Korpus, Corpuseye) sono stati pubblicati online e online possono essere esplorati. Secondo gli obiettivi del corpus, l'annotazione è in alcuni casi di tipo morfosintattico, mentre in altri sono stati annotati anche elementi di valore pragmatico[53]. Sembrano mancare esempi di annotazione polivalente, con un markup di tipo sia morfosintattico sia pragmatico.

Future analisi dei tratti della CMC dovrebbero puntare di più anche sui loro aspetti quantitativi e distributivi, possibilmente considerando fenomeni linguistici a più livelli, p.es. integrando l'analisi delle parti del discorso con uno studio dei tratti diafasicamente marcati della scrittura in rete. Come Biber (1994a) ha evidenziato, la frequenza delle parti del discorso e la loro combinazione sono molto utili per l'identificazione dei registri del parlato e dello scritto e per l'identificazione della loro posizione nel continuum dello spazio comunicativo. Anche la compresenza di tratti morfosintattici e pragmatici potrebbe essere indagata nel contesto di ricerche sulla tipologia testuale del web.

Realtà plurilingui

Nella linguistica dei corpora si discutono – e si sviluppano – vari tipi di corpus anche plurilingui, soprattutto nell'ambito della linguistica applicata (didattica, traduzione di lingue speciali). In questo modo, nella tipologia dei corpora plurilingui si trovano sia corpora paralleli (generalmente testo e testo tradotto, molto spesso da domini specializzati), sia corpora comparabili dal punto di vista pragmatico (funzione e temi) e sociolinguistico (Aston 1999).

Per un corpus composto da testi del web il markup linguistico dovrebbe coprire anche l'aspetto dell'identificazione di lingue diverse. Siccome il plurilinguismo è costitutivo per la lingua della rete, non si può fare a meno di rispettare il carattere mistilingue delle sue parti. Questo vale anche per elementi dialettali che si possono trovare isolati o in contesti dialettofoni veri e propri.

---

[53] È questo il caso del Dortmunder Chatkorpus: «ausgewählte Stilelemente chatbasierter Kommunikation (z.B. Emoticons, Adressierungen, Handlungs- und Zustandsbeschreibungen in Asterisken) wurden in den Daten mittels einer XML-Annotation ausgezeichnet» (Beißwenger 2008, 500).

Ai limiti del linguaggio naturale

Come si è descritto, da itWaC sono stati tolti tutti gli elementi di HTML. Ciononostante, è riconosciuta l'importanza delle informazioni sullo status di un elemento testuale codificato in HTML, p.es. titolo, paragrafo normale, paragrafo di un elenco, link ecc. Se si mantenesse questo tipo di informazione, si potrebbe indagare anche la natura linguistica degli elementi di navigazione. Questi (menu, links) non fanno certamente parte di un linguaggio naturale, ma presentano una variazione interna che può non solo contribuire a caratterizzare diversi tipi di pagine web, ma che dovrebbe essere un oggetto interessante anche per studi diacronici della lingua del web. Quale sarà, p.es., il rapporto tra *clicca* / *click* nei siti italiani fra dieci anni? Inoltre, un'espressione generata automaticamente può entrare nel linguaggio comune e assumere una funzione comunicativa in luogo di una funzione tecnica.

La comunicazione in rete non è solo di natura verbale. Pur non potendo approfondire tale aspetto in questo articolo, ci pare importante tenere a mente che le caratteristiche tipografiche e l'integrazione di immagini e altre illustrazioni sono una parte costituente della semiotica comunicativa in rete (cf. Storrer 2001, 45). Un altro aspetto per cui itWaC non è predisposto è la dimensione dell'interattività (WEB 2.0) che è da prendere in considerazione quando si tratta di sviluppare una tipologia contingente per i testi internet.

