Italian Journal of Linguistics
Rivista di linguistica

volume 13, numero 1, 2001

Sommario

Per corrispondenza e lavori proposti per la stampa indirizzare a: Pier Marco Bertinetto, Redazione di *Rivista di Linguistica,* Scuola Normale Superiore, Piazza dei Cavalieri, 7 - I-56126 Pisa (Italia) (Tel. 050/509111, e-mail: bertinet@sns.it; Fax 050/563513).
Le CONVENZIONI TIPOGRAFICHE sono consultabili presso <http://alphalinguistica.sns.it/RdL/stylesheet.html>.
Periodicità: semestrale. Abbonamento 2002: Italia € 67; Estero € 83 (spese postali comprese).
Per abbonamenti, cambi di indirizzo, informazioni indirizzare a: Pacini Editore - via Gherardesca, 1 - 56121 Ospedaletto (Pisa), Italia (Tel. 050/313011; Fax. 050/ 3130300).

Correspondence concerning contributions and editorial matters should to be addressed to: Pier Marco Bertinetto, Redazione di *Rivista di Linguistica,* Scuola Normale Superiore, Piazza dei Cavalieri, 7 - I-56126 Pisa, (Italy) (Tel. +39/050/509111; e-mail: bertinet@sns.it; Fax +39/050/563513).
The STYLE SHEET is available at <http://alphalinguistica.sns.it/RdL/stylesheet.html>.
Periodicity: twice a year. Subscription rates for 2002: Italy € 67; elewhere € 83 (packing and surface mail postage included). Advance payments required. Available on standing order.
All communications concerning subscriptions should be addressed to the Publisher: Pacini Editore - via Gherardesca, 1 - 56121 Ospedaletto (Pisa), Italy - Tel. +39/050/313011 - Fax. +39/050/3130300.

Autorizzazione del Tribunale di Torino n. 3948 del 9.6.1988; direttore responsabile: Pier Marco Bertinetto.
Impaginazione e stampa: Pacini Editore - Ospedaletto (Pisa)

Pubblicato con il contributo del CNR

La Cancellazione del Complementatore che/that

Carlo Conni

Il presente lavoro si propone di offrire una soluzione sintattica al fenomeno della cancellazione del complementatore (CD) nei termini dell'incorporazione della proiezione sincretica in cui il complementatore è generato. Secondo le linee fondamentali dell'analisi di Rizzi (1997), è possibile associare ad ogni proiezione che possiede tratti di accordo una proiezione pura Agr dove possa verificare i suoi tratti. Assumendo questa ipotesi, insieme ad un'articolazione fine del complementatore concepito come interfaccia fra la parte alta della frase, la subordinata e il contesto linguistico ed extra-linguistico, si prevede la possibilità di una proiezione sincretica di Forza e Finitezza dotata sia di tratti operatoriali che di accordo. L'obiettivo è quello di fornire un modello esplicativo unitario del fenomeno della CD capace di spiegare i differenti livelli di preferibilità e opzionalità del fenomeno, in inglese ed in italiano, sulla base delle differenti modalità di verifica dei tratti operatoriali e di accordo di *che* e *that*. La proposta sintattica di questo lavoro viene sviluppata nei §§ 4-5 parallelamente ad un'analisi degli aspetti più rilevanti della ricerca di Rizzi (1997), ed è preceduta, nel § 1, da alcune sintetiche indicazioni concernenti la tipologia dei contesti che ammettono cancellazione. Una discussione generale di alcune strategie alternative è sviluppata nel § 2, dove si discute l'ipotesi di Poletto (1994) come movimento del verbo in Comp, e nel § 3 attraverso un confronto con l'ipotesi della CD di Giorgi & Pianesi (1997) come movimento del congiuntivo in una proiezione ibrida Mood/Agr [1].

1. Dati ed ipotesi iniziali

Questa ricerca presenta un'ipotesi teorica esplicativa concernente un fenomeno dell'italiano standard e ancora più diffuso nell'inglese: la complementizer deletion o CD, un elemento che abitualmente introduce frasi subordinate a flessione finita, anche denominato 'congiunzione subordinante' o più semplicemente introduttore di frasi a tempo finito, relative, completive, soggettive o di altro tipo [2]. Considerazioni di natura distribuzionale hanno spinto numerosi linguisti a non accettare un'analisi di questo costituente come un pronome relativo. In italiano, si può osservare che mentre *cui* e *quale* si

Rivista di Linguistica, 13.1 (2001), p. 3-43 *(ricevuto nel settembre 2000)*

alternano liberamente fra di loro, non possono mai alternarsi con *che* negli stessi contesti

(1) La persona a cui / alla quale /* a che ho telefonato è Gianni

Il fenomeno della cancellazione del complementatore presenta delle variazioni dall'inglese all'italiano; lo scopo del presente lavoro sarà precisamente quello di cercare di fornire un quadro unitario capace di trattare queste variazioni di comportamento.

In modo preliminare, si può osservare come in italiano la cancellazione di *che* sembri essere condizionata da almeno tre fattori (i) dalla tipologia del verbo della frase principale, (ii) dal modo o dal tempo selezionato nella frase subordinata – il congiuntivo – che in alcuni rari casi può essere il condizionale o il tempo futuro [3], (iii) e, più specificamente, dal fatto che la frase subordinata non sembra richiedere una forza grammaticale autonoma e indipendente da quella della principale, come si verifica tipicamente nelle subordinate al modo congiuntivo [4].

Sebbene si verifichino delle differenze nei giudizi di accettabilità da parte dei parlanti a seconda delle aree regionali di provenienza - è il caso di frasi come (2b, 2l, 2m) quando il soggetto è realizzato in posizione preverbale [5] – affinché si manifesti CD il verbo deve appartenere ad una determinata classe mentre la frase solo in rarissimi casi può essere dipendente da un nome o da un aggettivo [6]:

(2) a. Credo (che) sia arrivato ieri
b. Immagino (che) Luigi abbia già avvisato Maria
c. Credo (che) sarà interessante
d. Suppongo (che) aiuterebbe anche Maria
e. Il fatto *(che) sia partito non vuole dire niente
f. La probabilità ?(che) si tratti di un errore è molto alta
g. Sono certo (che) possa farcela
h. Ho stabilito *(che) venga assunto immediatamente
i. Mario crede ?(che) arriverà più tardi
l. Gianni immagina ?(che) Mario torni domani
m. Lui suppone ?(che) Marco arrivi domani
n. So *(che) Gianni è partito

Osserviamo, innanzitutto, come nelle frasi rilevanti in (2) il fenomeno della cancellazione del complementatore sia chiaramente opzionale e non sembrino darsi casi in italiano di cancellazione obbligatoria di questo elemento [7]. Per quello che concerne le frasi subordinate dipendenti da verbi possiamo affermare che i verbi che ammettono

CD debbano appartenere ad un determinato gruppo semantico di tipo epistemico-modale, emblematicamente rappresentato dalla classe dei verbi di credenza. Per il momento, possiamo limitarci a classificare questi verbi come modalizzanti, verbi che normalmente richiedono il congiuntivo nella subordinata. A questo gruppo appartengono verbi come: *ammettere*, *arguire*, *dubitare*, *giudicare*, *immaginare*, *pensare*, *presumere*, *ritenere*, *sospettare*, *supporre*, ecc. Tuttavia, se consideriamo il caso dell'inglese, constatiamo che la cancellazione del complementatore è sempre ammessa in contesti dipendenti da verbi di qualsiasi tipo semantico, e inoltre si deve anche rilevare che a differenza dell'italiano l'inglese possiede sostanzialmente solo una forma flessionale specifica per il modo congiuntivo che prevede la perdita della desinenza *-s* alla terza persona singolare. La CD è esclusa ogniqualvolta la frase subordinata è dipendente da un nome.

(3) a. I think (that) John left
b. I decided (that) Mary came the following day
c. The belief *(that) John will come...
d. *(That) John will arrive is possible

La cancellazione del complementatore è possibile nelle frasi dichiarative dalle quali un elemento sia stato estratto come in (4a), obbligatoria con l'estrazione del soggetto come nelle interrogative esemplificate da (4b), o nelle relative restrittive sul soggetto formate a partire da una struttura subordinata come ($4e_{CP2}$) – si tratta del ben noto effetto '*that*-trace' dove un soggetto non può essere estratto attraverso un complementatore realizzato – ammessa nelle relative restrittive sull'oggetto come (4c)

(4) a. Who do you think (that) John will invite t?
b. Who do you say (*that) t likes Mary?
c. This is the man (that) I saw t
d. This is the man *(that) t knows Mary
e. This is the man [$_{CP1}$ (that) [I think [$_{CP2}$ (*that) [t likes Mary]]]]

Si potrebbe provare a generalizzare immediatamente e sostenere che ogniqualvolta il *that* non precede una traccia come in (4a), (4c), e ($4e_{CP1}$) la cancellazione debba essere un fenomeno opzionale. Ma questa generalizzazione verrebbe subito falsificata da strutture come (3c) e (3d) dove il *that* non precede una traccia e non è cancellabile. Queste due strutture ci mostrano che solamente una testa verbale è in grado di selezionare un complementatore che possieda le proprietà pertinenti alla cancellazione di *that*. Come spiegare il fatto che nelle relative come (4c) la CD è ammessa? La risposta più plausibile è che in questo tipo di strutture il complementatore è prodotto da un movimento-wh della testa della relativa a partire dall'interno del sintagma verbale. In questo modo non sussisterebbe dipendenza della relativa da una testa non verbale [8]. Per quanto concerne la relativa sul soggetto in (4d), in passato è stata avanzata l'ipotesi che la cancellazione di *that* debba essere esclusa per ragioni di processing. Sembrerebbe che i parlanti, in assenza del complementatore, non siano in grado di segmentare correttamente la frase e si trovino costretti a processarla (categorizzarla) ambiguamente come una frase indipendente dichiarativa e come una dipendente relativa. Tuttavia, questa ipotesi è resa implausibile dal fatto che in certi stadi precedenti dell'inglese, e forse in certe varietà attuali, come mi fa osservare un referee, la cancellazione è ammessa. L'analisi più plausibile, come vedremo più avanti, sembra essere quella grammaticale in termini di assenza di governo della traccia in posizione soggetto da parte di una testa appropriata [9].

Se iniziamo ad estendere la base dei nostri dati possiamo constatare come nelle frasi interrogative dirette in italiano, diversamente dall'inglese, la cancellazione di *che* sia del tutto opzionale e facoltativa. Tuttavia, si può osservare che la presenza consecutiva di due complementatori – come nel caso delle interrogative formate da strutture subordinate – determina un effetto di ridondanza, rilevabile anche nelle relative costruite sempre a partire da una subordinata con il modo congiuntivo, effetto che sembra peggiorare le frasi che mantengono il complementatore

(5) a. Chi credi (che) t abbia aiutato Maria?
b. Le persone che credo (che) t cerchino di fare il loro dovere non sono poche
c. Questa è la ragazza che credo (che) Gianni abbia conosciuto t ieri
d. Gianni, che/il quale penso (che) abbia superato l'esame, è partito ieri

Mentre una frase come (2a), nella variante senza *che,* può apparire non tanto più corretta grammaticalmente quanto stilisticamente più elevata in rapporto alla variante che ammette il *che*, le frasi in (5), nella variante senza il *che*, sembrano invece mostrare un grado di preferibilità più marcato ed evidente, determinato proprio dall'assenza di ridondanza a cui si è prima accennato. Ad una più attenta osservazione, si nota che in (5a) – un caso di estrazione del soggetto – quando *che* non viene omesso, il SN in posizione postverbale viene facilmente interpretato come un soggetto, una possibilità meno frequente qualora il complementatore venga cancellato. La relativa appositiva in (5d) appare ancora più gravemente compromessa dalla presenza del complementatore. Infine, se utilizziamo un verbo come *scoprire*, che non seleziona il congiuntivo nel complemento frasale, l'effetto di ridondanza scompare e la CD non è ammessa nemmeno nella relativa formata a partire dalla struttura subordinata

(6) Le persone che ho scoperto *(che) sono fuggite non erano italiane

La cancellazione del complementatore può verificarsi anche nel caso di comandi espressi al congiuntivo

(7) a. (Che) Gianni entri immediatamente!
b. (Che) chiamino un medico! [10]

e nelle frasi iussive di tipo augurativo

(8) a. (Che) la fortuna sia con voi!
b. (Che) ti venga un accidente!

Nelle frasi ottative, analizzabili come delle strutture subordinate esprimenti un desiderio ma senza che compaia il verbo principale, si osserva che il complementatore non è mai ammesso, mentre il soggetto deve sempre essere postverbale

(9) a. (*Che) Venisse Gianni ad aiutarci!
b. *Gianni venisse ad aiutarci!
c. Arrivasse in orario almeno una volta!

Considerato che semanticamente un'ottativa come (9a) corrisponde ad una frase dipendente da una principale come (10a), dove il verbo che regge la subordinata non è mai presente, e tenuto conto anche del fatto che non può trattarsi di particolari forme di dislocazione a sinistra della subordinata in quanto *che* sarebbe obbligatorio, come è mostrato in (10b) e nella frase (iii) riportata alla nota 8, diventa plausibile avanzare l'ipotesi che queste strutture senza complementatore corrispondano a delle strutture troncate prive della proiezione di accordo del soggetto. Il fatto che il soggetto debba sempre comparire in posizione postverbale può indurci a pensare che l'IP, o una parte di esso, sia assente come conseguenza della mancanza dello strato strutturale Comp. Questa ipotesi, che spiegherebbe allo stesso tempo l'omissione sistematica del complementatore *che* e la posizione rigidamente postverbale del soggetto, dovrebbe comunque fare i conti con la complessa natura delle frasi ellittiche. Si noti, inoltre, come sia possibile costruire una ottativa preceduta da *che* a condizione che il soggetto sia preverbale come in (10c)

(10) a. [*Vorrei / Desiderei che*] venisse Gianni ad aiutarci
b. *(Che) venisse Gianni ad aiutarci, lo vorrei
c. (Che) Gianni venisse ad aiutarci almeno una volta [11].

In termini descrittivi e ricapitolativi, possiamo affermare che in italiano una molteplicità di fattori complementari determinano la cancellazione del complementatore. Specificamente, si è constatata la possibilità di prospettare almeno tre condizioni rilevanti per la CD, che possiamo ridurre a due di carattere generale: (i) il verbo della frase principale deve appartenere ad una tipologia epistemico-modale, (ii) è preferibile che la frase subordinata non possieda una forza grammaticale autonoma come nel caso del congiuntivo. Tuttavia, nessuna di queste due condizioni è strettamente vincolante. Anche per la prima, che a prima vista sembrerebbe non violabile, è possibile esibire dei casi non conformi, mentre la seconda è soltanto una condizione di preferibilità come attestano i dati esposti in (2) e quelli forniti alla nota 3. La distribuzione del fenomeno della CD non è quindi esattamente isomorfica alle proprietà di selezione semantica dei verbi di tipo modale, verbi che tipicamente non presuppongono la verità del contenuto proposizionale della subordinata, che non lo assumono come dato ma piuttosto come possibile. In termini generali, sembra più prudente sostenere che al momento attuale non siamo ancora in grado di stabilire se la classe degli elementi grammaticali e dei tratti sintattici che determinano l'emergenza di una articolazione dello strato Comp adeguata alla cancellazione, corrisponda in modo diretto a determinate proprietà semantiche e modali del verbo della principale. Tuttavia, questo non esclude che sulla base delle proprietà di selezione semantica non si possano fare, in molti casi, predizioni accurate. Le ragioni di questa indeterminazione nella formulazione di condizioni descrittive necessarie e sufficienti per spiegare il fenomeno della cancellazione è da attribuirsi al fatto che la selezione dello strato Comp pertinente, all'interfaccia fra la matrice e la dipendente, deve soddisfare molteplici condizioni ed esigenze di diverso valore funzionale. Si tratterà, specificamente, di individuare quelle proprietà sintattiche e quella particolare rappresentazione strutturale dello strato del complementatore più idonee a trattare in maniera unitaria questo fenomeno.

In inglese, l'opzionalità del fenomeno della CD è ancor meno direttamente riconducibile alla natura del verbo della principale, né tantomeno al tipo di flessione della subordinata. In questa lingua, non si osserva una specifica flessione al congiuntivo e pertanto la frase subordinata sembra sempre disporre di una forza grammaticale autonoma. Un quadro generale rappresentativo delle distribuzioni di forza nelle due lingue e nei contesti rilevanti potrebbe essere il seguente

```
(11)   a. Strutture con subordinazione: principale     Comp        subordinata
                                            |             |
                                        + forza       + forza (inglese)
                                                      +/- forza (italiano)

       b     Strutture relative:          testa relativa  Comp   frase relativa
                                                           |
                                                      +forza (italiano/inglese)
```

Per trattare le analogie e le differenze fra le due lingue sulla base di questo quadro delle distribuzioni di forza, verrà utilizzata la particolare articolazione X-barra del complementatore proposta da Rizzi (1997) associata ad una analisi in termini di verifica di tratti. L'articolazione dettagliata di questa ipotesi teorica, dove il fenomeno della CD viene spiegato sulla base delle differenti modalità in cui le proprietà di forza degli enunciati vengono soddisfatte, sarà preceduta dal § 2 dove si cercherà di illustrare l'ipotesi sviluppata da Poletto (1994), dal § 3 dove si espongono alcuni elementi ricavati dal testo di Giorgi & Pianesi (1997) e dove le soluzioni proposte dalla Poletto vengono riconsiderate alla luce del frame minimalista di Chomsky (1995), mentre nel § 4 si prenderanno in esame alcuni aspetti fondamentali della teoria di Rizzi (1997) sull'articolazione fine della struttura della periferia sinistra della frase. Quest'ultimo lavoro, insieme alle Class Lectures del corso di Sintassi Comparativa tenute da Rizzi a Ginevra negli a.a. 93/94-94/95, costituiscono la base teorica delle ipotesi avanzate nel § 5.

2. CD come movimento di V in Comp

Nella proposta teorica di Poletto il fenomeno della CD è analizzato nei termini del fenomeno del *V2* in contesti subordinati, dove la flessione muovendo in C° occupa la posizione del complementatore impedendogli di venire generato direttamente in quella posizione. Si assume generalmente che il fenomeno del 'verb-second', tipico delle lingue germaniche, sia provocato dalla presenza di tratti morfologici di accordo/Agr in C° (cfr. Tomaselli (1990).

