Una nota sul dibattito relativo alla nozione di fonema

Andrea Paoloni

Il tema del dibattito tra Giovanna Marotta (GM) e Federico Albano Leoni (FAL) su un modello di analisi del linguaggio articolato ha da sempre suscitato il mio interesse. Molti anni fa cercavo di realizzare un sistema di riconoscimento del parlato (*ASR*: *Automatic Speech Recognition*) per l'italiano e ricordo che molti studi erano orientati ad individuare i cosiddetti invarianti. Il paradigma adottato era il seguente: se un uomo è in grado di riconoscere i fonemi, esisteranno, con ogni probabilità, alcune loro caratteristiche non influenzate dalle variazioni dovute al diverso parlante e al diverso modo di parlare (un interessante dibattito sul tema è contenuto in Perkell e Klatt, 1986). Si pensava che fosse possibile trovare una qualche funzione in grado di trasformare il segnale acustico in simboli collegati coi fonemi costituenti la frase da identificare. In seguito ad un incontro con i ricercatori del centro di fonetica sperimentale di Padova mi dedicai ad ulteriori studi di fonetica e fonologia ed in particolare soffermai la mia attenzione sulla teoria dei tratti distintivi: pensai che, identificando nel segnale tali tratti, sarebbe stato facile procedere alla trascrizione in alfabeto fonetico e conseguentemente alla trascrizione ortografica. Mi accorsi presto però che l'ipotesi da me formulata non era perseguibile perché i tratti non avevano corrispondenze fisiche (acustiche), e inoltre la tabella di definizione presentava dei vuoti: tali vuoti erano dovuti, mi spiegò il professor Trumper, alla non pertinenza di alcuni tratti per alcuni fonemi della lingua italiana. Inoltre osservai che i tratti distintivi dei suoni vocalici erano *diversi* da quelli dei suoni consonantici. Come in altre occasioni scoprii che la formalizzazione, la *prospettiva simbolica astratta* (Marotta, 2010: *passim*), era fallace in quanto non era possibile realizzare un programma in grado di utilizzare la formalizzazione proposta al fine di classificare i segnali.

Anche il modello della sintassi proposto da Chomsky non ha risolto il problema dell'analisi linguistica del parlato: infatti i sistemi di dettatura (*Via Voice*, *Dragon Dictate*, *Philips*) non fanno uso di un tale modello affidandosi, nel componente linguistico, ad un approccio puramente statistico

Ricevuto: Aprile 2012
Accettato: Settembre 2012

(Roe e Wilpon, 1993; De Mori, 2008; Juang, 1998). Peraltro anche nell'analisi del testo il modello proposto da Chomsky, che tante aspettative aveva suscitato per la sua impostazione rigorosa con precise indicazioni operative, non ha condotto allo sviluppo di sistemi di analisi del testo (*parser*) efficienti ed efficaci, e molti dei sistemi di analisi oggi utilizzati non possono fare a meno del contributo della statistica (Kay, 2011; Johnson, 2011).

Ma sugli insuccessi predittivi delle teorie linguistiche e sul successo delle teorie statistiche torneremo più avanti. Parliamo ora di alcune prime impressioni derivate dalla lettura del commento di Marotta (2010). Un primo punto degno di nota è l'osservazione che la teoria motoria trovi una qualche conferma dalla scoperta dei neuroni specchio: pensiamo sia un'osservazione interessante, che mette in evidenza come proprio dallo studio del comportamento cerebrale si potrebbero ottenere risposte sul funzionamento del sistema 'comunicazione verbale' nell'uomo. Si noti però che, essendo l'elaborazione del cervello fortemente parallela (a differenza di quella seriale dei computer), l'osservazione rafforzerebbe le teorie basate sugli *exemplars*: «nelle teorie basate sugli *exemplars*, i processi dell'elaborazione dell'informazione sono soltanto paralleli e soggetti a restrizione; non sono quindi né seriali né governati da regole» (Marotta, 2010: 294).