Desiderata

In una prospettiva ambiziosa, si potrebbe pensare a un markup a vari livelli, con cui i componenti linguistici e paralinguistici di una pagina web potrebbero essere evidenziati in modo più completo. Nell'insieme sembra che per i corpora linguistici della rete sia di grande importanza portare avanti anche per le lingue romanze lo sviluppo di nuovi sistemi di markup linguistico a livello non solo sintattico e morfosintattico, ma anche testuale (rispettando i vari elementi costituenti come titolo, link, lista, paragrafo, immagine), pragmatico e grafico; inoltre, questi sistemi devono in ogni momento essere aperti a riconoscere elementi di lingue o di varietà diverse.

Se il markup linguistico comprende anche aspetti caratteristici della CMC a livello grafico, lessicale e pragmatico, si potrebbero evidenziare aspetti tipologici ancora più adatti a intendere il processo che è stato chiamato tendenza all'ibridizzazione di scritto e parlato (p.es. Gadet 2008, 527) – e i risultati di questo processo. L'ampiezza di itWaC offre la possibilità di imparare molto sulla qualità di tratti non standard della lingua in rete e sulla loro distribuzione. E questo vale anche per un corpus come itWaC che comporta, come le

analisi della terza parte dimostrano, solo una piccola parte di pagine web con varianti caratteristiche della scrittura secondaria.

L'importanza di internet come risorsa di dati linguistici risulta dalla gamma di registri, dominî, temi e dalla loro attualità e dinamicità. Come si è discusso in questo articolo, la natura digitale dei dati ne facilita la raccolta, ma comporta anche nuovi problemi di differenziazione tra linguaggio naturale e linguaggio generato automaticamente. Un trattamento computazionale potrebbe aiutare a scoprire frequenze di n-grams che indicano un possibile carattere non-naturale (*rispondere citando*).

Da un corpus vasto come ITWAC nascono grandi desiderata che richiedono molto tempo e un lavoro pluridimensionale con metodi linguistici diversi. Anche se non ci si possono aspettare risultati rapidi e completi, un tale tipo di lavoro appare valido in quanto riesce a evidenziare nell'insieme del «noise» della lingua del web una melodia sua propria, polifonica e variegata.

Université de la Ruhr
Bochum

Annette GERSTENBERG

Bibliografia

Anis, Jacques, 2002. «Communication électronique scripturale et formes langagières», versione 31.08.2006, in: *Réseaux Humains / Réseaux Technologiques* 4, ‹http://rhrt.edel.univ-poitiers.fr› (20.03.2010).

Antonini, Anna, 1974-1975. «Il problema del gerundio», *Studi di grammatica italiana* 4, 85-107.

Arnold, Rafael, 2008. «KFÉ, Eskimo», *Irak*, *RJb* 59, 46-70.

Aston, Guy, 1999. «Corpus use and learning to translate», *Textus* 12, 289-314.

Aston, Guy, 2001. «Text categories and corpus users: a response to David Lee», *Language learning & technology* 5/3, 73-76.

Aston, Guy / Piccioni, Lorenzo, 2004. «Un grande corpus di italiano giornalistico», in: Bernini, Giuliano / Ferrari, Giacomo / Pavesi, Maria (ed.), *Atti del 3 congresso di studi dell'Associazione Italiana di Linguistica Applicata*, Perugia, Guerra, 289-310.

Barbera, Manuel / Corino, Elisa / Onesti, Cristina, 2007. «Cosa è un corpus? Per una definizione più rigorosa di corpus, token, markup», in: id. (ed.), *Corpora e linguistica in rete*, Perugia, Guerra, 25-88.

Baroni, Marco / Bernardini, Silvia / Ferraresi, Adriano / Zanchetta, Eros, 2009. «The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora», *Language Resources and Evaluation* 43, 209-226, ‹http://wacky.sslmit.unibo.it/lib/exe/fetch.php?media=papers:wacky2008.pdf› (07.03.2009).