L'italiano, nonché l'inglese, non sono lingue a 'verb-second', il problema nell'analisi di Poletto consisterà quindi nel determinare quali cause possano far scattare il movimento del verbo in Comp. La soluzione prospettata nell'articolo prende le mosse dal fatto che la classe speciale di verbi che permettono la CD assegna un tratto modale al complemento frasale e questo tratto +Mod viene realizzato sulla testa C°. Come spiegare il carattere opzionale del fenomeno della CD? Assumendo che il *che,* così come la flessione della subordinata, siano entrambi in grado di verificare il tratto modale in C°. Si stabilisce non solo che la flessione al congiuntivo possa realizzare i tratti modali in C° ma anche che il complementatore, normalmente generato in C°, possa soddisfare questa condizione. Sarà proprio la presenza di questi tratti modali a determinare il movimento del verbo, o, alternativamente, del *che* in Comp. In questo modo, si cerca di sviluppare un'analogia più generale fra l'italiano ed il fenomeno del V2: "the first piece of evidence for treating CD as a case of verb movement to C° is constitued by the parallel with embedded V2 in V2 languages like Standard German. The class of elements (verbs, adjectives or nouns) which permits CD in Italian is the same class which permits embedded V2 in German" (Poletto 1994: 5).

L'analisi della distribuzione di alcuni avverbi che modificano il complemento frasale induce Poletto ad ipotizzare l'esistenza di una proiezione ModP compresa fra AgrSP (che in Poletto è contrassegnata come AgrP) e CP

(12) a. Credo che sicuramente arrivi domani
b. *Credo sicuramente arrivi domani
c. Credo che arrivi sicuramente domani
d. Credo arrivi sicuramente domani.

La struttura frasale che si ottiene è la seguente: CP, ModP, AgrP, TP. Secondo Poletto, è l'accettabilità di una frase come (12c) a dimostrare la necessità di postulare un'ulteriore proiezione ModP. *Che* verrà quindi realizzato in C° mentre il verbo salirà in Mod° seguito dall'avverbio aggiunto ad AgrP. Tuttavia, l'ipotesi di spiegare il fenomeno della CD attraverso la presenza di V° in C° non sembra possedere pregnanza ed evidenza. Si può osservare, in prima istanza, come la flessione al congiuntivo salga in Agr°: è il caso di frasi come (12a), dove l'avverbio è aggiunto ad AgrP. In (12c), invece, il verbo salirà fino in Mod°, interpolandosi fra il complementatore e l'avverbio aggiunto ad AgrP, mentre in strutture con CD come (12d) il verbo dovrà salire fino in C°. L'interrogativo che deve essere posto è quindi il seguente: per quale ragione la flessione modalizzata in (12d) dovrebbe verificare i suoi tratti modali non nella proiezione modale ma in C°?

Se per spiegare una struttura come (12d) si assume che siano i tratti modali ad attirare il verbo in C°, determinando il fenomeno della CD, che cosa può attirare in (12c) il verbo in Mod° se non dei tratti di tipo Mod? Non solo, ma dovremo anche assumere che il soggetto possa, a seconda della posizione dell'avverbio, salire in Spec, AgrP, (13a), in Spec, ModP, (13b) e anche in Spec, CP, (13c), ponendo problemi di assegnazione di caso nominativo

(13) a. Credo che sicuramente Gianni arrivi
b. Credo che Gianni arrivi sicuramente
c. Credo Gianni arrivi sicuramente

Poletto presenta dei dati tratti dal dialetto piemontese di Torino che attesterebbero la possibilità di avere il soggetto in Spec,CP, ma non ne presenta per quello che concerne la possibilità di avere il soggetto in Spec,ModP

(14) a. Gnun ch'a s'bogia!
'Nessuno che si muova'

b. Mario ch'a s' presenta subit
'Mario che si presenti subito'.

Se si chiede ad un nativo di Torino competente del dialetto locale di illustrare le sue intuizioni su questi enunciati scopriamo che (14a) non rappresenta un ordine espresso nel modo imperativo con flessione al congiuntivo di terza persona: *che nessuno si muova*, ma è piuttosto un modo di dire, un'espressione idiomatica cristallizzata che sta a significare: *nessuno che faccia o cerchi mai di far qualcosa*. Nel caso contrario in cui si debba esprimere effettivamente un ordine, il quantificatore dovrà seguire sempre il complementatore. Se invece abbiamo un SN referenziale e il nostro scopo illocutorio è quello di impartire un ordine a qualcuno, la posizione del SN può precedere il complementatore o seguirlo. Si può anche osservare che in (14b), con il SN non quantificazionale che precede il complementatore, sebbene Poletto non sembri avvalorare questo fatto, la frase avrà una intonazione ascendente e un leggero stacco intonativo dopo il SN, esattamente come nelle costruzioni dislocate.

Non intendo addentrarmi ulteriormente nella questione anche per la delicatezza dei dati concernenti i dialetti che devono essere sempre presentati con molte precauzioni. Vorrei solo limitarmi ad aggiungere come nel mio dialetto nativo, quello di Piacenza, si verifichi esattamente la stessa situazione che abbiamo appena incontrato in (14a):

(15) a. Anson ch'a s'mova
b. Ch'anson s'mova

(15a) è solamente una sorta di esclamazione enfatica, riferibile ad una pigrizia generale dei presenti, mentre se intendiamo impartire un ordine dobbiamo sempre utilizzare (15b). La situazione appena menzionata in (15) si verifica comunque anche nell'italiano standard. Per il momento possiamo quindi presumere che i dati avanzati da Poletto non siano completi o perlomeno non forniscano evidenze sufficienti all'ipotesi che il soggetto possa trovarsi in Spec,CP. Vedremo più avanti, pur in un contesto teorico differente, come anche Giorgi & Pianesi (1997) arrivino a riconoscere che:

> "given that negative quantifiers cannot undergo topicalization, the (relative) acceptability of (i):
>
> (i) Mario crede nessuno arrivi stasera
>
> shows that the subject does not occupy a derived position. This is an important observation that any theory dealing with these facts should account for" (Giorgi & Pianesi 1995: cap. V).

L'assunzione che il soggetto si trovi in Spec,CP pone inoltre seri problemi di assegnazione di caso nominativo. Mentre nei contesti di tipo Aux-to-Comp come:

(16) a. Avesse Gianni aiutato Maria a cercare un lavoro
b. *Gianni avesse aiutato Maria a cercare un lavoro

si può ipotizzare che C° assegni caso nominativo sotto reggenza, in una struttura come:

(17) Credo Gianni abbia deciso di partire

dove secondo Poletto la flessione modale è in C° e il soggetto in Spec,CP, si deve assumere che C° assegni caso nominativo mediante Accordo Spec-Head, una ipotesi apertamente in conflitto con l'inaccettabilità di (16b). Nonostante queste evidenze sfavorevoli, Poletto sostiene che C° può assegnare caso anche attraverso Accordo Spec-Head e assume, diversamente dalla teoria sviluppata da Rizzi (1991) sul Criterio Wh, che in italiano Agr° possa assegnare caso sotto reggenza (under government). Come sappiamo, in Rizzi (1991) viene elaborato un criterio in grado di spiegare il complesso comportamento della flessione nelle strutture interrogative dirette con inversione, il Criterio Wh:

A. Un operatore Wh deve trovarsi in una relazione Spec-head con una testa +Wh

B. Una testa +Wh deve trovarsi in una relazione Spec-head con un operatore Wh.

Una conseguenza di questo principio è che la presenza della flessione verbale in C°, che ha il compito di veicolare i tratti +Wh, si verifica allo scopo di soddisfare il Criterio Wh, ma il movimento della flessione da I° a C° distrugge la corretta configurazione per l'assegnazione di caso nominativo. Questo fenomeno è attestato in frasi come la seguente

(18) *Cosa ha Gianni fatto?

La flessione, che da AgrS° si sposta in C°, non è in grado di assegnare caso nominativo sotto reggenza e per questa ragione il soggetto non può mai apparire fra l'ausiliare ed il verbo, nella posizione canonica in Spec,AgrSP, ma soltanto in posizione postverbale o dislocata a sinistra o a destra. Poletto, invece, sostiene che si possono rintracciare evidenze empiriche sufficienti per dimostrare che la flessione in Agr° può assegnare caso sotto reggenza. Le evidenze empiriche sono costituite dalle frasi seguenti che dovrebbero esprimere un registro ad un livello stilistico piuttosto elevato

(19) a. ?Cosa mai avrà Gianni fatto in quel frangente?
b. Cosa mai avrebbe Gianni potuto fare?

I diacritici riportati sono quelli assegnati da Poletto. Queste valutazioni non sono facilmente condivisibili: le frasi in (19) non sembrano essere chiaramente accettabili e resta ancora da dimostrare come sia possibile che C° assegni caso attraverso Accordo Spec-Head. Non dobbiamo dimenticare quanto sia cruciale questo passaggio se intendiamo sostenere che la CD sia un caso di movimento di V° in C°. La soluzione che viene prospettata prevede una seconda posizione Comp che si collocherebbe fra lo strato CP e AgrSP, come è stato anche proposto da Shlonsky (1994). Specificamente, si tratterebbe di una proiezione ibrida AgrC che riunirebbe in una sola posizione proprietà di tipo C e proprietà di tipo Agr, mentre il fenomeno della CD verrebbe configurato nei termini del movimento del verbo in AgrC°. Come rileva anche Bianchi (1999), il problema con questa ipotesi è che: "the deletion of the complementizer che cannot be directly derived from the movement of the verb to AgrC°, unless che is generated in AgrC° rather than in C°" (Bianchi 1999). La struttura massimale della frase sarà ora la seguente: CP, AgrCP, ModP, AgrSP, TP. Secondo Poletto, è possibile sostenere che: "the lower C can be realized as a complementizer, hence it must be a C position. Nevertheless it hosts a subject in its Spec and clitics on its heads, properties typically associated with an Agr head" (Poletto 1994: 23). Avremo che nei contesti Aux-to-Comp, participio assoluto e interrogativi, dove il soggetto può apparire alla destra dell'ausiliare, C assegnerà caso sotto reggenza mentre nei contesti che manifestano la cancellazione del complementatore, dove il soggetto può apparire alla sinistra del verbo come in (17), si dovrà assumere che C possa assegnare caso mediante Accordo Spec-Head. In quest'ultimo caso la testa coinvolta sarà AgrC° mentre in quelli precedenti sarà C°.

Tuttavia, nonostante queste revisioni, non emerge ancora chiaramente che cosa attiri il verbo in AgrC°: tratti +Mod, come all'inizio, o tratti +Agr? In ogni caso, dovranno essere dei tratti che possono venire soddisfatti indifferentemente sia dal verbo che da *che*, diversamente non si spiegherebbe la CD. È chiaro, comunque, che nell'analisi di Poletto il verbo può salire e terminare il suo movimento, a seconda dei casi, in Mod°, in contesti come (12c), in AgrC°, quando abbiamo CD, e in C° in contesti di tipo Aux-to-Comp.

3. CD come movimento del congiuntivo nella proiezione ibrida Mood/Agr

L'approccio di Giorgi & Pianesi (1997) rappresenta un tentativo di spiegare la distribuzione del congiuntivo e dell'indicativo nelle lingue romanze e germaniche su basi essenzialmente semantiche, utilizzando una strumentazione teorica mutuata dalla semantica dei mondi possibili e dagli sviluppi del frame minimalista di Chomsky (1995). È importante rilevare fin dall'inizio come in questo approccio si assegni un ruolo decisivo all'opposizione fra un contenuto proposizionale posto come 'possibile' ed un altro posto come 'dato'. Opposizione che in questo lavoro viene inizialmente riproposta nei termini della dicotomia tradizionale 'realis/irrealis' e reinterpretata come opposizione fra presupposizione *vs* assenza di presupposizione di verità del contenuto proposizionale espresso dalla subordinata:

> "according to such an account, the indicative is the mood of 'realis' contexts whereas the subjunctive is employed in 'irrealis' contexts. As we stated above, such a simple binary distinction will be shown to be too coarse grained in that it cannot account for crosslinguistic variation" (Giorgi & Pianesi 1997: cap. V).

Un'analisi del modo della frase subordinata, insieme alla particolare tipologia semantica del verbo della frase sovraordinata che seleziona il congiuntivo, conducono i due autori ad una classificazione delle proprietà dei differenti tipi di contesti che manifestano la CD. Sulla base di queste analisi e di alcuni principi formulati a partire da Chomsky (1995) le due versioni A e B del <Feature Scattering Principle> dove: "each feature can head a projection" si giunge a prospettare una morfosintassi del congiuntivo in termini di verifica di un tratto Mood che si realizza sopra una proiezione sincretica Mood/Agr. A questo punto l'analisi tocca più da vicino gli scopi del nostro discorso.

L'analisi della CD prende le mosse dal riconoscimento che questo fenomeno si manifesta tipicamente con il congiuntivo, ma si riconosce che questa non può essere ancora una condizione sufficiente. Come abbiamo visto anche nel § 1, la frase (2h) rappresenta dei contesti formati da verbi fattivi dove la presenza del congiuntivo non è un requisito sufficiente per la CD. Secondo gli autori, in contesti fattivi come nella frase seguente:

(20) Si rammarica *(che) sia partito

la cancellazione non sarebbe ammissibile poiché il contenuto proposizionale della subordinata è posto come dato [12], esattamente come nel caso in cui si richiede l'indicativo nella subordinata:

(21) Ha confessato *(che) è partito

Le differenze con l'approccio di Poletto (1994) concernono il fatto che il solo fattore in grado di determinare il movimento manifesto del verbo è la presenza di tratti forti +Agr insieme all'analisi della posizione assegnata al soggetto in frasi come:

(22) *Credeva Gianni fosse arrivato ieri

Secondo i due autori:

> "we observed that Italian speakers do not homogeneously share this judgment. For some speakers (22) is grammatical, while for others it is not. Moreover, this fact does not seem to be related to regional variation. Significantly, the two authors of this work do not have the same judgments, in spite of the fact that both originate from central Italy" (Giorgi & Pianesi 1997).

La testa che può ospitare rispettivamente la flessione al congiuntivo, nei casi di CD, e il complementatore *che*, non viene designata come AgrC ma apparterrà ad una categoria sincretica che proietta tratti modali e tratti di accordo, una proiezione Mood/Agr:

> "the subjunctive always moves to check the feature mood. Movement is overt when mood is projected syncretically with Agr, given that Agr feautures are strong. In these cases, therefore the checking of mood is parasitic on the checking of the phi-features. Movement is covert when mood is expressed by means of a complementizer, i. e. when the scattering B option has been taken, given that mood is weak" (Giorgi & Pianesi 1997).

Mentre in Poletto il soggetto nei contesti con CD appare in Spec,AgrC, secondo i due autori le differenze nei giudizi di accettabilità di (22) sarebbero invece determinate dal fatto che lo Spec della proiezione sincretica Mood/Agr è per alcuni parlanti una posizione A, accessibile al soggetto, mentre per altri una posizione A' limitata ad elementi di natura operatoriale.

Quello che costituisce un ulteriore significativo spostamento rispetto all'analisi di Poletto, è il tentativo di estendere questa ipotesi anche alle strutture interrogative con inversione. Se lo Spec della proiezione sincretica è una posizione di tipo A/A', allora è possibile che questa stessa posizione di Spec possa ospitare operatori Wh. Nei casi in cui la proiezione sincretica è utilizzata per delle strutture interrogative, avrà le caratteristiche di una proiezione di tipo Wh/Agr, dove il movimento manifesto della flessione nella testa di questa proiezione verrà determinato da tratti Agr forti, mentre il movimento dell'operatore Wh verrà determinato dai tratti Wh presenti sullo specificatore. Questa ipotesi della omogeneità fra le proiezioni Wh/Agr e Mood/Agr è basata su alcune analogie fra i contesti con CD e l'inversione nelle interrogative dove si osserva l'impossibilità di avere un soggetto fra l'ausiliare e il participio:

(23) a. *Chi ha Gianni visto?
b. *Credo sia Gianni arrivato

mentre nelle strutture Aux-to-Comp, dove l'ausiliare muoverebbe in una proiezione di tipo C ma non di tipo Wh/Agr, il soggetto può apparire a destra dell'ausiliare:

(24) Avendo Gianni incontrato Maria.

Ci sarebbero due posizioni Comp, ma solo una, quella sincretica Wh/Agr, sarebbe pertinente nelle strutture interrogative.

Non rientra negli obiettivi di questo lavoro approfondire una questione come l'inversione nelle frasi interrogative. Tuttavia, se si ritiene possibile che un'analisi delle interrogative possa venire sviluppata come un'immediata estensione dell'analisi del fenomeno della CD, allora è plausibile che le eventuali obiezioni mosse alle soluzioni proposte per la CD possano ripercuotersi sulla questione delle interrogative e viceversa. Abbiamo visto che secondo i due autori è cruciale il fatto che:

> "in CD cases the subject can never appear between the auxiliary and the participle, whereas this is possible in Aux-to-Comp constructions [...] with respect to this pattern interrogative constructions strongly resemble CD constructions [...] interrogative constructions instantiate the same kind of phenomena found in CD contexts: the subject cannot occur preverbally" (Giorgi & Pianesi 1997).

Ma le analogie non spiegano le differenze, come si può facilmente constatare dal fatto che nei contesti dove il *che* non viene cancellato e quindi il verbo non è salito in Comp nella testa Wh/Agr, il soggetto non può apparire a destra del participio:

(25) *Credo che sia Gianni partito ieri.

Sembra più plausibile pensare che in (23b) e (25) la posizione della flessione sia sempre la stessa e che nessun movimento di V in C sia intervenuto, mentre il soggetto, in entrambi i casi, semplicemente non abbia completato il suo movimento per raggiungere la posizione di assegnazione di caso in Spec,AgrSP determinando la malformazione delle due strutture. L'obiezione costituita dalla frase (25) mi sembra decisiva. Secondo i due autori, la devianza di (23a) e (23b) sarebbe determinata dal fatto che in (23a) è disponibile solo una posizione di Spec per il soggetto e l'operatore Wh, e questo spiegherebbe il fatto che il soggetto non può restare né fra l'ausiliare ed il verbo né salire in Spec,Wh/Agr. Quello che non appare chiaro è da che cosa esattamente sia determinata la devianza, se dal fatto che il soggetto non può salire nella posizione di Spec della testa Wh/Agr, perché è già occupato dall'operatore Wh, o dal fatto che nella posizione a destra dell'ausiliare non può ricevere caso nominativo poiché la flessione non assegna caso sotto reggenza. L'analogia fra contesti con CD ed interrogative sarebbe pertanto motivata dal fatto che in (23b) il soggetto deve salire nello Spec di Mood/Agr per ricevere caso nominativo, poichè la flessione sarebbe salita in quella testa, ma la frase (25) sembra però indicare fortemente che il soggetto in (23b) non è in Spec,AgrSP, così come la flessione congiuntiva non è in Mood/Agr°. Entrambi si troverebbero più in basso e nessuna assegnazione di caso sarebbe in gioco nella proiezione sincretica Mood/Agr o Wh/Agr. Secondo Giorgi & Pianesi la devianza di (26) si spiega col fatto che:

(26) *Che cosa Gianni ha fatto?