Un altro punto degno di nota è quello dove si afferma che: «la prosodia può trasmettere solo una gamma limitata di significati, molti dei quali non appartengono alla sfera linguistica in senso stretto, cioè sono extralinguistici» (Marotta, 2010: 288). Come esempio di significato extralinguistico viene proposto l'identificazione dello stato emozionale, ma riteniamo che un altro valido esempio di significato extralinguistico che si sarebbe potuto proporre è l'identificazione del parlante. Questa altra informazione extralinguistica, ovvero l'identità del parlante, viene veicolata principalmente dai suoni vocalici, e allora dovremmo dedurre che i suoni vocalici, che peraltro portano informazioni linguistiche ridondanti, «trasmettono solo una gamma limitata di informazioni, molte delle quali non appartengono alla sfera linguistica in senso stretto, cioè sono extra linguistiche?».

Ci sembra che le informazioni convogliate dalla prosodia siano fondamentali per la comprensione del parlato e non solo per l'identificazione dello stato emotivo del parlante.

Veniamo ora al punto in cui vengono presentati «due argomenti forti che testimoniano l'autonomia degli elementi fonologici indipendentemente dalla parallela presenza di lettere che li esprimono sul piano della scrittura» (Marotta, 2010: 290). Su questo punto concordo con FAL: a mio avviso

entrambi gli argomenti (che peraltro sono riducibili ad uno: esistono persone che parlano pur non conoscendo né la scrittura né la lettura), testimoniano l'opposto di quanto affermato da GM: le persone comunicano tra loro generando suoni che identificano oggetti o situazioni ma questi suoni sono le parole, non i loro costituenti, siano essi fonemi o sillabe. Per quanto attiene ai bambini che imparano a parlare il fenomeno detto lallazione può essere così descritto: «verso i 10-12 mesi la maggior parte dei bambini produce strutture sillabiche complesse che caratterizzano la cosiddetta lallazione variata (per es. *dadu*). Sempre a questa età compaiono i primi suoni simili a parole che, pur avendo una forma fonetica identica, assumono un significato specifico quando vengono utilizzate consistentemente in determinati contesti (ad es. il suono *nanà* prodotto in una situazione di richiesta). In genere, nei bambini che imparano l'italiano come lingua madre, la forma delle proto-parole è CV x 2 (come *tata* o *papa*)» (Sala, 2010: *www.assomensana.it/Comunicazione-e-disturbi/suoni-vocalizzi-lallazioni.php*).

Pur non essendo un esperto dello studio dell'apprendimento dei bambini mi sembra di poter ribadire che quello che apprendono sono parole e non certamente fonemi. Peraltro i fonemi consonantici non possono essere pronunciati isolatamente e conseguentemente si pronuncia comunque una sillaba. Non è sostenibile l'assunto che prima si apprendano i fonemi, poi le sillabe ed infine le parole: i bambini imitano la voce dei genitori e cercano di formulare i suoni linguistici di maggiore loro interesse: *mamma, pappa*, etc. Alcuni suoni di difficile realizzazione vengono, in una prima fase, sostituiti da suoni affini: /'ak:a/ per /'ak:wa/, /'loma/ per /'roma/, etc. Esattamente nello stesso modo i bambini imparano a cantare senza necessariamente conoscere l'esistenza delle note. Sostenere che l'esistenza della scala tonale sia dimostrata dal fatto che canti anche chi non conosce la notazione musicale, non mi sembra ragionevole. Ritornando alla nozione di fonema possiamo sicuramente dire che è un modello efficiente di notazione del parlato, così come la scala tonale è un modello efficiente di notazione della musica europea, ma l'impossibilità di definirne compiutamente le caratteristiche acustiche consente qualche dubbio sulla sua realtà ontologica, sulla sua esistenza come mattone con il quale viene costruito il linguaggio articolato.

Un ulteriore argomento di cui non comprendo la valenza è quello in cui si afferma che l'utilità dell'analisi fonetica è collegata alla capacità di portare evidenza empirica ad un'ipotesi precedentemente formulata, ipotesi che potrebbe essere smentita dai dati. «Un esempio concreto tratto dalle indagini che sto attualmente conducendo sulla nozione di prominenza e i suoi corre-

lati fisico-acustici: il parlante sa, in senso chomskiano classico, come e dove segnalare la prominenza, anche plurima, nell'ambito dell'enunciato. [...] La verifica linguistica di quanto si legge sullo spettro o sulla forma d'onda deve tuttavia essere necessariamente tipo percettivo» (Marotta, 2010: 296). Cosa significa che la verifica deve essere di tipo percettivo e quindi soggettivo? Non dovrebbe avvenire su dati strumentali oggettivi? La verifica dovrebbe basarsi, a mio avviso, sulla implementazione di un programma in grado di identificare automaticamente la prominenza con un moderato tasso di errore.