Baroni, Marco, [s.a., 2008]. «ItWaC – Italian Web Corpus», in: Kilgarriff, Adam / Rychlý, Pavel / Pomikálek, Jan, *SketchEngine*, University of Leeds / University of Sussex, Lexical Computing, ‹http://trac.sketchengine.co.uk/wiki/Corpora/ItWaC› (11.03.2008).

Baroni, Marco *et al.*, 2004. «Introducing the ‹la Repubblica› corpus: A large, annotated, TEI(XML)-compliant corpus of newspaper Italian», in: Lino, Maria Teresa, *et al.* (ed.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004, Lisbon, may 26-28)*, Paris, ELRA – European Language Resources Association, 1771-1774.

Beaufort, Richard / Roekhaut, Sophie / Cougnon, Louise-Amélie / Fairon, Cédrick, 2010. «A hybrid rule/model-based finite-state framework for normalizing SMS messages», Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 11-16 July 2010, 770-779. ‹http://www.aclweb.org/anthology/P/P10/P10-1000.pdf› (11.10.2010).

Beißwenger, Michael / Storrer, Angelika, 2008. «Corpora of computer-mediated communication», *HSK* 29.1, 292-309.

Beißwenger, Michael, 2008. «Linguistik im Internet: Korpora zur computervermittelten (internetbasierten) Kommunikation», *Zeitschrift für Germanistische Linguistik* 35/3, 496-503.

Bergenholtz, Henning / Mugdan, Joachim, 1989. «Korpusproblematik in der Computerlinguistik: Konstruktionsprinzipien und Repräsentativität», in: *HSK* 4, 141-149.

Bergh, Gunnar / Zanchetta, Eros, 2008. «Web linguistics», *HSK* 29.1, 309-327.

Bernardini, Silvia / Baroni, Marco / Evert, Stefan, 2006. «A WaCky Introduction», in: Baroni, Marco / Bernardini, Silvia (ed.), *Wacky! Working papers on the Web as Corpus*, Bologna, GEDIT, 9-40.‹http://wackybook.sslmit.unibo.it/pdfs/wackybook.zip› (07.03.2009).

Berruto, Gaetano, 2005. «Italiano parlato e comunicazione mediata dal computer», in: Hölker, Klaus / Maaß, Christiane (ed.), *Aspetti dell'italiano parlato*, Münster, LIT, 137-156.

Biber, Douglas, 1994a. «Representativeness in Corpus Design», in: Zampolli, Antonio / Calzolari, Nicoletta / Palmer, Martha (ed.), *Current issues in computational linguistics: In Honour of Don Walker*, Pisa, Giardini, 377-407.

Biber, Douglas, 1994b, «Using Register-Diversified Corpora for General Language Studies», in: Armstrong, Susan (ed.), *Using large corpora*, Cambridge, MIT Press, 179-201.

Bolasco, Sergio, 2005. «La reperibilità statistica di tendenze diacroniche nell'uso delle parole», in: De Mauro / Chiari, 335-354.

C-ORAL-ROM = Cresti, Emanuela / Moneglia, Massimo (ed.), 2005. *C-ORAL-ROM: integrated reference corpora for spoken romance languages*, Amsterdam, Benjamins.

Chiari, Isabella, 2005. «Linguistica e informatica: la linguistica dei corpora in Italia», *Bollettino di italianistica. Rivista di critica, storia letteraria, filologia e linguistica* 4, 101-118.

Chiari, Isabella, 2007. *Introduzione alla linguistica computazionale*, Bari, Laterza.

Corpus = AAVV, 2007. «Corpus: Bilan et perspectives», *Revue française de linguistique appliquée* 12/1.

de Beaugrande, Robert-Alain / Dressler, Wolfgang, 1981. *Einführung in die Textlinguistik*, Tübingen, Niemeyer.