> "our hypothesis on interrogative inversion is that, similarly to what we proposed for CD constructions, the subject cannot appear preverbally because Agreement and C form a syncretic category and as such there is only one Spec position available. Since the Wh-operator in Italian overtly moves there, the lexical subject has to appear elsewhere, namely inside the VP, whereas in CD structures pro can stay preverbally" Giorgi & Pianesi (1997).

Infine, una questione concernente la valutazione di alcuni dati empirici. Nel § 2 del presente lavoro viene riportato un passo tratto da Giorgi & Pianesi dove si evidenzia come la posizione del soggetto ben difficilmente può essere analizzata come una posizione derivata. Ritengo vi sia un'incongruenza fra il sostenere, da un lato, che la flessione al congiuntivo muova in C e dall'altro negare al soggetto una posizione derivata. C'è una incongruenza perché se ammettiamo, come viene fatto, che una struttura come (22) è almeno accettabile all'interno di una certa area di parlanti e la flessione è salita in

Mood/Agr°, il soggetto deve essere necessariamente in una posizione derivata, diversa da quella canonica in Spec,AgrSP, a meno che si intenda sostenere che (22) è un enunciato inaccettabile *tout court*.

In conclusione, se si perviene ad ammettere che nella proiezione Mood/Agr la posizione di specificatore ben difficilmente può accogliere un soggetto realizzato, questo significa che molto plausibilmente non si tratta di una proiezione di accordo, ma a questo punto il verbo non potrebbe più muoversi in questa testa ibrida per verificare i suoi tratti forti +Agr. Viene meno così anche il tentativo di spiegare la posizione adiacente all'operatore Wh assunta dal verbo nell'inversione nelle interrogative. Non trattandosi di una proiezione di tipo Agr, il verbo non raggiungerebbe la posizione adiacente all'operatore Wh per verificare dei tratti Agr, ma per altre ragioni.

4. Il sistema Forza-Finitezza nella periferia sinistra della frase

Nell'articolo *The fine structure of the left periphery* (1997), Rizzi ha sviluppato una struttura particolarmente articolata del sistema del complementatore. Secondo le linee fondamentali di questa analisi, sotto l'impatto del lavoro di J.Y. Pollock (1989) sulla struttura del sistema della flessione, lo strato Comp si è venuto sempre più arricchendo di diverse proiezioni che articolerebbero a livello sintattico anche proprietà di tipo interpretativo e illocutivo. La peculiarità fondamentale dello strato Comp è quella di trovarsi all'interfaccia, di qualificarsi come *trait d'union*, fra il contenuto proposizionale espresso da una frase subordinata e la frase principale sovraordinata o direttamente con il contesto di discorso linguistico-extralinguistico: "we expect the C system to express at least two kinds of information, one facing the outside and the other facing the inside" (Rizzi 1997: 3).

Come sappiamo, ogni enunciato rientra all'interno di un certo tipo frasale e l'area del complementatore sembra disporre della capacità di rappresentare quelle proprietà che rendono una frase una interrogativa piuttosto che una esclamativa, una iussiva o una ottativa. Queste proprietà rappresentano la forza grammaticale della frase in quanto la costituiscono come una frase autonoma e rientrante in una precisa tipologia frasale. Se consideriamo le strutture subordinate dove si verifica cancellazione osserviamo che è la frase principale a determinare il tipo frasale e a possedere la forza grammaticale. Sebbene i contesti di credenza della forma *credo che…* siano frasi di tipo dichiarativo non possiamo sostenere che in simili contesti modalizzanti il contenuto proposizionale subordinato venga asserito nello stesso modo in cui viene asserito in una frase semplice o in una subordinata come in (27):

(27) a. Ho saputo che Gianni vive qui
b. Gianni vive qui.

In questi casi di divieto di cancellazione la frase subordinata è indipendente dalla sovraordinata e dispone di una sua forza autonoma. Ritroviamo, in questo modo, nelle strutture subordinate, due sistemi formalmente distinti ma anche strettamente connessi. Se è l'area del complementatore ad avere il compito di rappresentare le proprietà di forza della frase e di connetterle con l'"interno' e con l'"esterno', nelle strutture con subordinazione avremo due sistemi: uno all'interfaccia con il contesto di discorso linguistico-extralinguistico in cui è collocato il parlante, ed un'altro con il compito di connettere il sistema della frase principale con quello della subordinata. Il primo sistema, quello connesso con l'esterno, responsabile direttamente della forza grammaticale della frase principale, potrà venire realizzato anche tramite particolari operatori non realizzati a livello morfologico, oppure attraverso indicatori di forza di tipo prosodico come l'intonazione:

(28) a. Esci?
b. Esci!

In entrambi i casi il complementatore dovrà essere in grado di esprimere sincreticamente informazioni provenienti dai due settori al fine di rendere possibile una connessione coerente. Se prendiamo il caso specifico delle subordinate al congiuntivo, l'interfaccia a contatto con la principale dovrà esprimere le proprietà di forza della principale mentre nella zona a contatto diretto con la subordinata dovrà rappresentare proprietà di tipo modale-temporale. Volendo usare una metafora, possiamo dire che il sistema del complementatore si comporta come un mediatore che per condurre felicemente in porto un affare fra due contraenti deve essere capace di far valere e soddisfare le esigenze di entrambi.

Ritornando all'analisi di Rizzi, si rileva come:

> "it appears that, at least in these language families, C expresses a distinction related to tense but more rudimentary than tense and other inflectional specifications on the verbal system: finiteness. Following much recent work (e.g., Holmberg & Platzack (1998)), I

> will then assume that the C system expresses a specification of finiteness, which in turn selects an IP system with the familiar characteristics of finiteness: mood distinctions, subject agreement licensing nominative case, overt tense distinctions" (Rizzi 1997: 3).

In questa analisi, si assume che, all'interno del sistema del complementatore, si trovino due proiezioni che rappresentano le proprietà di forza e finitezza della frase: quello che chiameremo il sistema di forza-finitezza. Il secondo sistema fondamentale del Comp sarà il sistema topic-focus. L'interazione di questi due sistemi determina, secondo Rizzi (1997), una struttura massimale del complementatore di questo tipo:

ForzaP, TopP, FocusP, TopP, FinitezzaP, IP.

Non esamineremo le proprietà del secondo sistema fondamentale del Comp. Sarà sufficiente notare che mentre il sistema di forza-finitezza rappresenta le relazioni selettive fra la parte alta della frase e l'IP il sistema topic-focus è indipendente da restrizioni di tipo selettivo. Mentre il sistema di forza-finitezza viene sempre attivato ogniqualvolta il VP principale seleziona una subordinata, il sistema topic-focus, al contrario, verrà attivato soltanto nel caso di clitic left dislocation, hanging topic*,* topicalizzazione o focalizzazione di un costituente.

Per quanto concerne le strutture che ci interessano più da vicino, l'articolazione strutturale del sistema di forza-finitezza ci pone nella condizione di poter rendere conto non solo delle intrinseche proprietà dei contesti con subordinazione ma anche di spiegare agevolmente le caratteristiche distribuzionali dei diversi tipi di complementatori e le loro interazioni con altri elementi che vengono a trovarsi nella parte sinistra della frase. Nelle frasi seguenti vengono presentate alcune delle proprietà distribuzionali manifestate dai complementatori *che* e *di* in rapporto alla possibile occorrenza topicalizzata di un avverbio o alla left dislocation di un XP:

(29) a. Credo [ForP *(che) [TopP sicuramente [FinP[IP Gianni arrivi domani]]]]
b. *Penso [FinP di [TopP probabilmente [IP andare al cinema]]]
c. *Credo [TopP questo libro [FinP che [IP lo leggerò domani]]]
d. Credo [ForP *(che) [TopP questo libro[FinP[IP lo leggerò domani]]]] [13].

Nelle frasi (29a) e (29c) la posizione più plausibile per il complementatore *che* è la testa della proiezione di forza. Il *che* può essere seguito da un avverbio mentre la proiezione di finitezza – proprio per le proprietà che deve esprimere e per la funzione che ricopre in rapporto all'IP – deve sempre trovarsi adiacente a quest'ultimo e l'interpolazione di un avverbio o di un sintagma dislocato è da escludersi. Le proprietà distribuzionali di *che* e *di* giustificano l'ipotesi che debbano essere disponibili almeno due proiezioni Comp differenti:

> "in Italian, and more generally in Romance, prepositional elements introducing infinitives such as di are generally considered the non-finite counterparts of the finite complementizer *che* [...] this distribution is hardly consistent with a theory assuming a unique C position , while it can be immediately expressed within the current articulated theory of C by assuming that *che* manifests the force position, while di manifests the finiteness position, hence they show up on opposite sides of the topic" (Rizzi 1997: 7).

Abbiamo una prima posizione Comp, quella di forza, che deve sempre occupare la parte più alta a contatto con la frase principale e una seconda posizione Comp che deve rispettare restrizioni di adiacenza con il sistema flessionale IP. All'interno di questi due poli Comp, si può articolare, a seconda dei casi e a partire dalla natura ricorsiva delle proiezioni topicali, il sistema topic-focus.

La domanda alla quale dobbiamo cercare di rispondere è la seguente: che cosa accade in tutti quei casi in cui il sistema topic-focus non viene attivato e le proiezioni di forza e finitezza vengono a trovarsi adiacenti? La soluzione avanzata da Rizzi è la seguente:

> "the proposed analysis assumes that force and finiteness can be expressed in a single head, and that this option is enforced by economy unless the activation of the topic-focus field makes it not viable. Alternatively, one could consider the possibility that the force-finiteness is 'agglutinative' as many other syntactic subsystems seem to be, hence it always involves two distinct heads" (Rizzi 1997: 47).

Abbiamo già visto, in contesti teorici diversi, che l'ipotesi di una proiezione Comp sincretica ha dei precedenti nella letteratura sull'argomento. A ragioni di natura 'economica' è possibile affiancarne altre che concernono più direttamente la natura del legame che il sistema del complementatore stabilisce fra la parte alta e la parte bassa della frase, sembra inoltre che il fenomeno della CD possa costituire un evidenza importante in favore dell'ipotesi sincretica.

Prima di ritornare al fenomeno della CD vorrei fornire qualche elemento dell'analisi di Rizzi sulla questione degli effetti di antiadiacenza. Come sappiamo dalle indagini condotte sull'inglese, affinché una traccia lasciata in posizione soggetto sia legittimata è indispensabile che certi requisiti di località vengano soddisfatti, nella fattispecie che una testa con capacità di governo, una testa con tratti +Agr, si trovi adiacente alla traccia soggetto mantenendola in una posizione sotto reggenza. Le modalità di soddisfazione di queste condizioni sono formalmente definite da ECP (Empty Category Principle). Un tipico contesto che viola questo principio è l'effetto '*that*-trace' poiché una testa C in cui non siano stati attivati tratti Agr, una testa realizzata come *that*, non è in grado di legittimare una traccia in posizione soggetto. Questo fenomeno si osserva nel caso dell'estrazione del soggetto in una interrogativa, dove il passaggio del soggetto nello specificatore della proiezione Comp adiacente alla traccia attiva i tratti Agr su questa testa, ne impedisce la realizzazione come *that* e la mette nella condizione di governare propriamente la traccia soggetto:

(30) Who$_i$ do you think [CP t'$_i$ (*that) [IP t$_i$ left ?]].

Questo fenomeno ci interesserà più da vicino quando cercheremo di fornire un'ipotesi alternativa della cancellazione del complementatore *that*. Vediamo, ora, quello che succede nel caso degli effetti di antiadiacenza. La caratteristica di queste strutture, tipicamente delle relative sul soggetto formate a partire da una frase subordinata, è che un avverbio si interpone fra la traccia e il *that* determinando un contesto più favorevole alla traccia:

(31) a. An amendement which I think (*that) t will be law next year
b. An amendement which I think *(that) next year t will be law.

L'articolazione del sistema di forza-finitezza fornisce una importante soluzione a questo complesso rompicapo. Perché, nella frase (31b), sulla traccia in posizione soggetto, ECP viene soddisfatto mentre questo non può accadere in (31a)? L'ipotesi è che l'avverbio topicalizzato attivi il sistema topic-focus determinando la dissociazione delle due teste di forza e finitezza. In questo modo, la proiezione di finitezza adiacente all'IP mette a disposizione una testa in grado di legittimare la traccia in posizione soggetto:

(32) An amendement which I think[ForP that [TopP next year[FinP t' fin°+Agr[IP t will be law]]]].

Questa analisi del fenomeno dell'antiadiacenza fornisce delle prove in favore dell'ipotesi che le due specificazioni di forza-finitezza, qualora il campo topic-focus non venga attivato, debbano trovarsi riunite in una sola proiezione sincretica. Si consideri la rappresentazione strutturale di (31a) dove le specificazioni di forza-finitezza vengono rappresentate prima sopra due teste distinte e poi sopra una proiezione sincretica:

(33) a. An amendement which I think [ForP (*that) [FinP t' fin°+Agr [IP t will be law]]]

b. An amendement which I think [For-FinP (*that) [IP t will be law]].

Se (33a) fosse la rappresentazione corretta della struttura del complementatore di (31a) con il *that* generato nella testa di forza, allora non si spiegherebbero i positivi effetti di antiadiacenza provocati dalla presenza dell'avverbio in (31b), e specificamente dalla presenza della testa di finitezza che sarebbe in grado di governare la traccia in posizione soggetto in (31b) ma che misteriosamente perderebbe questa capacità in (33a). Se, al contrario, assumiamo l'ipotesi sincretica espressa in (33b), dove il *that* è nella testa sincretica, siamo in grado di spiegare i fenomeni dell'antiadiacenza e la malformazione dell'opzione con il *that* come conseguenza di un tipico effetto '*that*-trace'. Nei contesti in cui il sistema topic-focus non viene attivato la generazione del complementatore non può verificarsi che in una testa che rappresenti sincreticamente le due specificazioni di forza e finitezza. L'analisi condotta da Rizzi sugli effetti di antiadiacenza costituisce una forte evidenza in favore dell'esistenza di una proiezione sincretica nei contesti che manifestano un effetto '*that*-trace', ed è sulla base di questa ipotesi che deve essere affrontato il fenomeno della CD. Questi due fenomeni sembrano essere strettamente connessi e la cancellazione opzionale di *that,* sempre disponibile nei contesti con subordinazione in inglese, non potrà prescindere né da una analisi delle proprietà e dei tratti presenti sulla proiezione sincretica né da una adeguata rappresentazione dei tratti specifici che caratterizzano intrinsecamente questo tipo di complementatore. Rizzi individua i seguenti tratti specifici che possono essere posseduti da *that*:

> "suppose that the force-finiteness system can be expressed by a single item drawn from the functional lexicon. In English, for embedded finite declaratives we have the alternation that/0. I will continue to assume that the latter, but not the former, is consistent with Agr:

(i) That = +decl, +fin
0 = +decl, +fin, (+Agr).

Suppose now that the topic-focus field is activated in the C system. The force specification must be manifested by that above the topic, on the other hand, finiteness must be manifested by a zero C head (fin) under the topic. So, we should revise (i) in the following way:

(ii) That = +decl, (+fin)
0 = (+decl), +fin, (+Agr)" (Rizzi 1997: 27-28).

Nell'analisi di Rizzi, *that* non possiede intrinsecamente dei tratti specifici ma li assume a seconda dei contesti in cui occorre, che si trovi proiettato in una testa sincretica o in quella di forza. Ad esempio, nel caso dell'estrazione del soggetto in una interrogativa come in (30), che è il solo caso di CD obbligatoria, la forma zero del Comp disporrebbe di tratti Decl, Fin e Agr. I tratti Agr sulla forma zero sarebbero opzionali poiché sono attivati dal passaggio del soggetto nello Spec della testa sincretica, mentre in altri casi di CD, dove non si verifica il passaggio del soggetto, la forma zero non avrebbe tratti Agr ma solo tratti Decl e Fin. Sembra esserci quindi uno stretto rapporto fra la CD obbligatoria e l'attivazione di tratti Agr. In (34) abbiamo un'altro caso in cui si evidenzia chiaramente come l'opzione della cancellazione del *that* è ammessa solo quando il *that* viene realizzato nella testa sincretica:

(34) I think *(that) next year John will leave the country.

L'attivazione del campo topic-focus provoca la separazione delle due specificazioni di forza e finitezza e il *that* viene a trovarsi nella testa più alta di forza impedendone la cancellazione. Sulla base di questi dati si può avanzare l'ipotesi che la CD sia connessa alla presenza sincretica e contemporanea delle specificazioni di forza e finitezza e che solo grazie alla presenza di quest'ultima sia possibile attivare dei tratti Agr all'interno di una testa Comp. Secondo Rizzi, la testa di finitezza ha il compito di esprimere distinzioni di modo, di tempo, distinzioni legate, nella maggior parte dei casi, a fenomeni di accordo fra la flessione e il soggetto e di assegnazione di caso nominativo. Si può osservare, inoltre, come in inglese *that* occorra sempre in contesti contrassegnati da queste proprietà, mentre in presenza di forme non temporalizzate venga utilizzato il complementatore *for* che occuperebbe la proiezione più bassa del sistema del complementatore, quella di finitezza, come viene mostrato nell'esempio seguente dove un avverbio topicalizzato non può mai intervenire fra *for* e l'IP:

(35) a. ...for, (*tomorrow), John to leave
b. ...that, (tomorrow), John will leave.

Questa analisi sembra essere perfettamente in grado di trattare questi fenomeni dell'inglese e può venire estesa anche all'italiano. Sarà sufficiente introdurre alcune estensioni che tengano conto delle particolarità dell'italiano.

L'ipotesi che stiamo sviluppando assume che il fenomeno della CD è intrinsecamente connesso alla presenza della testa sincretica di forza e finitezza. Abbiamo visto che questo fenomeno, nell'inglese, pur rimanendo nella maggioranza dei casi opzionale, è sempre possibile nei contesti subordinati mentre in italiano dipenderebbe dalla presenza di un certo gruppo di verbi che prevalentemente richiedono il congiuntivo nella subordinata. L'ipotesi è che in inglese la variante sincretica venga assegnata per 'default' nei contesti con subordinazione mentre in italiano sia sottoposta a delle restrizioni selettive determinate dal tipo di verbo che appare nella principale. Se la CD può essere utilizzata come un test per accertare la presenza della proiezione sincretica, allora in italiano, in tutti quei contesti in cui la subordinata non può manifestare il modo congiuntivo, come ad esempio nella frase seguente:

(36) Ha detto *(che) è / *sia partito ieri

possiamo pensare che questo tipo di proiezione sia assente e che il complementatore *che* sia generato nella testa di forza. Ma che ne è della specificazione di finitezza? Nel cercare di rispondere a questa domanda dobbiamo partire dalla constatazione che in italiano le frasi dichiarative possono manifestare solamente il modo indicativo o il condizionale mentre il congiuntivo, non avendo forza dichiarativa, è limitato alla subordinata. Una delle particolarità delle frasi come (36) è che la subordinata possiede in modo autonomo una sua forza, mentre quella al congiuntivo non può presentarsi isolata:

(37) a. È partito ieri
b. *Sia partito ieri.