Un punto che mi ha particolarmente colpito è quello dove si afferma: «ma questo percorso è valido se la lingua è un sistema comunicativo più che una struttura logico-grammaticale» (Marotta, 2010: 296). Ora io avevo sempre creduto che la lingua fosse un sistema comunicativo particolarmente efficiente, in grado di dare all'animale che la possiede (l'uomo) un vantaggio competitivo di straordinaria potenza. Cosa sia e a cosa serva la struttura logico-grammaticale 'lingua' se non come modello della realtà linguistica, nello stesso modo in cui i modelli fisici della realtà vengono utilizzati per comprenderla, misurarla e modificarla, non mi è dato comprendere.

A chiusura di queste osservazioni vorrei riprendere il discorso sugli insuccessi delle teorie linguistiche quando si scontrano con la possibilità di implementarle in un automa. I primi tentativi di realizzare macchine parlanti e macchine in grado di comprendere il parlato si sono basati, come sempre avviene, sul funzionamento della macchina uomo. La macchina di von Kempelen cercava di riprodurre l'apparato fonatorio; i primi sistemi di riconoscimento delle cifre utilizzavano filtri e sistemi di segmentazione del parlato che in qualche modo volevano riprodurre il funzionamento dell'orecchio (cfr. Klatt, 1987; Pettorino e Giannini, 1999; *YouTube*: «*Kempelen's speaking machine*»).

Queste macchine, basate sull'imitazione dell'apparato vocale umano, hanno avuto lo stesso insuccesso dei sistemi di volo basati sull'imitazione del volo degli uccelli: come tutti sanno l'aeroplano non batte le ali. I sistemi di riconoscimento del parlato oggi in uso utilizzano due componenti, denominati analizzatore acustico-fonetico e analizzatore linguistico. Entrambi i componenti, come abbiamo già detto, non si basano su modelli teorici di tipo linguistico, bensì su modelli statistici addestrati su un *corpus* di segnale vocale di adeguata dimensione. Non possiamo non ricordare quanto diceva Fred Jelinek (1998): «Every time I fire a linguist, the performance of the speech recognizer goes up».

Anche i sistemi di sintesi da testo (*TTS*: *Text-To-Speech*) non fanno certo uso del modello fonema. Utilizzando tale modello, nonostante l'applicazione di algoritmi che simulano la coarticolazione tra i fonemi adiacenti, non si riesce ad ottenere una qualità di parlato accettabile. La strada adottata è stata dapprima quella di usare unità della dimensione della sillaba (*difoni*), per poi passare ad unità sempre più grandi quali parole o gruppi di parole, come ad esempio: *la casa*, *al mio*, *di più*, etc. (cfr. Taylor, 2009).

I sistemi di riconoscimento e sintesi della voce fanno uso di modelli statistici e non di modelli linguistici. Poiché tali sistemi calcolano la probabilità sulla base degli esempi tratti da un *corpus* avanziamo l'ipotesi che la teoria linguistica che suggeriscono sia la teoria dei prototipi o la *exemplar theory*. Inoltre poiché gli *ASR* calcolano la probabilità che si sia pronunciata una certa frase componendo le informazioni di tipo acustico con quelle di tipo lessicale, suggeriscono una qualche ipotesi 'gestaltica' del riconoscimento.