Debatin, Bernard, 1998. «Analyse einer öffentlichen Gruppenkonversation im Chat-Room. Referenzformen, kommunikationspraktische Regularitäten und soziale Strukturen in einem kontextarmen Medium», in: Prommer, Elizabeth / Vowe, Gerhard (ed.), *Computervermittelte Kommunikation: Öffentlichkeit im Wandel*, Konstanz, 13-37.

Döring, Martin / Osthus, Dietmar / Polzin-Haumann, Claudia (ed.), 2004. *Medienwandelund romanistische Linguistik. Akten der gleichnamigen Sektion des XXVIII. Deutschen Romanistentages (Kiel, 28.9.-3.10.2003)*, Bonn, Romanistischer Verlag.

EAGLES = Sinclair, John, 1996. *EAGLES. Preliminary recommendations on Corpus Typology. EAGLES Document EAG-TCWG-CTYP/P*, pdf-Dokument. University of Birmingham, School of English.

Ernst, Gerhard, 2009. «Googlemetrie?! Versuch einer quantitativen Erfassung der Akzeptanz von Orthographiereformen (Deutsch, Französisch, Rumänisch)», in: Gabriele Blaikner-Hohenwart *et al.* (ed.), *Ladinometria. Festschrift für Hans Goebl zum 65. Geburtstag*, vol. 2. Salzburg *et al.*, Universität, 43-60.

Evert, Stefan, 2009. «Corpora and Collocations», *HSK* 29.2, 1212-1248.

Ferrari, Angela / Zampese, Luciano, 2006. «Aperture al gerundio: valori modali e configurazioni informative», *Cuadernos de filología italiana* 13, 46-71.

Frei, Henri, 2007. *La grammaire des fautes*, ed. orig. 1929. Rennes, ennoïa.

Gadet, Françoise, 2008. «Ubi scripta et volant et manant», in: Stark, Elisabeth (ed.), *Romanische Syntax im Wandel*, Tübingen, Narr, 513-529.

Gansel, Christina, 2002. «Verba dicendi», *HSK* 21.2, 1562-1569.

Gerstenberg, Annette, 2004. «‹Digitare in Piazza›. Zur Sprache im italienischen Chat», in: Dahmen, Wolfgang / Holtus, Günter / Kramer, Johannes / Metzeltin, Michael / Schweickard, Wolfgang / Winkelmann, Otto (ed.), *Romanistik und neue Medien. Romanistisches Kolloquium XVI*, Tübingen, Narr, 309-326.

Gerstenberg, Annette, 2009. *Arbeitstechniken für Romanisten. Eine Anleitung für den Bereich Linguistik*, Tübingen, Niemeyer.

Giltrow, Janet / Stein, Dieter, 2009. «Genres in the Internet. Innovation, evolution, and genre theory», in: id. (ed.), *Genres in the Internet*, Amsterdam, Benjamins, 1-25.

Grands Corpus = AAVV, 1999. «Dossier: Grands Corpus: diversité des objectifs, variété des approches», *Revue française de linguistique appliquée* 4/1.

Habert, Benoît / Nazarenko, Adeline / Salem, André, 1997. «Introduction», in: id., *Les linguistiques de corpus*, Paris, Colin, 7-18.

Hathout, Nabil / Montermini, Fabio / Tanguy, Ludovic, 2008. «Extensive data for morphology: using the World Wide Web», *FLS*, 18/01, 67-85.

Heid, Ulrich, 2008. «Corpus linguistics and lexicography», *HSK* 29.1, 131-153.

ITWAC = WaCky, 2009. *Corpora: itWaC*, [Korpusdownload: 07.03.2009]. Bologna, School of Modern Languages for Interpreters and Translators, Università di Bologna, ‹http://wacky.sslmit.unibo.it/doku.php?id=corpora› (27.11.2009).

Jakobs, Eva-Maria, 2003. «Hypertextsorten», *Zeitschrift für Germanistische Linguistik* 31/2, 232-252.

Kilgarriff, Adam, 2007. «Googleology is bad science», *Computational Linguistics* 33/1, 147-151.