L'ipotesi è la seguente. Mentre in inglese certe distinzioni di modo sono presenti in modo limitato, in italiano si manifestano e devono corrispondere differentemente alle specificazioni di forza e finitezza, quest'ultima, in particolare, esprimerà più direttamente le proprietà dei contesti modalizzati manifestati dal modo congiuntivo mentre la testa di forza sarà direttamente connessa a modi assertivi come l'indicativo. Qualora la frase subordinata sia al modo indicativo, come in (36), è plausibile che la specificazione di finitezza non venga rappresentata in quanto non sarebbero in gioco proprietà di tipo modale. La connessione fra la parte alta della struttura e la parte bassa sarà quindi assicurata dalla sola proiezione di forza che esprimerà la proprietà rilevante di ambedue le strutture. Torneremo più avanti sopra questo aspetto delle strutture che non ammettono CD. La rappresentazione strutturale di (36) è:

(38) Ha detto [ForP *(che) [IP è partito ieri]].

Pertanto, mentre in inglese la proiezione di finitezza si fa carico di rappresentare le proprietà di tempo e modali disponibili in quella lingua, in italiano avrà il compito di rappresentare le proprietà modali espresse dal congiuntivo, mentre le proprietà di forza degli altri modi, come ad esempio l'indicativo, verranno rappresentate specificamente dalla proiezione di forza. Le differenti strategie utilizzate dall'inglese e dall'italiano nel rappresentare sulle proiezioni di forza e finitezza le proprietà modali, temporali e di forza della subordinata, possono venire schematizzate nel modo seguente:

(39) inglese: FinP = (a) forme finite di differenti modi verbali che esibiscono anche proprietà modali
(b) forme non finite

italiano: FinP = (a) forme finite al congiuntivo e altri modi che manifestano proprietà modali
(b) forme non finite

ForP = (a) forme finite di differenti modi verbali che manifestano proprietà di forza.

5. CD come incorporazione della proiezione di Forza-Finitezza nella proiezione Agr associata

Come dovrebbe essere già emerso da alcune indicazioni sviluppate nel precedente paragrafo a proposito dello stretto rapporto fra la CD e la necessità della presenza di una proiezione sincretica, nella soluzione che si sta sviluppando per il fenomeno della CD il movimento del verbo in una proiezione di tipo Comp non può giocare alcun ruolo esplicativo. La scarsa plausibilità dell'ipotesi che il soggetto possa occupare lo specificatore di una proiezione Comp, e quindi di essere sottoposto ad un ulteriore movimento dalla posizione canonica nello Spec,AgrSP, oppure l'idea che il complementatore venga generato all'interno di una proiezione di accordo mentre soltanto la flessione congiuntiva muoverebbe nella posizione ibrida AgrC, insieme alle difficoltà di applicazione di questo modello all'inglese dove la CD è una opzione sempre possibile, spingono ad adottare le linee fondamentali dell'ipotesi più conservativa e al tempo stesso più consistente con i dati empirici sviluppata da Rizzi.

L'ipotesi esplicativa del fenomeno della CD avanzata in questo paragrafo è in termini di incorporazione del complementatore, ma non secondo le modalità illustrate da Pesetsky (1995) [14]. La soluzione che si intende privilegiare è fondata su alcune osservazioni sviluppate da Rizzi circa la plausibilità di ipotizzare l'esistenza di una proiezione pura Agr associata a tutte quelle teste che possono disporre di tratti Agr, vale a dire di un contenuto semantico-modale non puramente funzionale. L'associazione con Agr, secondo Rizzi, deve essere limitata alle teste del sistema temporale-aspettuale che visibilmente ammettono un accordo morfologico realizzato nell'IP. Nel sistema Comp, quindi, l'opzione esiste per FinP e per la testa sincretica che include quella di finitezza ma non per la pura testa di forza. I tratti Agr presenti su questo tipo di teste verrebbero verificati, nelle condizioni appropriate, all'interno delle proiezioni Agr ad esse direttamente associate. È plausibile che la proiezione sincretica di forza-finitezza che denomineremo For-FinP, possa venire associata ad una proiezione pura Agr al cui interno possa verificare i suoi tratti Agr. La specificazione di finitezza con i suoi contenuti modali/temporali e la sua capacità di legittimare una traccia nei contesti appropriati sembra disporre di elementi sufficienti per pensare che sia intrinsecamente costituita con tratti Agr, rendendola un candidato plausibile per la classe di teste che possono venire associate ad una proiezione pura di accordo. La natura chiaramente operatoriale/A' della specificazione di forza esclude, invece, che questa proiezione possa venire associata ad una proiezione Agr.

Se il fenomeno della CD corrisponde effettivamente alla distribuzione della proiezione sincretica, diventa plausibile concludere che la cancellazione del complementatore consista nell'incorporazione della testa sincretica nella proiezione Agr associata per verificare i suoi tratti di accordo. Come abbiamo visto, il fenomeno della CD è in diversi casi intrinsecamente opzionale/facoltativo, possiamo quindi pensare che una testa dotata di tratti Agr non costituisca di per se stessa una condizione sufficiente perché i tratti di accordo debbano venire obbligatoriamente verificati. Possiamo stabilire la condizione seguente:

(41) *Condizione minimale d'incorporazione del Comp*
CD opzionale = verifica dei tratti +Agr di For-FinP mediante l'incorporazione di For-FinP in Agr,For-FinP.

Oppure, possiamo ipotizzare che i tratti siano comunque verificati ma secondo due strategie diverse a seconda delle preferenze del parlante. Nella prima, utilizzando il modello in (41) dell'incorporazione della testa sincretica nella proiezione di accordo associata, nella seconda, direttamente mediante la realizzazione di un complementatore nella testa sincretica. Questa seconda ipotesi sarebbe anche consistente con la situazione che si verifica in (51), dove *that* verifica i suoi tratti venendo generato nella proiezione FinP. Nel caso della estrazione del soggetto nelle interrogative:

(42) Who do you think (*that) t left?

il fenomeno della CD obbligatoria è attestato esattamente in corrispondenza del verificarsi della relazione di Accordo Spec-Head fra la testa sincretica e la traccia del soggetto che veicola tratti Agr e che nell'analisi di Rizzi (1990) attiva tratti Agr nella testa Comp. Questo tipo di relazione di Accordo Spec-Head si verifica solo in questo tipo di contesti mentre negli altri che abbiamo già considerato non si verifica nessuna relazione di questo genere e la CD è solo opzionale, come nella frase seguente:

(43) I think (that) John left.

All'interno della nostra ipotesi sulla CD, la presenza di una relazione di Accordo Spec-Head nella proiezione For-FinP come in (42), determina l'incorporazione obbligatoria della testa nella proiezione Agr associata:

(44) *Condizione massimale d'incorporazione del Comp*
CD obbligatoria = relazione di Accordo Spec-Head fra i tratti presenti nella proiezione sincretica For-FinP che ne determina l'incorporazione nella proiezione Agr associata.

Si tratta, ora, di articolare più dettagliatamente il meccanismo che innesca l'incorporazione della testa sincretica nella proiezione Agr associata. A questo scopo sono indispensabili alcuni chiarimenti preliminari sulla linea di analisi e le assunzioni che s'intendono sostenere. Nell'analisi di Rizzi (1997), quando è la testa sincretica ad essere selezionata, *that* viene a disporre di tratti dichiarativi e di finitezza. Sembra ragionevole definire i tratti dichiarativi come tratti di forza (For) con un valore di tipo operatoriale/A', mentre i tratti di finitezza (Fin) come tratti di accordo/A:

(45) a. tratti For = tratti A'/operatoriali
b. tratti Fin = tratti A/accordo.

Questa categorizzazione è motivata dal fatto che mentre la proiezione di finitezza ospita dei tratti Agr questo non può accadere nel caso della sola proiezione di forza.

La ragione di questa scelta di ricondurre i vari tipi di tratti alla più tradizionale dicotomia A/A' – in questo contesto rappresentati come una opposizione tra proprietà di tipo operatoriale e proprietà di tipo Agr – è di mantenere un sostanziale parallelismo fra la duplice natura delle posizioni sintattiche-strutturali e i vari tratti pur nelle loro specifiche differenziazioni. In effetti, il rischio che si corre in una proliferazione indiscriminata di tratti è la perdita della specificità e del primato del valore geometrico-formale che connota la sintassi. Nella proiezione sincretica, a causa della sua natura ibrida, si verificherà la seguente distribuzione di tratti:

(46) a. Spec,For-FinP = tratti +For/operatoriali (+Op) e tratti +Fin/accordo (+Agr)

b. For-Fin° = tratti +For/operatoriali (+Op) e tratti +Fin/accordo (+Agr).

In inglese, quando si verifica l'estrazione del soggetto nelle interrogative, avremo che il soggetto nello Spec,For-FinP veicolerà allo stesso tempo tratti Agr e tratti operatoriali, mentre la testa For-Fin° ospiterà, a sua volta, tratti Agr e tratti operatoriali che verranno soddisfatti dal complementatore *that.* Si verifica una situazione dove tutti i tratti presenti sullo specificatore e sulla testa della proiezione sincretica vengono soddisfatti contemporaneamente determinando un configurazione di Accordo Spec-Head. Questa singolare situazione, che si verifica nel caso dell'estrazione del soggetto nelle interrogative dirette, soddisfa la condizione massimale (44) determinando l'incorporazione obbligatoria della testa sincretica nella proiezione Agr associata:

(47) Who$_i$ do you think [For-FinP[Spec $_{(+Op/+Agr)}$ t'$_{i\ (+Op/+Agr)}$][For-Fin° $_{(+Op/+Agr)}$ (*that) $_{(+Op/+Agr)}$][IP t$_i$ left?]].

Quello che è importante rilevare è che in questa analisi *that* dispone in modo intrinseco dei tratti For/A' e dei tratti Fin/A:

(48) That = tratti +For/+Op e tratti +Fin/+Agr.

In frasi come (43), la cui struttura è data in (49), *that* verifica i corrispondenti tratti sulla testa sincretica mentre i tratti presenti sullo specificatore non vengono verificati da nessun elemento. L'opzionalità del fenomeno della CD, in questo tipo di contesti, è determinata proprio dal fatto che i tratti presenti sull'intera proiezione vengono verificati solo sulla testa ma non sullo specificatore:

(49) I think [For-FinP[Spec $_{(+Op/+Agr)}$][For-Fin° $_{(+Op/+Agr)}$ (that) $_{(+Op/+Agr)}$][IP John left]].

In (49) la CD è soltanto opzionale poiché la condizione massimale d'incorporazione non è soddisfatta. Si può constatare, infatti, che in inglese la CD in contesti subordinati come (49) è molto utilizzata nel parlato quotidiano mentre in registri di discorso più formali o nello scritto non viene generalmente ammessa. Per contro, nel caso delle relative restrittive sull'oggetto come (4c), la cui rappresentazione strutturale è (50), pur verificandosi una situazione analoga a (49) in quanto l'oggetto transitando nello Spec della proiezione sincretica non verifica tutti i tratti presenti sulla posizione di Spec,For-FinP non disponendo di tratti +Agr ma solo di tipo +Op, si può constatare come la cancellazione venga ammessa anche a livello scritto e in registri più elevati. La ragione è da ricondursi al fatto che all'interno della proiezione sincretica viene a stabilirsi una relazione di Accordo Spec-Head almeno fra i tratti di tipo operatoriale:

(50) This is the man$_i$ [For-FinP[Spec $_{(+Op/+Agr)}$ t'$_{i\ (+Op/-Agr)}$][For-Fin° $_{(+Op/+Agr)}$ (that) $_{(+Op/+Agr)}$] [IP I saw t$_i$].

Sembra possibile formulare una classificazione gerarchica dei livelli di preferibilità della cancellazione compresa fra i due estremi rappresentati dalla condizione minimale e massimale d’incorporazione. I livelli intermedi di preferibilità si giustificherebbero sulla base delle differenze nella verifica dei tratti come verrà illustrato in (52). Veniamo, ora, al caso delle relative restrittive sul soggetto come (4d) dove l’effetto ‘*that*-trace’ non sembra manifestarsi. In queste strutture ci attenderemmo che la cancellazione del complementatore sia obbligatoria in quanto il soggetto estratto dalla sua posizione di base veicolerebbe tratti Agr determinando la relazione di Accordo Spec-Head ed ECP verrebbe soddisfatto. Come sappiamo avviene il contrario e la cancellazione di *that* è vietata. Avevamo già accennato nel §1 al fatto che la cancellazione del complementatore nelle restrittive sul soggetto sembrava bloccata da ragioni di processing, tuttavia, sulla base della strumentazione che abbiamo sviluppato, è possibile avanzare una soluzione sintattica di questo fenomeno. Nelle strutture come (51) non viene selezionata la testa sincretica ma l’opzione dissociata di forza e finitezza:

(51) This is the man [ForP $_{(+Op)}$ *(that) [FinP $_{(+Agr)}$ [IP t knows Mary]]].

Il complementatore *that* verifica i suoi tratti +Agr nella testa di finitezza ed in seguito si sposta nella testa di forza per verificare i tratti +Op. La testa di forza, essendo una testa A’, non può essere associata ad una proiezione pura di accordo, non verificandosi le condizioni minimali richieste in (41) per l’incorporazione il complementatore non può essere cancellato. L’ipotesi è che in (51) la traccia soggetto debba essere governata da una testa verbale. Quando si verifica la presenza di una testa sincretica il complementatore può incorporarsi nella proiezione Agr determinando la scomparsa dello strato Comp, mentre la proiezione di accordo resta trasparente al governo di una testa verbale. In (51), tuttavia, la traccia non può essere governata da una testa verbale e la sola testa disponibile per il governo resta quindi quella di finitezza adiacente all’IP. Si può facilmente verificare come nelle relative sul soggetto ma formate a partire da una struttura subordinata come (4e), o nelle interrogative sul soggetto come (47), entrambe sottoposte alla condizione massimale d’incorporazione, sia sempre disponibile una testa verbale per il governo della traccia e *that* deve essere obbligatoriamente cancellato. Nelle relative sul soggetto non resta dunque che la soluzione di dissociare le teste di forza e finitezza, al fine di permettere il governo della traccia.

Prima di estendere questa analisi anche all’italiano è possibile constatare come la verifica dei tratti di tipo operatoriale e di accordo, rispettivamente sia sullo specificatore che sulla testa della proiezione sincretica, si riveli perfettamente consistente con i giudizi anche più accurati dei parlanti nonché con i manuali di stile per quanto concerne i contesti anche formalmente più appropriati in cui la CD può venire applicata. Queste conclusioni, come abbiamo visto, sembrano sostenibili sulla base di un'analisi delle possibili combinazioni della verifica dei tratti all'interno della proiezione sincretica. I casi non attestati non vengono considerati:

(52)	Spec	Testa		
a.	+Op +Agr	+Op +Agr	=	estrazione del soggetto: interrogativa/relativa-subordinata/CD obbligatoria (condizione massimale)
b.	+Op -Agr	+Op +Agr	=	estrazione dell'oggetto: interrogativa/relativa /CD opzionale
c.	-Op -Agr	+Op +Agr.	=	struttura subordinata/CD opzionale (condizione minimale)

(52c) è il caso esemplificato da strutture come (43/49). La CD, in queste frasi, dove 2 dei 4 tratti presenti nella proiezione sincretica non vengono verificati, è del tutto naturale e ampiamente utilizzata nei contesti di parlato spontaneo-colloquiale mentre non è consigliata nella lingua formale-scritta e nei registri orali elevati. Al contrario, nel caso (52b), dove vengono verificati 3 tratti su 4, attestato da frasi come (4c/50), non solo la CD è ammessa nei vari registri della lingua orale ma anche in quella scritta. Sembra pertanto possibile concludere che la CD è un fenomeno meno opzionale e libero di quanto potesse apparire 'prima facie', e il merito dell'analisi in tratti, all'interno del contesto teorico complessivo in cui ci stiamo muovendo, è esattamente quello di mettere a disposizione una griglia categoriale molto fine tale da rendere conto di fenomeni sottili e diversificati.

Abbiamo visto in precedenza, nello schema (39), che in italiano, quando la frase subordinata non manifesta proprietà di tipo modale, la specificazione di finitezza non viene rappresentata. È il caso di frasi come (38) dove la flessione è al modo indicativo ed è la specificazione di forza a connettere la frase principale con la subordinata in quanto proprietà e tratto comune alle due strutture [15]. In inglese, invece, la subordinata disporrà nello stesso tempo di proprietà di forza e di finitezza che potranno venire rappresentate sincreticamente poiché il *that* possiede sia tratti For che tratti Fin e può quindi verificare entrambe le specificazioni all'interno di un'unica proiezione. Se ora consideriamo le frasi in (53), si osserva che, mentre in (53a) il contenuto proposizionale non è posto come possibile, in (53b) è da interpretarsi come modalizzato proprio per le caratteristiche del verbo principale. Queste differenze, in inglese, non verrebbero quindi registrate a livello del modo verbale della subordinata che è sempre il medesimo in entrambi i casi. È plausibile pensare, allora, che la subordinata disponga, al medesimo tempo, di proprietà di forza legate essenzialmente alla presenza del modo indicativo e di proprietà di finitezza legate alle proprietà temporali ed eventualmente modali del contenuto proposizionale, come abbiamo visto nello schema (39):

(53) a. John knows (that) Mary left
b. John thinks (that) Mary left.

In entrambi i casi la selezione della proiezione sincretica sarà effettuata per default come conseguenza del fatto che *that* può verificare direttamente entrambe le specificazioni. Quello che si verifica in italiano è che la specificazione di finitezza registra le proprietà modali della subordinata come conseguenza della presenza in questa lingua del modo congiuntivo. Nei contesti in cui questa forma flessionale non è presente, e non si manifestano quindi proprietà di tipo modale, *che* deve lessicalizzare direttamente la testa di forza e questo spiegherebbe immediatamente anche il divieto di cancellazione del complementatore in questi contesti. Si tratta, ora, di mostrare come funzioni nell'italiano la verifica dei tratti all'interno della proiezione sincretica in rapporto ai vari tipi di frase che manifestano CD.