I sistemi di traduzione automatica, dopo aver a lungo perseguito l'approccio basato sulla conoscenza (*Rule-based Machine Translation*), implementando *parser* sintattici e lemmatizzatori, hanno trovato poi miglior soluzione nell'approccio statistico (*SMT*: *Statistical Machine Translation*) o meglio nell'approccio *exemplar-based* (Koehn, 2010; *Moses Core Project 2012*: *www.statmt.org/mosescore/*). In altri termini, i sistemi di traduzione oggi disponibili basano il loro funzionamento su grandi *corpora* paralleli (di cui uno è la traduzione dell'altro) e cercano in tali *corpora* la frase da tradurre e la corrispondente frase tradotta. Su questa base funziona il traduttore di *Google* che, seppure con vistose lacune, tuttavia è in grado di farci comprendere di cosa si parli nella lingua a noi ignota. Non ci sorprende che questo approccio empirico abbia ottenuto risultati migliori di quello basato su regole, perché ciò è quanto già avvenuto nel campo della sintesi della voce (*TTS*), dove si è compreso che per migliorare la qualità era necessario unire tratti di segnale sempre più grandi in modo che contenessero al loro interno quelle variazioni locali continue che non sono realizzabili giustapponendo segmenti discreti. Basandosi su *corpora* paralleli di opportune dimensioni è possibile trovare la traduzione più appropriata per un numero sempre più elevato di frasi senza utilizzare alcun modello linguistico esplicito. L'ipotesi gestaltica viene avvalorata anche dagli studi sull'intelligibilità del parlato quando fortemente degradato da rumore; tali studi dimostrano che la comprensione è legata alla prevedibilità della frase, o del frammento di frase, da identificare (Romito, 2005). Senza entrare nel merito di come questo avvenga, perché lo ignoro, sembra che la comprensione sia possibile integrando gli stimoli acustici, in particolare

quelli prosodici, con informazioni sulla lingua utilizzata e sul contesto della comunicazione. Ascoltando più e più volte un segnale rumoroso ad un tratto si riesce a decrittare una frase e quando ciò è avvenuto, sulla base del testo decrittato l'ascolto di conferma è poi facile. Anche questo fenomeno induce a pensare ad una sorta di *Gestalt*, dove una volta che siamo riusciti a separare il segnale utile dallo sfondo rumoroso, l'operazione viene agevolmente ripetuta. Il fenomeno è facilmente verificabile a livello visivo: nelle immagini ambigue nelle quali è possibile visualizzare due diverse figure, se ne visualizziamo una abbiamo poi difficoltà ad identificare l'altra (Massironi, 1998). Naturalmente questo processo gestaltico può condurre a clamorosi fraintendimenti quando qualche presupposto utilizzato nella decrittazione si dimostra errato, come è avvenuto in un esperimento da me condotto nel quale avevo lasciato credere che la frase da riconoscere fosse in italiano mentre era in arabo: tutti gli ascoltatori hanno fornito trascrizioni, in parte simili, in italiano. Nello stesso esperimento, un altro segnale da decrittare era stato costruito partendo da un'onda pseudo-laringea che nell'arco di circa un secondo variava da 115 a 145 Hz; tale forma d'onda era stata ulteriormente manipolata inserendo tratti fortemente attenuati ogni 200 ms circa ed equalizzandola in modo da esaltare le zone 'formantiche'. Anche questo suono è stato trascritto dal gruppo di ascolto, formato da esperti trascrittori, con frasi (*in italiano*) o tratti di esse che presentavano tra loro alcune significative similitudini ad ulteriore prova che anche in assenza di fonemi si possono 'sentire' parole (Paoloni e Zavattaro, 2007). Inutile dire che se avessi sottoposto il materiale ad esperti spagnoli avrei ottenuto trascrizioni nella loro lingua.

Mi stupisce che nel dibattito gli autori non abbiano fatto cenno alle indicazioni che vengono dai numerosi sistemi commerciali come il trascrittore della *Dragon* (*Dragon dictate*), il traduttore della *Google*, o il sistema di dialogo della *Apple* (*SIRI*), tutti in grado di fornire buone prestazioni senza far uso di modelli linguistici teorici espliciti. Naturalmente le informazioni sulla lingua sono presenti nei *corpora* utilizzati per addestrare i modelli statistici, che a loro volta sono stati opportunamente adattati per migliorare le prestazioni in termini di accuratezza e tempo di calcolo (Rabiner e Juang, 1993; Jelinek, 1998; Jurafsky e Martin, 2008). In conclusione, a mio avviso, così come avviene per le scienze fisiche, un modello teorico della lingua dovrebbe confrontarsi con la sua implementazione su un sistema automatico, e la validità del modello dovrebbe dipendere dal risultato dell'implementazione. In altri termini l'implementazione potrebbe essere la via per 'falsificare' il modello teorico proposto.