Kilgarriff, Adam / Grefenstette, Adam, 2003. «Web as Corpus», *Computational Linguistics* 29/3, 1-15.

Kilgarriff, Adam / Rychly, Pavel / Smrz, Pavel / Tugwell, David, 2004. «The Sketch Engine», in: Williams, Geoffrey / Vessier, Sandra (ed.), *Proceedings of the 11 Euralex International Congress*, Lorient, Université de Bretagne-Sud, Faculté de Lettres et de Sciences Humaines, 104-116.

Koch, Peter / Oesterreicher, Wulf, 2007. *Lengua hablada en la Romania*, Madrid, Gredos.

Kunstmann, Pierre / Stein, Achim (ed.), 2007. *Le Nouveau Corpus d'Amsterdam. Actes de l'atelier de Lauterbad, 23-26 février 2006*, Stuttgart, Steiner *La Repubblica Corpus*, Bologna, Scuola Superiore di Lingue Moderne per Interpreti e Traduttori / Università di Bologna 2004. ‹http://dev.sslmit.unibo.it/corpora/corpus.php?path=&name=Repubblica› (28.05.09).

Lee, David YW, 2001. «Genres, Registers, Text Types, Domain, and Styles: Clarifying the Concepts and Navigating a Path Through the BNC Jungle», *Language Learning and Technology* 5/3, 37-72.

Lobin, Henning (ed.), 2001. *Text im digitalen Medium. Linguistische Aspekte von Textdesign, Texttechnologie und Hypertext Enginering*, Opladen, Westdeutscher Verlag.

Lonzi, Lidia, 1991. «Frasi subordinate al gerundio», in: Renzi, Lorenzo / Salvi, Giampaolo (ed.), *Grande grammatica italiana di consultazione, [vol. 2 recte]: I sintagmi verbale, aggettivale, avverbiale. La subordinazione*, Bologna, Mulino, 571-592.

Lorenzetti, Luca / Schirru, Giancarlo, 2006. «La lingua italiana nei nuovi mezzi di comunicazione: SMS, posta elettronica e Internet», in: Gensini, Stefano (ed.), *Fare comunicazione*, Roma, Carocci, 71-98.

Lüdeling, Anke / Evert, Stefan / Baroni, Marco, 2007. «Using web data for linguistic purposes», in: Hundt, Marianne (ed.), *Corpus linguistics and the web*, Amsterdam, Rodopi, 7-24.

Mehler, Alexander / Gleim, Rüdiger, 2006. «The Net for the Graphs: Towards Webgenre Representation for Corpus Linguistic Studies», in: Bernardini / Baroni, 191-224.

Pierozak, Isabelle, 2000. «Approche sociolinguistique des pratiques discursives sur internet ‹ge fé dais fotes si je voeux›», *Revue française de linguistique appliquée* 5/1, 89-104.

Pierozak, Isabelle, 2003. *Le français tchaté. Une étude en trois dimensions – sociolinguistique, syntaxique et graphique – d'usages IRC*, 3 vol. Pdf. Université de Picardie [Université d'Aix-Marseille], Laboratoires d'Études Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique, ‹http://www.u-picardie.fr/LESCLaP/spip.php?rubrique43› (12.03.2008).

Pistolesi, Elena, 1997. «Il visibile parlare di IRC (Internet Relay Chat)», *Quaderni del Dipartimento di Linguistica – Università di Firenze* 8, 213-246.

Pistolesi, Elena, 2004. *Il parlar spedito. L'italiano di chat, e-mail e SMS*, Padova, Esedra.

Policarpi, Gianna / Rombi, Maggi, 1983. «Altre metodologie per la sintassi: tipi di gerundio e tipi di participio», in: Albano Leoni, Federico, *et al.* (ed.), *Italia linguistica: idee, storia, strutture*, Bologna, Il Mulino, 309-331.