Il dato più evidente dal quale dobbiamo partire sembra essere il seguente. Mentre in inglese si danno alcuni casi di CD obbligatoria – ad esempio l'estrazione del soggetto nelle interrogative – questo fenomeno non si verifica mai in italiano. Nel § 1, considerando le frasi in (5) analoghe a (54), avevamo rilevato che nei casi di estrazione del soggetto, sia nelle interrogative che nelle relative, la cancellazione del complementatore era preferibile:

(54) a. Chi credi (?che) t abbia visto Maria?
b. Le persone che credo (?che) t cerchino di fare il loro dovere non sono poche

Inoltre, in contesti con subordinazione retti da verbi di credenza come (55a), la CD non solo è puramente opzionale ma, qualora si verifichi, la frase appare stilisticamente più elevata di quella dove *che* non viene omesso, la soluzione nettamente più utilizzata nei registri colloquiali. Si noti come sia esattamente il contrario di quanto accade in inglese dove la cancellazione di *that* è un fenomeno estremamente diffuso nel parlato quotidiano:

(55) a. Credo (che) sia arrivato ieri
b. Credo (che) Gianni abbia visto Maria.

Sono questi i fenomeni che un'analisi in termini di verifica dei tratti di tipo operatoriale e di tipo Agr presenti nella proiezione sincretica deve spiegare. Assumiamo che in italiano il complementatore *che* sia marcato intrinsecamente solo da tratti Fin di tipo Agr:

(56) Che = tratti +Fin /+Agr.

La verifica dei tratti in caso di estrazione del soggetto in una struttura interrogativa come (54a) sarà la seguente:

(57) Chi$_i$ credi [For-FinP[Spec $_{(+Op/+Agr)}$ t'$_{i\ (+Op/+Agr)}$][For-Fin° $_{(+Op/+Agr)}$ (?che) $_{(-Op/+Agr)}$][IP t$_i$ abbia visto Maria ?]].

La rappresentazione a parentesi etichettate di (55a) è (58):

(58) Credo [For-FinP[Spec $_{(+Op/+Agr)}$][For-Fin° $_{(+Op/+Agr)}$ (che) $_{(-Op/+Agr)}$][IP sia arrivato ieri]].

La presenza di tratti Agr sul complementatore *che* ne rende possibile la cancellazione opzionale mediante incorporazione poichè la condizionale minimale è soddisfatta. Quello che non si verifica nella interrogativa sul soggetto in (57), in rapporto alla sua equivalente in inglese, è la relazione di Accordo Spec-Head con la conseguente verifica di tutti i tratti presenti nella proiezione sincretica, mentre la leggera preferibilità della versione con cancellazione verrebbe effettivamente spiegata dal fatto che tre tratti su quattro verrebbero verificati. In (58) l'opzionalità della CD, connotata dalla scelta di un registro più formale e quindi più limitato in rapporto all'analogo fenomeno nell'inglese, è invece riconducibile al fatto che solamente il tratto Agr viene verificato sulla testa sincretica. Vediamo, ora, il caso dell'estrazione dell'oggetto nelle interrogative e nelle relative restrittive. Si tratta di giudizi molto sfumati e sottili ma sembra comunque attestabile la leggera preferibilità della CD in caso di estrazione del soggetto in rapporto a quella dell'oggetto:

(59) a. Chi credi t abbia visto Maria?
b. Chi credi Maria abbia visto t?
c. Queste sono le persone che credo t abbiano aiutato Maria
d. Queste sono le persone che credo Maria abbia aiutato t.

Nell'estrazione dell'oggetto i tratti Agr sopra lo specificatore della testa sincretica non vengono verificati in quanto l'elemento interrogativo estratto dalla posizione oggetto dispone soltanto di tratti di tipo operatoriale:

(60) Chi$_i$ credi [For-FinP[Spec$_{(+Op/+Agr)}$ t'$_{i\ (+Op/-Agr)}$][For-Fin°$_{(+Op/+Agr)}$ (che)$_{(-Op/+Agr)}$][IP Maria abbia visto t$_i$?]].

Così come abbiamo proceduto in (52) per l'inglese, ricapitoliamo anche per l'italiano il quadro generale della verifica dei tratti Op/Agr in rapporto alla preferibilità della CD:

(61)	Spec	Testa		
a.	+Op	-Op	=	estrazione del soggetto: interrogativa/relativa/CD opzionale
	+Agr	+Agr		
b.	+Op	-Op	=	estrazione dell'oggetto: interrogativa/relativa/ CD opzionale
	-Agr	+Agr		(condizione minimale)
c.	- Agr	+ Ag	=	strutture subordinate: CD opzionale (condizione minimale).

Secondo questa classificazione in termini di maggiore o minore preferibilità della CD, risulta che nel caso dell'estrazione dell'oggetto la CD sarebbe preferibile in rapporto alla CD in una frase contenente una subordinata, come nelle frasi in (55). Questa conclusione può apparire a prima vista insoddisfacente se confrontiamo la sottile ma pur minore preferibilità della CD, nel caso di estrazione dell'oggetto come in (59b), in rapporto alla CD in strutture subordinate a soggetto nullo come (55a). Il confronto pertinente deve essere effettuato con una subordinata con soggetto lessicale realizzato come (55b), che qui ripetiamo come (62b):

(62) a. Queste sono le persone che credo Gianni abbia aiutato t
b. Credo Gianni abbia aiutato Maria.

Per quale ragione, dunque, una frase come (62b) risulterebbe preferibile quando il soggetto non viene realizzato? La risposta a questo interrogativo costituirà la parte conclusiva del presente lavoro.

Abbiamo visto nei paragrafi precedenti come le soluzioni prospettate da Poletto e Giorgi & Pianesi avessero un sostanziale elemento comune e cioè la possibilità che il verbo spostatosi in una qualche proiezione ibrida di tipo Comp potesse assegnare caso nominativo al soggetto attraverso una relazione di Accordo Spec-Head. In entrambe le prospettive la proiezione rilevante era una proiezione di tipo ibrido, AgrC per Poletto e Mood/Agr per Giorgi & Pianesi, ed in entrambi i casi si finivano per ammettere le due seguenti condizioni:

(i) C° assegna caso sotto reggenza
(ii) C° assegna caso attraverso Accordo Spec-Head.

Una soluzione che già di per sé stessa solleva numerosi problemi (cfr. Rizzi 1991), e che non permette di risolvere da sola il problema posto da frasi come (62b). Giorgi & Pianesi arrivavano a supporre che alcuni parlanti accettassero il soggetto realizzato nello Spec della proiezione ibrida poiché ‘interpretavano’ quella posizione come una posizione A, mentre i parlanti che non condividevano questo giudizio ‘interpretavano’ quello specificatore come una posizione A’. Nell’analisi che si è qui sviluppata la condizione (ib) è esclusa, gli unici contesti di assegnazione di caso in Comp sono legati al fenomeno Aux-to-Comp e sono di tipo sotto reggenza.

Sembra essere chiaro che il problema rilevato da un certo gruppo di parlanti, in rapporto a frasi come (62b), è un problema di caso e come tale deve essere risolto. La soluzione più plausibile sembra essere la seguente. Quando i tratti Agr presenti nella testa della proiezione sincretica di forza-finitezza vengono verificati, innescando il processo opzionale di incorporazione della testa sincretica nella proiezione Agr associata, la proiezione sincretica For-FinP scompare a causa dell’incorporazione. Quello che rimane a livello strutturale è la sola proiezione Agr, che molto probabilmente è trasparente al governo. Quello che si verifica è che alcuni parlanti potranno registrare e interpretare la normale assegnazione di caso accusativo sul complemento frasale, da parte del verbo della principale, come un’anomala assegnazione di caso accusativo al soggetto della subordinata, che verrebbe ad interferire, a sua volta, con la normale assegnazione di caso nominativo da parte della flessione della subordinata. La possibilità che Agr possa assegnare caso al soggetto della subordinata è esclusa, poiché Agr assegna caso attraverso Accordo Spec-Head.

Carlo Conni

Indirizzo dell'Autore: Université de Genève - Département de Philosophie 2,
rue de Candolle, 1211 - CH - connicarlo@hotmail.com
Via Botti, 10 - 29017 Fiorenzuola (PC) - I

Summary

The aim of this work is to give a unitary theoretical account of the phenomenon of complementizer deletion (CD) both in Italian and in English. In order to achieve this goal two strategies are pursued (i) the verification of Agr and operator features (ii) the incorporation of a syncretic head of force-finiteness in the Agr associated projection according to the theoretical assumptions developed by Rizzi (1997). Under these assumptions, the Comp system has become more articulated by adding an array of X-bar projections which bears illocutionary and interpretative properties. A relevant characteristic of the Comp system is to lie at the interface between the propositional content of the embedded sentence and the main clause and, on the other hand, the linguistic and extralinguistic context of the utterance. Given a Comp structure articulated as follows: ForceP, TopP, FocusP, TopP, FinitenessP, IP, the properties listed above should be represented alternatively on the Force projection, which represent the phrasal type, and on the Finiteness projection, which in turn selects the IP system with the familiar characteristics of finiteness: mood distinctions, subject agreement licensing nominative case, overt tense distinctions. When the topic-focus system is triggered the complementizer will be generated in the head of Force and the deletion is excluded. In cases, however, where the topic-focus system is not activated, this makes available a syncretic head of force-finiteness provided with Agr features which will be checked by incorporation of the syncretic projection in the associated Agr projection. The complementizer deletion will occur as a consequence of this incorporation:

(i) CD = For-FinP moves to Agr,For-FinP in order to check its Agr features.

The CD phenomenon is fundamentally facultative. CD is obligatory only in English in the case of interrogatives with subject extraction. In this case, a configuration of Spec-head-agreement between the syncretic head and the subject trace specified with Agr features will be licensed According to the current assumption, *that* will check operator and Agr features on the head while the trace will check its features in the Spec position. In English, the syncretic For-FinP projection is a default option in every subordinate structure. As a consequence of limited mood distinctions, *that* is intrinsically specified with both operator and Agr features. In Italian, the syncretic option is more constrained because of the verbal type of the main clause, as well as mood and tense distinctions of the subordinate clauses which, in certain cases, does not posess an independent Force projection. Therefore, the complementizer *che* is intrinsically specified only with Agr features. In this article two general schemata are drawn for each language to account for the differences in acceptability of the CD in three relevant contexts: subject extraction, object extraction and subordinate structures. The suggested explanation of the phenomenon is preceded by a general analysis of some alternative solutions. In particular, in paragraph 2, we examine the hypothesis of the CD as a case of movement of V to Comp, in terms of the typical phenomenon of 'verb-second' (Poletto 1994), and in paragraph 3, as a case of movement of subjunctive to a hybrid projection Mood/AgrP (Giorgi & Pianesi 1997).

Note

[1] Desidero ringraziare in modo particolare Luigi Rizzi, non solo per i preziosi consigli e i commenti a questo lavoro ma ancora di più per avermi offerto la possibilità di apprezzare il valore della ricerca scientifica. Ringrazio vivamente anche Alessandro Zucchi per i suoi importanti commenti a versioni precedenti di questo articolo. Vorrei ringraziare anche due anonimi referees per il loro attento e rigoroso contributo.

[2] Con 'frase completiva' e 'soggettiva' s'intendono frasi come: *che partirà domani* o: che *Gianni parta domani*. Cfr. G. Graffi (1994: 116).

[3] In diversi contesti la cancellazione del complementatore in presenza di una subordinata al tempo futuro sembra perfettamente ammissibile; è il caso di (2c) dove la frase è stativa, mentre ad esempio con frasi teliche come (2i) la CD ad alcuni parlanti appare meno preferibile. La CD è solitamente ammessa anche qualora il modo della subordinata sia il condizionale, come in (2d), tuttavia non è implausibile pensare che la CD, come attestano i giudizi dei parlanti, sia pienamente ammissibile, anche stilisticamente, solo in presenza di una subordinata al modo congiuntivo. Per quanto concerne l'esempio (2d), rappresentativo della maggioranza dei giudizi, non è da trascurarsi il fatto che alcuni parlanti manifestino delle difficoltà ad accettare la CD con il verbo della subordinata al condizionale, un fenomeno che non si verifica con il congiuntivo dove si registra una generale unanimità di giudizio:
(i) ? Credo uscirebbe dalla competizione
(ii) ? Credo scriverebbe a Dario.

[4] Con forza grammaticale autonoma dobbiamo pensare a quell'insieme di proprietà che determinano l'autosufficienza della frase e indirettamente la sua appartenenza ad una tipologia autonoma e specifica. Ritengo che non sarebbe implausibile parlare di forza frasale. La mancanza di forza grammaticale autonoma delle subordinate al modo congiuntivo si evince chiaramente dai seguenti esempi:
(i) Ho saputo *(che) ha chiesto di uscire
(ii) Ha detto *(che) è partito ieri
(iii) È partito ieri
(iv) Credo (che) sia partito ieri
(v) *(Che) sia partito ieri.
Come mi ha fatto notare un referee, è molto significativo che quando il verbo reggente ammette sia l'indicativo che il congiuntivo, la cancellazione del *che* sia pos-

sibile solo con il congiuntivo:
(v) Sono certo *(che) è partito
(vi) Sono certo (che) sia partito
(vii) Trovo *(che) è simpatico
(viii) Trovo (che) sia simpatico.
Anche l'esempio in (2n) è teso a mostrare, paradigmaticamente, come il modo indicativo, indipendentemente dalla tipologia del verbo che seleziona la subordinata, sia refrattario alla cancellazione. Il *che*, nel modo congiuntivo, ha il ruolo di fornire la forza della frase e la sua cancellazione ne implica direttamente la perdita come si evidenzia in (v). Nelle strutture subordinate al congiuntivo come (iv), la CD è ammessa poiché la forza della frase è comunque garantita dalla frase principale. Il fatto che in queste strutture subordinate il *che* sia eliminabile ci conduce a pensare che non occupi una testa di forza autonoma. L'obiettivo è dunque quello di individuare una rappresentazione sintattica adeguata del fenomeno della CD a partire da due constatazioni fondamentali: (i) che la forza della frase deve essere sempre assicurata, (ii) che là dove la CD non è ammessa si presume la presenza autonoma di un proiezione di forza che non può essere eliminata.

[5] Quello che si osserva è che alcuni parlanti, quando si verifica la cancellazione del *che* con soggetto preverbale realizzato, interpretano erroneamente la normale assegnazione di caso accusativo al complemento frasale della principale come una erronea assegnazione di caso accusativo direttamente al soggetto della subordinata.

[6] Il solo caso che ho trovato in cui la CD, nel giudizio di alcuni parlanti, è parzialmente ammissibile con una frase dipendente da un nome è (2f). Si noti tuttavia, nell'esempio seguente:
(i) La probabilità *(che) incontri Maria è molto bassa
come la presenza di un soggetto non impersonale, realizzato o meno, diversamente da (2f), renda la CD inaccettabile. Secondo Rizzi, invece, anche frasi come (2f) sono comunque inaccettabili. È chiaro che le strutture dipendenti da nomi dispongono di un solo centro di forza della frase collocato a contatto con la subordinata e che la cancellazione del *che* determina la perdita di questo centro.

[7] In frasi come (2h), e più in generale in contesti come: *vietare che_,ordinare che_*, ecc., la CD non è ammessa in quanto la frase subordinata, sebbene mostri una flessione al congiuntivo, appare piuttosto come una frase al modo imperativo e quindi dotata di una sua forza anche in rapporto alla principale:
(i) Venga assunto immediatamente
(ii) Ho vietato *(che) venga ammesso
(iii) Non venga ammesso.

[8] Fenomeni analoghi a (3c) e (3d) sono osservabili anche in italiano:
(i) Il fatto *(che) sia partito è sorprendente
(ii) *(Che) possa venire da noi è improbabile
(iii) *(Che) possa venire da noi, non lo credo proprio.
Ma in italiano, a differenza dell'inglese, nelle strutture relative non è possibile la cancellazione di *che*:
(iv) La vettura *(che) Gianni ha comprato ieri è formidabile
(v) Il ragazzo *(che) vive con Maria è simpatico.
In (i), (ii) e (iii) la CD non è ammessa perché, diversamente, la frase verrebbe privata della sua forza grammaticale. Il fatto che in (iii) la cancellazione sia vietata (si tratta un caso di dislocazione a sinistra), sembra dimostrare che la frase dislocata non è generata nel sintagma verbale e poi spostata, ma piuttosto che è generata all'inizio della frase e di conseguenza la forza della frase dislocata deve essere rappresentata là dove è presente il livello del complementatore. Nelle frasi relative (iv) e (v) la cancellazione determinerebbe egualmente la perdita della

forza della frase, forza che sembra essere assicurata dalle proprietà del complementatore adiacente alla testa della relativa che in italiano non è il risultato di un movimento-wh dal sintagma verbale. Per quello che concerne le relative in inglese, si assume l'analisi di Lasnik & Stowell (1991), dove il movimento della testa della relativa istanzia un movimento-wh. Questa ipotesi trova una evidenza anche nel fatto che nelle strutture relative è possibile utilizzare un elemento morfologicamente-wh in sostituzione di *that*, un fenomeno non riscontrabile in italiano:
(vi) This is the man who t will leave tomorrow
(vii) This is the book which John can't stand t.

[9] Su questo punto, cfr. Bever (1977) e Vincenzi (1991).

[10] Come si è già osservato nella nota 7, la presenza di comandi, e presumibilmente del modo imperativo, impedisce che la forza della frase venga rappresentata direttamente dal *che*, come accade invece nel congiuntivo, rendendone facoltativa la presenza come in (7).

[11] Un focus sopra il SN in posizione postverbale sembra rendere la frase accettabile:
(i) Che venisse GIANNI ad aiutarci almeno una volta.

[12] Il diacritico in (20) è quello assegnato da Giorgi & Pianesi. Tuttavia, non si può non constatare che contesti come il seguente:
(i) Mi dispiace (che) sia partito
(ii) Mi rammarico (che) sia partito
rappresentano dei casi problematici per le tesi di Giorgi & Pianesi e in generale per qualsiasi ipotesi che pone come condizione necessaria per la CD la presenza di verbi non-fattivi o modali nella frase matrice. In (i) la verità del contenuto proposizionale subordinato è presupposta e il contenuto è quindi posto come dato. Contesti come (iii) implicano sempre (iv):
(iii) Mi spiace che P
(iv) Il parlante sa che P.
Può quindi essere problematico sostenere che la cancellazione del complementatore dipende dal fatto che il contenuto subordinato è posto come dato a causa delle proprietà di selezione semantica dei verbi di tipo modale. Si potrebbe sostenere, come mi ha fatto notare un referee, che se è vero che la classe dei verbi che ammettono la cancellazione può essere identificata in larga misura cone la classe dei verbi non-fattivi, è possibile, tuttavia, che questa corrispondenza non sia parte della grammatica dell'italiano, ma che serva solamente come euristica per colui che apprende l'italiano per identificare la classe dei verbi portatori del tratto sintattico che rende possibile la CD. Secondo questa ipotesi, non esisterebbe nessuna regola dell'italiano che associ una interpretazione semantica ai tratti sintattici caratteristici della CD. Non mancano, tuttavia, fenomeni che indicano l'esistenza di complessi rapporti, come ad esempio il fatto che in contesti come (i), (ii), dove la CD è ammessa, a differenza di (20), il soggetto della frase principale coincida con il parlante, mentre in presenza di un 'reporter' dell'enunciato la possibilità di cancellazione si degrada:
(v) Marco e Luigi si dispiacciono ??(che) abbia litigato con Maria
(vi) A Gianni dispiace *(che) Marco abbia litigato con Maria.
Nell'analisi di Giorgi & Pianesi, l'articolazione del complementatore in differenti proiezioni è determinata, in ultima analisi, dalle qualità dei tratti che sono in gioco. Quando il verbo principale è di tipo fattivo, avremo un tratto +fact che ha la proprietà di non poter apparire sincreticamente con un tratto +mood, come conseguenza dell'applicazione del 'Feature Scattering Principle'. I due tratti dovranno essere rappresentati su due proiezioni diverse e la CD sarà esclusa poiché il complementatore lessicalizzerebbe la testa con il tratto +fact. Il fatto che in contesti

come (i) la CD sia ammessa sembra dimostrare che la presenza di un tratto +fact non sia una condizione sufficiente per impedire la CD.