Bibliografia

Abney, S. (2011), *Data-Intensive Experimental Linguistics*, in «Linguistic Issues in Language Technology», 6, 2, pp. 1-27.

Anusuya, M.A. e Katti, S.K. (2009), *Speech Recognition by Machine: A Review*, in «International journal of computer science and information security», 6, 3, pp. 181-205.

Baldwin, T. e Kordoni, V. (2011), *The Interaction between Linguistics and Computational Linguistics*, in «Linguistic Issues in Language Technology», 6, 1, pp. 1-6.

Cater, J.P. (1983), *Electronically Speaking: Computer Speech Generation*, Howard W. Sams & Co., Indianapolis.

Nivre, J., Hall, J., Kubler, S., McDonald, R., Jens Nilsson, J., Riedel, S. e Yuret, D. (2007), *The CoNLL (2007) Shared Task on Dependency Parsing*, in Eisner, J. (2007, *ed.*), *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007 (Prague, June 2007)*, Association for Computational Linguistics, pp. 915-932.

De Mori, R. (2008), *Stato dell'arte e prospettiva della comunicazione uomo macchina*, Fondazione Ugo Bordoni, Roma.

Hajičová, *E.* (2011), *Computational Linguistics without Linguistics? View from Prague*, in «Linguistic Issues in Language Technology», 6, 6, pp.1-20.

Jelinek, F. (1998), *Statistical Methods for Speech Recognition (Language, Speech, and Communication)*, The MIT Press, Cambridge MA.

Johnson, M. (2011), *How relevant is linguistics to computational linguistics?*, in «Linguistic Issues in Language Technology», 6, 7, pp. 1-23.

Juang, B.H. e Chen, T. (1998), *The past, present, and future of speech processing*, in «IEEE Signal Processing Magazine», 15, 3, pp. 24-48.

Jurafsky, M. e Martin, J.H. (2008), *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, Upper Saddle River.

Kay, M. (2011), *Zipf's Law and l'Arbitraire du Signe*, in «Linguistic Issues in Language Technology», 6, 8, pp.1-25.

Klatt, D. (1987), *Review of text-to-speech conversion for English*, in «Journal of the Acoustical Society of America», 82, 3, pp. 737-793.

Koehn, P. (2010), *Statistical machine Translation*, Cambridge University Press, Cambridge UK.

Lee, C.H., Soong, F.K. e Paliwal, K.K. (1996, *eds.*), *Automatic Speech and Speaker Recognition: Advanced Topics*, Kluwer, Boston.

Marotta, G. (2010), *Sulla (presunta) morte del fonema*, in «Studi e Saggi Linguistici», 48, pp. 283-304.

Massironi, M. (1998), *Fenomenologia della percezione visiva*, Il Mulino, Bologna.

Paoloni, A. e Zavattaro, D. (2007), *Intercettazioni telefoniche e ambientali. Metodi, limiti e sviluppi nella trascrizione e verbalizzazione*, Centro Scientifico Editore, Torino.

Perkell, J.S. e Klatt, D.H. (1986, *eds.*), *Invariance and variability in speech process*, Lawrence Erlbaum Associates, London.

Pettorino, M. e Giannini, A. (1999), *Le teste parlanti*, Sellerio Editore, Palermo.

Rabiner, L. e Juang, B.H. (1993), *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs.

Roe, D.B. e Wilpon, J.G. (1993), *Whither Speech Recognition: The Next 25 years*, in «IEEE Communications Magazine», 31, 11, pp. 54-62.

Romito, L. (2005), *Il contesto, l'intelligibilità, il rapporto segnale-rumore*, in Cosi, P. (2005, a cura di), *Misura dei parametri. Aspetti tecnologici ed implicazioni nei modelli linguistici* (Atti del Primo Convegno AISV - Associazione Italiana di Scienze della Voce, Padova 2-4 dicembre 2004), EDK Editore, Brescia, pp. 539-566.

Taylor, P. (2009), *Text-to-speech Synthesis*, Cambridge University Press, Cambridge UK.

Warren, R.M. e Warren, R.P. (1970), *Auditory illusions and confusions*, in «Scientific American», 223, pp. 30-36.

Andrea Paoloni
Fondazione Ugo Bordoni
Viale del Policlinico 147
00161 Roma (Italy)
pao@fub.it