R = R Development Core Team, *R: A Language and Environment for Statistical Computing (R 2.7.1)*, Vienna, R Foundation for Statistical Computing 2008. ‹http://www.r-project.org› (04.07.2008).

Rehm, Georg, 2001. «Automatische Textannotation. Ein SGML- und DSSSL-basierter Anastz zur angewandten Textlinguistik», in: Lobin, 179-195.

Rehm, Georg, *et al.*, 2008. «Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems», in: Proceedings of LREC 2008, May 28-30, Marrakech, Morocco.

Santini, Marina, 2007. *Automatic Identification of Genre in Web Pages*, Thesis submitted for the degree of Doctor of Philosophy. Brighton (UK), University of Brighton.

Schafroth, Elmar, 2009. «Wörterbücher des Italienischen im Vergleich. Zur aktuellen Situation der italienischen Lexikographie», *Italienisch* 61, 72-93.

Schmitz, Ulrich, 2009. «Schrift an Bild im World Wide Web. Articulirte Pixel und die schweifende Unbestimmtheit des Vorstellens», in: Linguistik-Server Essen (LINSE): *Publikationen*, Essen, Universität Duisburg-Essen, Campus Essen, 1-29, ‹http://www.linse.uni-due.de/linse/publikationen/Aufsatz Schmitz/Aufsatz Schmitz2009.pdf› (29.10.2009).

Searle, John Roger, 1991 [1965]. «What Is a Speech Act?», in: Davis, Steven (ed.), *Pragmatics. A reader*, New York, Oxford University Press, 254-264.

Shepherd, Michael / Watters, Carolyn, 1998. «The Evolution of Cybergenres», in: AAVV: *Proceedings of the Thirty-First Annual Hawaii International Conference on System Sciences*, vol. 2, [1-13].

Sinclair, John, 2005. «Corpus and Text – Basic Principles», in: Wynne, 1-12.

Storrer, Angelika, 2001. «Kohärenz in Text und Hypertext», in: Lobin, 33-65.

Tognini-Bonelli, Elena, 2001. *Corpus linguistics at work*, Amsterdam, Benjamins.

TUSTEP: Tübinger System von Textverarbeitungsprogrammen, Tübingen, ZDV 2010. ‹http://www.tustep.uni-tuebingen.de/› (31.01.2010).

WaCky = Baroni, Marco, *et al.*, *WaCky*, Bologna, Scuola superiore di interpreti e traduttori 2008-2010. ‹http://wacky.sslmit.unibo.it› (20.03.2010).

WebGenreWiki = Santini, Marina / Sharoff, Serge / Rehm, Georg / Mehler, Alexander (ed.), *Web Genre Wiki*, s.l., s.d. [2010]. ‹http://www.webgenrewiki.org› (26.03.2010) dp3.

Wenz, Karin, 1998. «Formen der Mündlichkeit und Schriftlichkeit in digitalen Medien», *Linguistik Online* 1/1, ‹http://viadrina.euv-frankfurt-o.de/~wjournal/wenz.htm> (25/05/2000).

Wynne, Martin (ed.), 2005. *Developing Linguistic Corpora: a Guide to Good Practice*, Oxford, Oxbow Books, ‹http://ahds.ac.uk/linguistic-corpora› (31.05.2009).

Zampesi, Luciano, 2004. «Aspetti semantico-testuali del grundio modale in apertura di frase», in: Ferrari, Angela (ed.), *La lingua nel testo, il testo nella lingua*, Torino, Istituto dell'Atlante Linguistico Italiano, 79-116.

Zanchetta, Eros / Baroni, Marco, 2005. *Morph-it! A free corpus-based morphological resource for the Italian language*, *Proceedings of Corpus Linguistics 2005*, Birmingham, University of Birmingham.

Zingarelli = Zingarelli, Nicola, 2009. *Lo Zingarelli 2010. Database online*, Bologna, Zanichelli.