[13] Il complementatore *di* deve sempre occupare la testa di finitezza adiacente all'IP. Nell'italiano contemporaneo non è mai cancellabile, mentre nell'antico toscano poteva esserlo facilmente:
(i) Mi parrebbe _ rimanere troppo sola
(ii) Se io credessi _ poter aggiungere (Scorretti, 1981).

[14] Pesetsky, in questo libro, accenna ad una sua vecchia ipotesi sulla natura del fenomeno della CD. Secondo questo autore si può constatare come in inglese la variante zero del complementatore abbia la distribuzione caratteristica delle tracce, vale a dire di occorrere in contesti di governo, adiacente al verbo principale. L'idea di Pesetsky è che la variante zero del complementatore è consistente con ECP ed è una sorta di affisso che si incorpora sul verbo. Questa analisi, a cui accenna Pesetsky, è consistente con numerosi dati empirici dell'inglese, ma non lo è con il caratteristico fenomeno della cancellazione opzionale del complementatore nelle frasi relative sull'oggetto:
(i) This is the man (that) I saw.
Pesetsky, in una conferenza tenuta al DIPSCO di Milano nel 1994, ha proposto una teoria della CD nei termini della 'Teoria dell'Ottimalità'. Mi limito solo ad accennare all'esistenza di questa ipotesi alternativa, secondo la quale la cancellazione di *that,* nei contesti subordinati e nelle frasi relative, è un effetto del 'telegraphic speech', dello stesso tipo che troviamo negli stadi iniziali dell'acquisizione del linguaggio e nelle afasie. L'omissione di parole funzionali non implicherebbe l'assenza della corrispondente categoria funzionale, ma sarebbe piuttosto l'effetto di un principio più generale che impedisce la pronuncia di parole funzionali. Il principio è il seguente:
(ii) Telegraph: un morfema funzionale non deve essere pronunciato
(iii) Deletion: marca x [+ silente].

[15] È possibile sostenere che in italiano, in frasi come (38), la struttura del complementatore all'interfaccia fra la principale e la subordinata preveda anche la proiezione di finitezza ma dissociata da quella di forza:
(i) Ha detto [ForP *(che) [FinP [IP è partito]]].
Come abbiamo già detto, assumiamo che in italiano *che* dispone intrinsecamente solo di tratti Fin/Agr. Nelle strutture come (i) la rappresentazione indipendente della testa di forza permette al complementatore di verificare questa specificazione, lessicalizzando direttamente quella testa. Al contrario di *that*, che dispone sempre intrinsecamente di tratti For/+Op e Fin/+Agr e ha la possibilità di verificare questi tratti anche nella proiezione sincretica, il complementatore *che* dispone solo di tratti Fin/+Agr. La sola strategia per verificare i tratti di forza sulla specificazione di forza è quella in cui *che* lessicalizza direttamente la specificazione di forza in una proiezione indipendente. Se si adotta questa ipotesi, si deve convenire che il *che* dovrà essere generato nella proiezione di finitezza, dove potrà verificare i suoi tratti Fin. A questo punto, l'incorporazione nella proiezione Agr, presumibilmente associata a FinP, non sarà possibile poiché *che* deve spostarsi ancora nella forza. Quello che si può osservare, in questo caso, è che il movimento del complementatore *che* non sarà dettato dà ragioni di tipo 'egoistico', per verificare dei propri tratti, ma, contrariamente al principio GREED, seguirà motivazioni di tipo 'altruistico' per verificare le proprietà di forza della subordinata rappresentate sulla proiezione di forza. Lasnik (1995) si schiera a favore dell'ipotesi altruistica: "Chomsky (1995) argues that movement conforms to GREED: items move only to satisfy their own requirements. I conclude the last resort condition is not GREED but Enlightened Self-Interest: items move either to satisfy their own requirements or those of the position they move to" Lasnik (1995).

Riferimenti bibliografici

BEVER, Thomas G. (1977), *An integrated theory of linguistic ability*, Hassocks Sussex, The Harvester Press.

BIANCHI, Valentina (1999), *Consequences of antisymmetry: Headed Relative Clauses*, Berlin, London, Mouton de Gruyter.

CHOMSKY, Noam (1995), *The Minimalist program*, Cambridge Mass., MIT Press.

CULICOVER, Peter W. (1993), "The Adverb Effect: Evidence against ECP accounts of the that-t effects", *NELS*, 23: 97-111.

DE VINCENZI, Marica (1991), *Syntactic parsing strategies in italian. The minimal chain principle*, Dordrecht, Kluwer Academic Press.

GIORGI, Alessandra & Fabio PIANESI (1997), *Tense and aspect: from semantics to morphosyntax*, New York, Oxford University Press.

GRAFFI, Giorgio (1994), *Sintassi*, Bologna, Il Mulino.

LASNIK, Howard & Tim STOWELL (1991), "Weakest Cross-over", *Linguistic Inquiry*, 22: 687-720.

LASNIK, Howard (1995), "Case and Expletive Revisited: On Greed and other Human Failing", *Linguistic Inquiry,* 26: 615-633.

PESETSKY, David (1995), *Zero syntax*, Cambridge Mass., MIT Press.

PESETSKY, David (1994), "Telegraphic Effects and Optimality", conferenza tenuta al DIPSCO di Milano.

POLETTO, Cecilia (1994), "Complementizer Deletion and Verb Movement in Italian", manoscritto, Università di Padova e Venezia.

POLLOCK, Jean-Yves (1989), "Verb Movement, UG and the Strucutre of IP", *Linguistic Inquiry*, 20: 365-424.

RIZZI, Luigi (1990), *Relativized minimality*, Cambridge Mass., MIT Press.

RIZZI, Luigi (1991), "Residual Verb Second and the Wh Criterion", *Technical Reports in Formal and computational Linguistics*, Università di Ginevra, 2: 1-28.

RIZZI, Luigi (1997), "The Fine Structure of the Left Periphery", in Liliane Haegeman ed. (1997), *Elements of grammar*, Dordrecht, Kluwer Academic Press.

SCORRETTI, Mario (1981), "Complementizer Ellipsis in 15th Century Italian", *Journal of Italian Linguistics*, 6: 35-46.

TOMASELLI, Alessandra (1990), *La sintassi del verbo finito nelle lingue germaniche*, Padova, CLESP.

SHLONSKY, V.R. (1994), "Semitic Clitics", *Gen GenP*, Università di Ginevra, 1: 1-11.

Il ruolo dell'udito nella comunicazione linguistica. Il caso della prosodia

Federico Albano Leoni

L'articolo tratta del ruolo dell'ascoltatore e dell'udito nella comunicazione parlata, con particolare attenzione all'intonazione. Dopo una breve presentazione del problema e una sommaria rassegna degli studi linguistici sull'argomento (§ 1), viene fornita una descrizione delle caratteristiche dell'apparato uditivo, della sinergia tra udito e voce e del ruolo degli schemi primitivi e degli schemi appresi nella decodifica del segnale linguistico (§§ 2-4). Nel § 5 viene discusso il ruolo dell'intonazione vista come un potente componente della comunicazione parlata, la cui importanza è universalmente riconosciuta, ma il cui studio e la cui descrizione appaiono ancora problematici a causa della sua intrinseca variabilità, della sua interconnessione con il continuum degli aspetti pragmatici e della mancanza di una tradizione di annotazione scritta. Nel § 6 vengono presentate alcune conclusioni e alcuni suggerimenti per un programma di lavoro sull'intonazione da condurre sul parlato spontaneo e tenendo conto delle verifiche percettive da parte dei parlanti[*].

1. Premessa

È nota a tutti l'asimmetria tra il lessico del parlare, che è ricco, e il lessico del sentire linguistico, che è povero (De Mauro 1994b), al punto che per quest'ultimo spesso manca, come in italiano o in inglese, un verbo specifico. Uno dei motivi, forse il principale, di questa asimmetria risiede, a parere di molti (p. es. Ryalls 1996: 3-4), nella profonda differenza tra il parlare e l'udire: il primo è in gran parte esterno, visibile, percepibile e autopercepibile; il secondo è interiore, invisibile, sfuggente. Posso vedere chi parla (anche se non sento quello che dice), ma non posso vedere l'ascolto; posso osservare e percepire alcuni movimenti del mio apparato fonatorio (laringe, bocca, lingua, labbra), ma non posso vedere o percepire i movimenti dell'apparato uditivo (timpano, ossicini, perilinfa, membrana basilare ecc.) mio o dei miei interlocutori. Sembra dunque ragionevole supporre che un processo palese ed evidente sia più presente alla coscienza dei parlanti e arrivi quindi a una rappresentazione verbale più articolata di un processo meno palese ed evidente.

Questa asimmetria ha altre conseguenze, oltre a quella della diversa rappresentazione dei due processi sul piano lessicale. Qui ci interessa quella per cui i processi della comunicazione parlata, raffi-

(ricevuto nel giugno 2001)

gurati, ad esempio, nel circuito della *parole* di Saussure (1922: 27), sono stati studiati prevalentemente, a volte esclusivamente, osservando l'emittente o osservando il messaggio, cioè rispettivamente il parlante o il testo, come si può facilmente verificare osservando a caso manuali o opere di riferimento che riflettono il senso comune corrente in linguistica e in fonologia (come Akmajian *et al.* 1995; Simone 1995; Nespor 1993; Kenstowicz 1994; Goldsmith 1995; Laver 1991 ecc.) [1]. È come se una parte non trascurabile della linguistica manifestasse un pensiero "primarily anchored to saying-without-listening" (Corradi Fiumara 1990: 3).

Naturalmente nessuno può pensare che i linguisti debbano occuparsi obbligatoriamente anche dell'udito o dell'ascolto o del ricevente, ma non si può non osservare una preferenza, domandarsi quali ne siano le ragioni e quali ne siano gli effetti sulle nostre conoscenze.

Un motivo di questa preferenza è certamente il fatto che la rappresentazione scritta della lingua, oggetto di studio obbligato in passato e prevalente ancora oggi, valorizza e stabilizza il prodotto, il testo, ma non induce a studiare la ricezione uditiva. Un secondo motivo, anche importante, è dato dalla natura interiore e invisibile della percezione, che richiede tecniche e metodi di osservazione, elicitazione e studio che la linguistica ritiene estranei ai suoi apparati.

Questa preferenza si manifesta in maniera molto chiara anche negli studi fonetici, dove per molto tempo è stato evidente il predominio della descrizione articolatoria (cioè del produttore) finalizzata alla descrizione e alla trascrizione del prodotto (cioè del testo). A partire dai primi decenni del Novecento è andato diffondendosi lo studio del versante acustico del messaggio (nato nella sua forma moderna con l'abate Rousselot alla fine dell'Ottocento e reso oggi facile dalla diffusione di strumenti economici e di uso relativamente semplice). Ma anche i risultati di queste analisi sono usati prevalentemente o per descrivere la struttura dei suoni in sé (cioè, ancora una volta, il prodotto), o per descrivere la relazione tra la struttura acustica e i gesti articolatori che l'hanno generata (cioè, ancora una volta, il produttore). Insomma, nella ricerca linguistica l'udito e il ricevente sono spesso trascurati perché in passato erano considerati argomenti propri dei filosofi e oggi sono considerati argomenti degli psicologi o di quanti si occupano della fisiologia della percezione [2].

In questo articolo non affronterò la storia di questo interessante problema [3] ma mi limiterò a qualche osservazione preliminare, prima di passare all'argomento che vorrei trattare.

Tra la fine dell'Ottocento e i primi del Novecento sembrò che l'udito e il ricevente richiamassero l'attenzione dei linguisti. Saussure, sia pure solo per cenni, aveva intuito la necessità di una svolta decisa. Ecco due passi molto istruttivi (sempre attuali i commenti di De Mauro alle note 113 e 114 dell'edizione italiana; ricordiamo che per Saussure l'aggettivo *acoustique* significa 'uditivo'):

> Beaucoup de phonologistes s'attachent presque exclusivement à l'acte de phonation, c'est-à-dire à la production des sons par les organs (larynx, bouche, etc.), et négligent le côté acoustique. Cette méthode n'est pas correcte: non seulement l'impression produite sur l'oreille nous est donnée aussi directement que l'image motrice des organs, mais ancore c'est elle qui est la base naturelle de toute théorie (Saussure 1922: 63).

> La délimitation des sons de la chaîne parlée ne peut donc reposer que sur l'impression acoustique; mais pour leur description, il en va autrement. Elle ne saurait être faite que sur la base de l'acte articolatoire, car les unités acoustiques prises dans leur propre chaîne sont inanalysables (*ibid.*: 65).

Ancora più esplicite sono alcune sue osservazioni contenute nel manoscritto di Harvard (Marchese ed. Saussure 1995: 98-100, 105, 113, e in particolare p. 98, da cui cito):

> Un acte phonatoire est un fait (ou plutôt) un ensemble de faits physiologiques correspondant à un fait phonétique déterminé. Les fait phonétique nous étant a son tour donné par la sensation auditive, c'est d'après cette dernière seule que nous fixons les actes phonatoires.

Saussure aveva dunque assunto una posizione decisa a favore del primato degli aspetti uditivi e la sua intuizione aveva trovato conferma nella ricerca fonetica successiva, nei modelli sulle strutture foniche delle lingue, negli studi sui disturbi del linguaggio, nelle osservazioni sul fonosimbolismo (oltre che, naturalmente, negli studi di psicolinguistica). Rappresentanti illustri di questa linea erano stati, tra gli altri, fonetisti e linguisti come Gunnar Fant e Roman Jakobson. Mezzo secolo dopo Saussure, Jakobson & Halle (1956: 33-34) si erano espressi con grande chiarezza:

> In order to decode the message, its receiver extracts the distinctive features from the perceptual data. The closer we are in our investigation to the destination of the message, the more accurately can we gauge the information conveyed by the sound chain. This determines the operational hierarchy of levels in their decreasing perti-

> nence: perceptual, aural, acoustical and motor (the latter carrying no direct information to the receiver except for the sporadic help of lip-reading) […] [4].

Questa linea acustico-uditiva non ebbe però successo tra i linguisti e subì una battuta d'arresto nel corso degli anni Sessanta, quando, per vari motivi, la ricerca fonetica tornò ad orientarsi verso gli aspetti articolatori. Questo cambiamento di direzione è ben mostrato da tre fatti emblematici.

Il primo fu l'affermarsi di una teoria della percezione nota come *Mothor Theory* (Liberman *et al.* 1963, 1985) che riconduce i processi della percezione linguistica alla riproduzione interiore dei gesti articolatori [5], recuperando così, forse senza saperlo, una tradizione filosofica e psicologica ottocentesca che risaliva a Maine de Biran (Albano Leoni & Dovetto 1996) [6]. Naturalmente Liberman era lontanissimo dal ritenere che la sensazione uditiva fosse irrilevante ai fini della percezione linguistica e la sua posizione era molto più complessa, come appare dalla sua autobiografia scientifica, posta come introduzione a un volume del 1996 che raccoglie i suoi scritti più significativi (Liberman 1996: 1-44). Ma ugualmente essa ha facilitato, o rinforzato, la tendenza alla rimozione degli aspetti uditivi.

Il secondo è che la fonologia generativa, fin dal suo nascere (Chomsky & Halle 1968), ridimensionava drasticamente il ruolo dell'udito:

> […] what the hearer 'hears' is what is internally generated by rules. That is, he will 'hear' the phonetic shape determinated by the postulated syntactic structure and the internalized rules (Chomsky & Halle 1968: 24)

Probabilmente in conseguenza di ciò, pur conservando l'impostazione binarista nella descrizione dei suoni, Chomsky e Halle cancellavano dalle matrici ogni tratto acustico-uditivo e generalizzavano quelli articolatori.

Il terzo è la pubblicazione del lavoro di Lenneberg (1967) che, pur dichiarando nell'introduzione che non avrebbe trattato dei meccanismi concreti della produzione e ricezione del linguaggio (sia perché troppo tecnici, sia perché irrilevanti ai fini della teoria che intendeva esporre), dedica invece uno spazio molto ampio alla anatomia e fisiologia dell'apparato fonatorio (cap. 2, II; 3, II-V) ma solo una veloce e marginale allusione all'udito (cap. 3, V, 1b-c), pur trattando dei fondamenti biologici del linguaggio. Ciò potrebbe sorprendere, ma se si considera che in appendice al volume di Lenneberg compariva un saggio di Chomsky, il silenzio sulla percezione appare abbastanza ovvio.

Naturalmente questi autori non possono ignorare che l'udito è l'ingresso principale per il destinatario del messaggio, ma ritengono che esso non contribuisca agli aspetti profondi della comprensione o perché questi sono basati su schemi innati, o perché essi sono strettamente associati ai meccanismi neuromuscolari della produzione.

Si chiudeva così un circolo che, almeno in linguistica, non si è ancora riaperto.

L'ascolto e la percezione uditiva della lingua sono invece oggetto di studio e di discussione in un ambito disciplinare che risulta dalla felice commistione di psicoacustica, psicolinguistica, neurolinguistica e fonetica, come appare, p. es., dai lavori di Hess (1983), Moore (1988), Gernsbacher (1994), Goodman & Nusbaum (1994), Hardcastle & Laver (1997, specialmente nei contributi di Delgutte, di Moore e di McQueen & Cutler) e di Pickett (1999). Da qui prenderò spunto per fare qualche riflessione sul ruolo dell'udito e del ricevente dal punto di vista linguistico, senza presumere di dire cose nuove, ma solo sistemando e accostando notizie e dati che vengono da settori di studio diversi [7].

Cercherò quindi di ricordare non solo che l'udito svolge un ruolo importante, almeno da comprimario, nella comunicazione audioverbale, ma anche che in esso si incarna con particolare evidenza quello che Simone (1992: 47-48) ha chiamato il "principio del determinismo fisico", cioè l'insieme degli effetti che la natura psico-fisica dei nostri apparati di produzione e di ricezione provoca sulla struttura e l'uso delle lingue. In questo senso, dunque, questo lavoro vorrebbe essere anche un contributo indiretto alla discussione sulla iconicità linguistica, intesa nel senso di 'limitazione dell'arbitrarietà'. So bene che questa discussione, mutati i termini, è antica (sintesi in Simone 1992 e 1995, Gensini 1993), ma qui considererò la questione solo nella prospettiva odierna.

Cercherò infine di trarre qualche conclusione, partendo dall'udito e dalla prosodia, su alcuni aspetti generali della decodifica e della semiosi linguistiche. Il punto di vista che qui vorrei presentare può essere espresso con le parole di Rosen & Fourcin (1986: 373):

> In evolution, the senses of vibration and hearing preceded speaking, and so it is inescapable that basic auditory constraints moulded the way speech developed. In short, trying to understand the auditory processes involved in the perception of speech is likely to lead not only to a better undestanding of hearing, but also, of speech itself.

2. Perché l'udito è importante per i linguisti?

La voce e l'udito concorrono alla realizzazione della comunicazione parlata ma non sono del tutto alla pari. Se si osservano l'anatomia e la fisiologia degli apparati vocale e uditivo si vede che l'apparato vocale è costituito da organi (polmoni, bronchi e trachea, laringe, faringe, cavità orale, lingua, denti, labbra, cavità nasali) che hanno come funzioni primarie la respirazione e/o la nutrizione e per i quali il parlare è una funzione secondaria. L'apparato uditivo è costituito invece da organi (padiglione, timpano, ossicini, coclea, organo del Corti) che hanno come funzione primaria quella dell'udire (e del localizzare la sorgente sonora). L'ascolto è dunque coessenziale alla nostra biologia, il parlare no. Non è inoltre privo di rilievo il fatto che se non si sente non si parla (i sordomuti sono in realtà sordi): il parlare implica l'udire (ma l'udire, almeno in teoria, non implica il parlare) [8].

L'apparato uditivo è completo e funzionante prima della nascita: il feto sente e discrimina rumori e suoni interni ed esterni già al settimo mese della gestazione e il neonato sente e discrimina suoni, voci e ritmi con grande finezza (Mehler & Dupoux 1990 [1992]: 163-223; Hawkins 1999; molta bibliografia in Pennisi 1994: 268-75). L'apparato vocale, invece, si completa dopo la nascita, quando la laringe si abbassa e consente la formazione di una cavità faringale relativamente ampia, fondamentale per la produzione di vocali diverse. Inoltre esso diventa perfettamente operante solo dopo un tirocinio piuttosto lungo [9].

Queste osservazioni servono a dire che nella architettura degli strumenti della comunicazione verbale umana, l'udito ha una preminenza biologica e funzionale, anche se, naturalmente, tra le nostre capacità vocali e le nostre capacità uditive esiste una stretta correlazione neurologica, oltre che funzionale [10].

Tenendo presenti queste premesse, nel prossimo paragrafo cercherò di mostrare come si manifesta, in concreto, l'interazione tra voce e udito.

3. Il campo uditivo umano e le caratteristiche acustiche del segnale linguistico

La fig. 1 rappresenta il campo uditivo umano, cioè l'insieme dei suoni mediamente percepibili da un umano normoudente ed è ricavato dalle risposte di numerosi soggetti alla somministrazione di segnali acustici variati in frequenza e intensità.

La rappresentazione consiste in un diagramma in cui ogni punto è determinato da un valore in frequenza (in ascissa) e un valore in intensità (in ordinata). L'intervallo di frequenza va da circa 16 a circa 20.000 Hz, mentre l'intervallo di intensità va da -4 a 140 dB. Il limite superiore del campo uditivo è la soglia del dolore, che qui non ci interessa; il limite inferiore, cioè la soglia di udibilità, mostra invece il variare dell'intensità necessaria per percepire stimoli acustici alle diverse frequenze. Se lo si osserva con attenzione, si vede che la zona in cui il rapporto energia/frequenza è ottimale, cioè quella in cui il nostro orecchio percepisce suoni anche di debole e debolissima intensità, va *grosso modo* dai 2000 ai 5000 Hz (dove vengono percepiti suoni al livello della soglia minima di energia); ma anche la fascia che va da 200 a 7000 Hz è molto sensibile perché per percepire questi suoni è sufficiente un'energia pari a circa 10 dB [11].

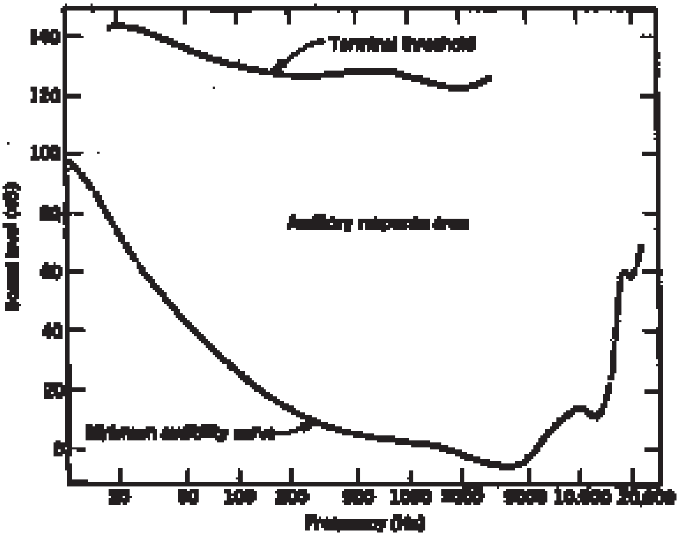

Fig. 1. Il campo uditivo umano (da Moore 1988: 48).

Il campo uditivo rappresenta dunque il primo, ovvio vincolo alla qualità acustica dei suoni linguistici i quali devono porsi necessariamente all'interno dei suoni udibili e preferibilmente all'interno di quella fascia di frequenza in cui risultino più facilmente percepibili.

A questa osservazione ne va aggiunta una seconda. Se il campo uditivo può essere considerato uno spazio fisico continuo, in cui ogni punto è definito da due valori 'oggettivi' (intensità e frequenza), dal punto di vista delle capacità umane di discriminare tra suoni diversi per altezza e volume (capacità fondamentale anche per gli usi linguistici) [12], esso è articolato al suo interno in modo discontinuo, e le modalità della discriminazione si distribuiscono lungo scale 'soggettive', determinate, ancora una volta, dalla anatomia e dalla fisiologia dell'apparato uditivo, e in particolare dalla caratteristica della tonotopicità e dal fenomeno delle bande critiche (Zwicker 1961).

I tests psicoacustici mostrano inoltre che la capacità di discriminazione dell'orecchio decresce al crescere della frequenza del segnale. La figura 2 presenta in ascissa le frequenze del segnale e in ordinata la differenza in Hz necessaria perché due segnali vengano riconosciuti come diversi o perché venga riconosciuta una variazione di altezza: alle frequenze basse, fino a 250 Hz, il nostro orecchio è in grado di discriminare tra toni puri che differiscono in frequenza anche per una frazione di Hz. Al crescere della frequenza aumenta la differenza necessaria perché due suoni vengano percepiti come diversi. Considerazioni simili possono essere fatte per la discriminazione di differenze di intensità e di durata. Questo meccanismo ha il suo fondamento nella anatomia e nella neurofisiologia dell'orecchio interno.

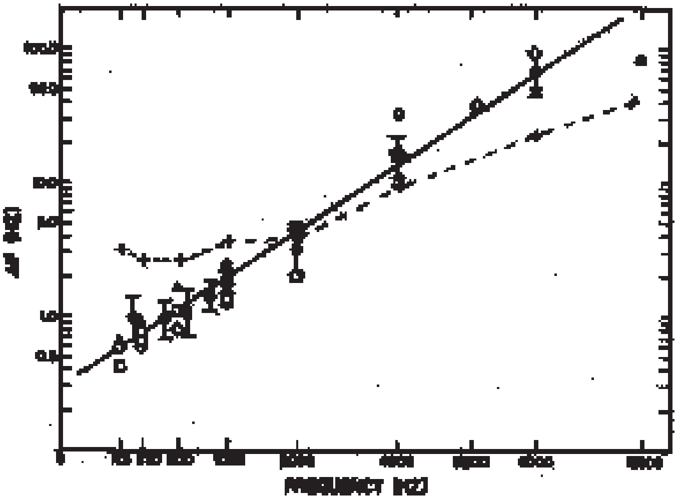

Fig. 2. Capacità di discriminazione dell'orecchio in rapporto alla frequenza del segnale (da Moore 1988: 160 con modifiche).

I dati che ho presentato finora sono basati sull'osservazione di toni puri. Essi forniscono indicazioni sulle caratteristiche fisiologiche della percezione, dalle quali non si può prescindere, ma non si può dimenticare che i toni puri sono praticamente assenti dal segnale linguistico che è costituito da oscillazioni complesse quasiperiodiche e aperiodiche.

Lo studio della percezione e riconoscimento dei suoni linguistici è molto più complicato. Prescindendo qui dal problema generale della percezione primitiva e della percezione basata su schemi (v. *infra* p. 13), basterebbe ricordare, a proposito della variabilità osservata degli spettri dei suoni vocalici, la questione del timbro metamerico (Bregman 1990: 122) [13], o la questione della percezione del *pitch*, ossia del correlato uditivo della frequenza fondamentale (che nei segnali linguistici è funzione delle variazioni di velocità del meccanismo laringeo). Qui mi soffermerò sul secondo problema.

La psicoacustica mostra che la percezione del *pitch* in un segnale periodico complesso avviene secondo modalità parzialmente diverse, a seconda che esso si trovi a frequenze basse o a frequenze alte. Qui considererò il primo caso perché esso è appunto quello generato dal meccanismo laringeo. Il *pitch* di bassa frequenza viene riconosciuto o perché è fisicamente presente una prima armonica di ampiezza percepibile, o perché viene ricostruito dall'orecchio attraverso un'analisi delle prime cinque armoniche. Questo secondo procedimento è importante perché la frequenza fondamentale della voce umana, per esempio di un maschio adulto, si aggira intorno ai 150 Hz, e può arrivare ai 400 Hz nella voce di un bambino [14] e può essere di debole intensità. Se ora si torna al campo uditivo si vede che le frequenze al di sotto dei 500 Hz sono percepibili solo ad una intensità pari ad almeno 20 dB. Ma la terza, quarta e quinta armonica di un segnale complesso che abbia la prima armonica a 150 Hz si collocano tra 450 e 750 Hz, cioè in una fascia di frequenze al tempo stesso meglio percepibili e nelle quali l'orecchio ancora discrimina variazioni anche piccolissime. È questo il complesso meccanismo che consente il riconoscimento dell'intonazione, sul quale tornerò più avanti [14]. L'orecchio è dunque attrezzato per percepire e discriminare tanto toni puri periodici, quanto suoni complessi quasi-periodici, affetti da rumore e da piccole oscillazioni di frequenza (dette j*itter*) intorno a un valore centrale, nei quali alcune componenti possono essere mascherate o cancellate o troppo deboli ed è dunque in grado di sopperire alle imperfezioni dell'apparato fonatorio:

> It is easy to show that, with respect to both measuring range and accuracy of the system, the ear outperforms the speech tract by far; the main constraints on the range of F0 and the accuracy of the

> realization of a periodic signal in speech are due to the process of speech production (Hess 1983: 63)

Si deve quindi osservare che la pertinenza percettiva delle variazioni di uno stato fisico prodotte dalla fonazione (cioè la base materiale su cui si esercita la capacità linguistica di distinguere) non è determinata solo dalle variazioni in sé, ma anche dai meccanismi psicoacustici. Quindi per capire come funziona un messaggio verbale non è sufficiente studiare tutte le caratteristiche dei gesti articolatori che l'hanno prodotto, o tutte le caratteristiche spettrali che ne sono il risultato, ma è necessario valutare queste ultime sulla base della loro significanza percettiva.

> No matter how systematically a phenomenon may be found to occur through a visual inspection of F0 curves, if it cannot be heard, it cannot play a part in communication ('t Hart *et al.* 1990: 25).

In altre parole, è lecito affermare che l'apparato uditivo condiziona due volte il meccanismo della comunicazione audioverbale: una prima volta, come abbiamo visto, determinando lo spazio fisico complessivo all'interno del quale la comunicazione avviene in modo ottimale; una seconda volta, determinando il livello minimo della percepibilità delle variazioni fisiche.

Il livello minimo della percepibilità delle variazioni fisiche non coincide però necessariamente con il livello minimo della loro pertinenza linguistica. Ciò significa, ancora riguardo al *pitch*, che la percezione di una variazione della melodia di una frase richiede una variazione della frequenza superiore alla soglia della discriminabilità meramente psicoacustica (Hess 1983: 78-79).

Se si osservano ora i tracciati spettroacustici [16] di un numero grande a piacere di enunciati linguistici naturali, si vede facilmente che l'energia non si distribuisce a caso nello spettro.

Se prendiamo le figure 3 e 4 come esempio rappresentativo della struttura acustica di enunciati umani, si vede che la banda di frequenza tra 200 e 5000 Hz è quella in cui si colloca la parte maggiore e più rilevante dei suoni linguistici prodotti dall'apparato fonatorio. Al di sopra di questo limite l'energia è in genere poca, limitata ad alcune fricative e, comunque, grazie alla ridondanza, non strettamente necessaria al riconoscimento dei suoni, come mostra, tra l'altro, la comunicazione telefonica analogica nella quale il canale fisico taglia le frequenze superiori a 2500/3000 Hz e inferiori a 500 Hz.

La correlazione tra prodotto articolatorio e capacità uditive è dunque particolarmente evidente. L'energia acustica si trova, per così dire, là dove l'orecchio è in grado di riceverla meglio. La messa in relazione delle capacità uditive umane e della distribuzione dell'energia acustica nel parlato mostra una ottima sinergia tra produzione e ricezione.

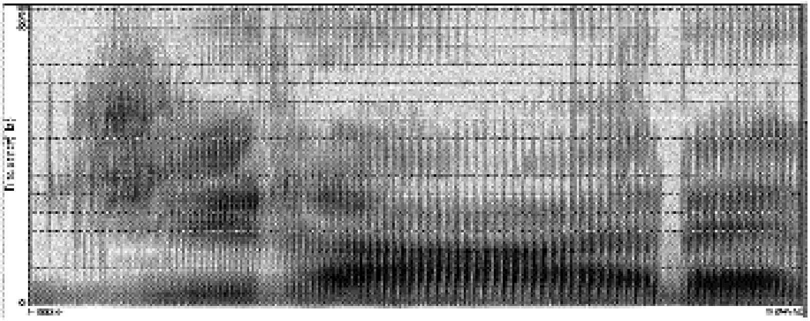

Fig. 3. Diagramma tempo/frequenza della parola italiana 'girare'.

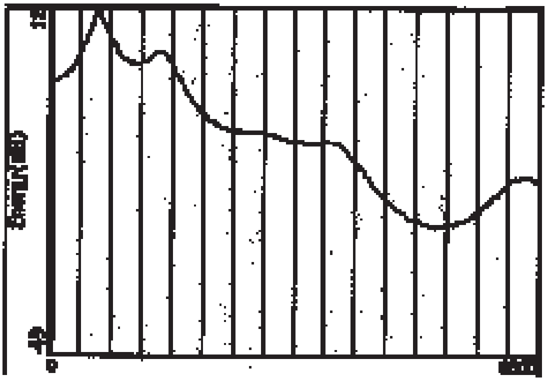

Fig. 4. Diagramma frequenza/energia del punto centrale della vocale [a] indicato dal puntatore nella fig. 3.

4. La percezione del parlato

Il problema oggi in discussione per quanto riguarda la percezione del messaggio fonico è se questa dipenda a) dalla attivazione di un modulo uditivo innato e generale, b) dalla attivazione di un modulo uditivo specifico innato preposto alla percezione di suoni linguistici già categorizzati, c) dalla attivazione di un modulo complesso che porta a riconoscere i suoni a partire dai gesti articolatori necessari per produrli (che qui non considererò per quanto ho detto sopra alla nota 6).

Il problema, che pure avrebbe un risvolto filogenetico, viene in genere studiato osservando il comportamento infantile (a partire dalla fase perinatale), il comportamento di animali, il comportamento di adulti in contatto con lingue diverse dalla lingua madre. Il primo punto di vista è quello che oggi richiama la maggiore attenzione.

Mehler & Dupoux (1992: 215) dichiarano che "oggi prevale l'idea che il sistema di percezione della parola sia innestato su quello dei suoni acustici". Questa posizione si basa sull'assunto di una priorità dell'apparato uditivo generale, a partire dal quale si sviluppa la capacità di percepire e riconoscere il parlato. Sembra infatti accertato che la capacità di categorizzare suoni (linguistici e non) da parte del feto e del neonato si costruisce su esperienze uditive prelinguistiche e quindi biologicamente determinate.

Miller & Eimas (1994: 43, 48) sono invece più cauti e ritengono che non si possa ancora decidere con certezza

> "whether the mechanisms underlying phonetic perception form a system specialized for the processing of speech or whether processing is accomplished by the general auditory system. [Allo stato attuale] the evidence at hand does not provide a compelling case for choosing between an account of underlying mechanism based on specialized processing and an account based on the operation of the general auditory system".

Nusbaum & Goodman (1994: 328) sembrano più vicini a posizioni 'innatiste' che riducono il ruolo dell'udito a quello di selezionare a posteriori i foni più utili partendo da un repertorio universale predeterminato e precategorizzato:

> The infant may be born with few a priori expectations about the way to distribute attention in analyzing the acoustic structure of speech. Experience of listening to speech directs this attention toward those properties that are most relevant for the phonetic and phonological processing of speech in the infant's native language environnement. This would be consistent with the data that suggest that infants are born with a universal phonetic inventory and that perceptual experience eliminates the nonnative contrasts from this inventory, mantaining the native contrasts alone".

Pinker (1994: 256-57), che peraltro non è uno specialista e basa le sue considerazioni su esperimenti di Mehler, Eimas e Jusczyk, non si pronuncia esplicitamente sulla esistenza di un inventario universale innato, ma dà un grande rilievo al tirocinio uditivo prenatale e, in particolare, alla percezione della melodia.

La questione è dunque se l'udito sia un ricettore passivo, che assegna le esperienze sensoriali alle categorie innate preformate e preesistenti, o se invece esso sia attivo e concorra a formare le categorie. Ma anche scegliendo questa seconda ipotesi, che per molti aspetti sembra più forte, la questione rimane complessa non appena si passi dall'osservazione del neonato a quella dell'adulto.

Gli studi di psicoacustica mostrano infatti che l'attività di decodifica del percetto uditivo è basata su due meccanismi: un meccanismo primitivo, innato, basato sulle capacità uditive in sé, e un meccanismo basato su schemi, appreso, idiolinguistico (Bregman 1990: *passim*). Il primo fornisce la base al secondo [17], ma il secondo integra la base sensoriale. Su questo punto tornerò più avanti.

Quindi, malgrado differenze di posizione, la sollecitazione uditiva sembra essere il punto di partenza dello sviluppo delle capacità linguistiche (intese come capacità di codifica e decodifica audioverbali e non necessariamente come capacità cognitive generali) nel bambino.

5. Il caso esemplare dell'intonazione

Il senso di tutte queste osservazioni diventa particolarmente evidente quando esse vengono calate nell'analisi prosodica.

Fino ad anni recenti la linguistica aveva dedicato alla prosodia, cioè all'insieme dei fenomeni ritmici e intonativi delle lingue, un'attenzione minore di quella dedicata non solo alla fonetica e fonologia segmentali ma anche alla morfologia, alla sintassi e alla semantica. Con l'eccezione della scuola britannica, che da tempo aveva sviluppato una tradizione di studi prosodici su base uditiva (cito come esempio rappresentativo Cruttenden 1986), i lavori dedicati a questo settore erano pochi, come si può vedere osservando, ad esempio, le date di pubblicazione dei lavori citati da Bolinger (1989), che aumentano esponenzialmente a partire dagli anni '60; inoltre questi lavori, dedicati per lo più all'inglese, erano basati su osservazioni qualitative, spesso non sistematiche, per lo più concentrate sull'osservazione del rapporto tra alcune funzioni elementari sintattico-pragmatiche (come l'interrogazione) e alcuni profili intonativi. Solo da pochi anni gli studi di prosodia, e in particolare quelli dedicati all'intonazione sono aumentati drasticamente, grazie all'interesse per questi aspetti da parte delle fonologie più recenti (autosegmentali, prosodiche) [18], al punto che ora la situazione si è ribaltata. Oggi chi studia il parlato riconosce alla prosodia un ruolo determinante tanto nella ideazione ed esecuzione della stringa, quanto nella sua segmentazione e interpretazione da parte dell'ascoltatore.

La prosodia è un fenomeno difficile da studiare per molti motivi. Il primo è certamente il fatto che tutte le variabili fisiche che la determinano (il tempo e l'intensità per il ritmo, la frequenza per l'intonazione) si dispongono lungo un *continuum* non segmentato *a priori* (come è invece, grazie alla scrittura, per fonologia, morfologia, lessico e, in parte, sintassi) [19].

Il secondo motivo è che l'intonazione, componente importantissimo della prosodia, non è analizzabile in una prima e seconda articolazione (anche Bertinetto 1981: 27): un segmento intonativo di prosodia ritagliato dal suo contesto è in sé totalmente privo di significato e di funzione, perché i valori che determinano la prosodia, e dunque anche l'intonazione, sono sempre e tutti radicalmente relativi, valutabili e interpretabili solo in rapporto a ciò che segue e a ciò che precede all'interno dell'intera unità prosodica considerata. Quindi, di un segmento vocalico in sé, del quale posso descrivere in modo 'oggettivo' l'altezza, la durata e l'intensità (oltre che il timbro), non posso dire se sia linguisticamente acuto o grave, lungo o breve, intenso o debole. Inoltre, allo stato attuale delle nostre conoscenze, abbiamo ancora difficoltà a segmentare i profili intonativi di molti enunciati naturali, e ad assegnare a una determinata porzione una determinata funzione o un determinato contenuto comunicativo: in molti casi è come se il profilo intonativo e il corrispondente contenuto semantico e pragmatico, pur perfettamente integrati nel sistema comunicativo, si muovessero in una dimensione olistica [20].

Un terzo motivo è la natura particolare della fortissima variabilità prosodica. La variabilità è naturalmente una caratteristica fondamentale di tutte le manifestazioni foniche (e non solo foniche) delle lingue. Ma, mentre la variabilità nella realizzazione dei segmenti che costituiscono la stringa è, in qualche modo e almeno in parte, predicibile e riconducibile a una qualche altra forma di variazione (diatopica, o diastratica, o diafasica che sia), per cui essa non concorre a determinare il contenuto semantico dell'enunciato [21], la variazione prosodica, a parità di condizioni diatopiche, diafasiche e diastratiche, è sempre il riflesso di una sia pur lieve differenza nelle intenzioni comunicative del parlante: ad una variazione prosodica corrisponde sempre una variazione semantico-pragmatica dell'enunciato, perfettamente chiara a chi ascolta [22]. Ciò è detto molto chiaramente in 't Hart *et al*. (1990: 110-114):

> Intonation features have no intrinsic meaning […] We speculate that these choices [cioè quelle fra i vari contorni] are influenced by the attitudinal meaning that a speaker wants to add a literal meaning of his utterances. But the actual encoding of his attitudinal meaning into an individual pitch contour is evidently governed by so many pragmatic and situational factors that we are still looking for a manageable experimental paradigm in which to tackle this complicated issue.

Ma, infine, ciò che fa della prosodia qualche cosa di speciale tra gli strumenti della comunicazione audioverbale è che, mentre una stringa segmentale non può esistere se non dentro uno schema prosodico, uno schema prosodico può esistere senza contenere una sequenza segmentale (come quando mugoliamo a bocca chiusa una melodia [mmmmm]), o può esistere appoggiandosi a una sequenza segmentale artificiale e asemantica conservando una sua capacità comunicativa, come sapeva, p.es., Italo Calvino [23]. Per usare termini antichi, ma molto efficaci, la *phoné* esiste senza la *diálektos*, ma la seconda non può esistere senza la prima (Laspia 1997: 59-69).

La prosodia, oltre a disporre di una sua grammatica (perfettamente nota a chi parla e a chi ascolta, ma ancora poco nota ai linguisti), che consente la trasmissione di un numero grandissimo di sensi a partire da una stessa sequenza segmentale, è anche in grado di dominare tanto la componente sintattica, grazie ai complessi processi della focalizzazione e della messa in rilievo, quanto la componente semantica (la prosodia può far sì che una doppia affermazione neghi) [24].

Lo studio della prosodia si colloca dunque in uno spazio delimitato da un lato dalla psicoacustica e dall'altro dalla rilevanza comunicativa ('t Hart *et al*. 1990: 5).

Come si vede, questa componente di straordinaria importanza nella comunicazione orale: a) si realizza attraverso un meccanismo articolatorio estremamente economico, basato su una variazione di poche decine di Hz rispetto alla frequenza fondamentale propria di ciascuno e determinata dalla anatomia individuale, di una variazione che può anche essere di pochi millisecondi nella durata dei segmenti, e di piccole variazioni nella pressione dell'aria espiratoria a ridosso dei diversi ostacoli glottidali o supraglottidali; questo meccanismo a) è perfettamente congruente con le capacità uditive; b) riflette dinamiche fonatorie generali di ordine probabilmente biologico: la sua articolazione in unità tonali è legata ai cosiddetti 'gruppi espiratori'; le unità tonali si concludono sempre (tranne che nei casi di marcatezza) con una declinazione naturale di tutti gli indici (intensità e frequenza tendono a zero e l'eloquio rallenta), e anche la marcatezza (cioè l'andamento terminale ascendente) non è che il rovesciamento di un andamento naturale e dunque anch'esso non 'arbitrario'.

La questione della segmentazione del discorso in unità tonali e della individuazione del punto focale all'interno di ciascuna unità tonale è forse l'aspetto determinante per individuare la centralità del ruolo dell'udito nella decodifica del parlato. Ciò è confermato anche dall'importanza della prosodia nell'apprendimento linguistico da parte dei bambini e nello sviluppo della loro capacità di segmentare il *continuum* della catena parlata (processo che avviene senza che il bambino abbia ancora alcun controllo consapevole sul meccanismo laringeo, e quando il suo controllo degli schemi prosodici della sua lingua è ancora ad uno stadio aurorale). Hawkins (1999: 195) presenta con grande chiarezza gli aspetti ontogenetici del problema:

> The most global properties of the baby's to-be-native language are prosodic, and some of these may even be learned at or before birth. Rhythm and pitch patterns can be heard in utero, since they are what is left in the signal when the mother's voice reaches the uterus, low-pass filtered through her body. [...] their response patterns were the same regardless of whether the speech was heard unfiltered, or low-pass filtered so that only the prosodic information remained [...] Other evidence suggests that babies are sensitive to the language-specific prosodic properties that cue clause boundaries by about 6 months of age and to those marking phrasal boundaries by 9 months.

Quando gli esperimenti vengono condotti con materiale acustico più complesso (ma ancora artificiale) appare subito che le modalità di discriminazione di altezza sono molto complesse e tengono conto della durata del segnale e anche delle armoniche di frequenza più alta. Quando poi si passa ad osservare la percezione e la discriminazione dei movimenti del *pitch* nel parlato naturale (condizioni sperimentali delicatissime e ancora poco praticate), agli occhi dell'osservatore appare una specie di circolo vizioso: è certamente la percezione uditiva che mette in moto il processo di analisi e di decodifica, ma quasi contemporaneamente il segnale linguistico mette in moto attività cognitive, più alte, di analisi del contesto, del cotesto, delle aspettative dei parlanti ecc., che si sovrappongono alla percezione meramente uditiva e, per così dire, la guidano nella decodifica secondo schemi idiolinguistici. Ciò è vero non solo per il riconoscimento dei segmenti e delle sequenze di segmenti, ma è vero anche per i movimenti del *pitch*: questi, ferme restando le soglie fisiche di durata, altezza e intensità, al di sotto delle quali non si ha percezione o non si ha discriminazione, possono venire trascurati nella decodifica linguistica anche quando siano psicoacusticamente discriminabili.

Questo circolo vizioso è oggi il vero problema teorico degli studi sull'intonazione e non è facile romperlo. Esso consiste, come abbiamo ricordato, nella compresenza e nella cooperazione dei meccanismi di percezione primitiva e di percezione basata su schemi, di cui parla Bregman (1990): le diverse posizioni degli studiosi dipendono in ultima analisi dal fatto che alcuni riducono drasticamente o annullano la percezione primitiva, mentre altri la prendono in considerazione, pur differendo nel dosaggio delle due componenti.

6. Alcune conclusioni

La linguistica può dare un contributo importante alla soluzione di questo nodo che non può essere considerato solo di pertinenza della psicoacustica. Non si deve dimenticare infatti che il segnale di cui ci si occupa è dotato, oltre che di proprietà prosodiche e pragmatiche, anche di proprietà semantiche, lessicali, sintattiche e morfologiche.

Il contributo della linguistica deve partire da alcune semplificazioni. Attualmente la principale difficoltà nello studio della prosodia dipende essenzialmente a) dalla sua intrinseca variabilità e b) da una interrelazione tra variazioni prosodiche e variazioni attitudinali del parlante che si dispongono ambedue lungo un *continuum* non facilmente discretizzabile (o comunque molto meno facilmente discretizzabile di quanto non sia il *continuum* segmentale sul piano fonologico, sul piano morfologico, sul piano sintattico e forse anche sul piano semantico, per il quale si dispone almeno di una unità operazionale come la parola). Mettere in correlazione due variabili continue, relativamente poco note, e cercare di stabilire una gerarchia delle covariazioni è un'impresa disperata senza due operazioni preliminari [25].

La prima è che è necessario assumere, come ipotesi di lavoro, che una delle due variabili possa essere considerata provvisoriamente la variabile indipendente. È ragionevole pensare che questa possa essere una qualche unità del parlato, come il turno nelle interazioni dialogiche, per la quale esistono, a partire dalla teoria degli atti linguistici di Austin e di Searle, le classificazioni elaborate dalla pragmatica: queste, per quanto controverse e, a volte, opinabili, si basano tuttavia, allo stato attuale, su un apparato più complesso e meno incerto di quelle elaborate dalla prosodia: il ruolo provvisorio di variabile indipendente può essere loro assegnato anche in base alla considerazione che le unità pragmatiche possono essere classificate e descritte a prescindere dalla implementazione prosodica, ma il contrario non è vero.

La seconda è che ciascun insieme di enunciati raccolti sotto una determinata etichetta pragmatica deve essere ulteriormente analizzato per mezzo di tutti i sofisticati strumenti di analisi (semantica, sintattica, morfologica, funzionale) di cui dispone la linguistica: questa ulteriore analisi è necessaria perché la prosodia non può essere considerata solo in funzione della pragmatica ma deve essere considerata anche in funzione delle strutture linguistiche.

La paziente e graduale costruzione di queste griglie può fornire lo schema, relativamente robusto, rispetto al quale iniziare, su basi più certe, la classificazione delle regole prosodiche.

Questo programma di lavoro richiede ancora due corollari. Il primo è che ritengo una condizione ormai irrinunciabile che il materiale da analizzare sia costituito da enunciati naturali, estratti da raccolte sistematiche di parlato spontaneo (che nella grande maggioranza dei casi è parlato dialogico), con tutte le difficoltà che ciò può provocare: questo materiale verrebbe ad accompagnarsi a quello già disponibile sul parlato letto o recitato. Il secondo è che nei protocolli di analisi un posto preminente venga assegnato alle verifiche percettive condotte sui parlanti, intesi non come parlanti ideali senza corpo e senza storia, ma come parlanti concreti e definiti in base a caratteristiche esplicite. In questo i linguisti possono guardare ad altri specialisti che considerano importante che gli esperimenti siano espliciti e riproducibili e che i risultati siano verificabili. La convalida del ricevente è importante anche da un punto di vista teorico: nell'interscambio parlato la materia fonica diviene forma semiotica attraverso l'integrazione tra l'agire intenzionale del parlante e la convalida del ricevente. Senza quest'ultima si cadrebbe nel solipsismo.

Se gli schemi prosodici da utilizzare in questo confronto debbano essere rappresentati in forma di *pitch levels* o in forma di *pitch movements* è oggi oggetto di vivace discussione scientifica: il primo modello (la cui codifica formale più nota e diffusa è quella detta ToBI) ha dalla sua il prestigio indiscusso di buona parte della linguistica generativa più recente); il secondo ha dalla sua il prestigio della scuola olandese. Le opinioni sulle capacità esplicative dei due modelli possono essere diverse. Ma per evitare che questa disputa divenga una disputa teologica è necessario passare alle verifiche.

Indirizzo dell'autore: CIRASS - Università di Napoli Federico II, v. Porta di Massa 1, I-80133 Napoli - fealbano@unina.it

Note

[*] Ringrazio Pier Marco Bertinetto che ha letto questo articolo e mi ha suggerito numerosi miglioramenti.

[1] Naturalmente non è sempre così: il tema della fatica del capire è presente in molti lavori di De Mauro (p. es. 1994); in Pinker (1994), si leggono pagine molto efficaci sulla acustica dei suoni (155-163), sulla loro variabilità (173-175), sulle strategie di riconoscimento (175-180) e sul ruolo preminente assegnato all'udito nell'ontogenesi del linguaggio (256-57). Allo stesso modo si osserva che una parte della semiotica è giustamente interessata, *mutatis mutandis*, al ruolo del ricevente e della decodifica (p. es. Eco 1979). Diversa è la situazione quando si osservino lavori specialistici, come vedremo più avanti.

[2] Naturalmente ci sono eccezioni: per la situazione italiana, oltre a Uguzzoni (1990 a, b) si può vedere la rassegna in Albano Leoni (in stampa).

[3] Questa storia sarebbe molto interessante e, per le vicende del Settecento e dell'Ottocento, è stata anche impostata (Gessinger 1994:485-631 e *passim*, Loi Corvetto 1992, 1995, 1998, Dovetto 1998, Albano Leoni & Dovetto 1997, Tani 2000: 104-105, 131-132). È ancora tutta da studiare, da questo punto di vista, la posizione di Paul, di Saussure, degli strutturalisti, della linguistica chomskiana e postchomskiana.

[4] Lo stesso punto di vista è rappresentato ancora in Jakobson & Waugh (1979: *passim*), che costituisce una sorta di ricapitolazione del pensiero e degli studi di Jakobson sull'aspetto fonico delle lingue.

[5] "the perception of speech is tightly linked to the feedback from the speaker's own articulatory movements" (Liberman *et al.*, 1963: 4).

[6] Bertil Malmberg (1967: 168) aveva visto con chiarezza i limiti di questa posizione: "Il est intéressant de constater qu'une nouvelle génération, élevée avec les résultats acoustiques modernes et trop jeune pour connaître l'histoire, plus ancienne, de la phonétique articolatoire et sa fallite avec Meyer, Russel etc,. se sent maintenant tentée, devant cette même complexité des faits qui avait enlevé trente ans plus tôt à l'articulation son rôle comme base de classement primarie, de recourir à la physiologie et au sens musculaire pur expliquer des phénomènes qui, aux yeux de celui qui entreprend son analyse d'un point de vue phonologique, n'impliquent aucun problème". A queste osservazioni si potrebbe aggiungere che questa prospettiva articolatoria, almeno nella sua forma corrente, è problematica per almeno quattro motivi: il peso dato al ruolo del segmento configura un orientamento *bottom-up* molto rigido; il ruolo della prosodia, non riconducibile certo alle sensazioni articolatorie è azzerato; le esperienze uditive perinatali, che certamente precedono ogni capacità articolatoria, non vengono riconosciute; la presenza di deficit articolatori anche gravi non inficia la percezione linguistica. Perplessità esplicite sulla *Mothor Theory* sono anche in Jakobson & Waugh (1979:*passim*).

[7] Il ruolo dell'udito è considerato in un recente libro italiano di psicolinguistica (Orsolini 2000) dedicato all'apprendimento della lingua nei bambini.

[8] Naturalmente qui prescindo del tutto dai problemi e dalle controversie circa la rieducazione dei sordi (Pennisi 1994: 21-93), come anche da quelli sulla equipotenza di codici diversi dall'audioverbale.

[9] Questa storia ontogenetica sembrerebbe riflettere, come si sa, quella filogenetica (Lieberman 1975: 153-165; Lieberman & Blumstein 1988: 205-213): la posizione della laringe nei primati, immediatamente al di sotto della lingua, corrisponde a quella del feto umano e in ambedue i tipi la cavità faringale è ridottissima. Ancora alla storia filogenetica appartiene il fatto che il diametro della laringe

umana rispetto a quello della trachea si è modificato a svantaggio della respirazione e a vantaggio della fonazione: il minor diametro della laringe favorisce il costituirsi di una pressione subglottidale, indispensabile per l'attivazione del meccanismo laringeo. Infine, non si può ignorare il fatto che mentre l'anatomia e la fisiologia dell'apparato uditivo umano sono in fondo molto simili a quelle di altri mammiferi, i rispettivi apparati fonatori sono invece molto diversi. Mi sembra che questi siano tutti indizi del fatto che nel processo di messa a punto degli strumenti per la comunicazione umana, l'interazione tra produzione e ricezione deve essere stata molto forte e che l'assegnazione all'udito di un ruolo subalterno non ha molto fondamento.

[10] La mia incompetenza mi aiuta qui a tenermi alla larga dal problema, forse di impossibile soluzione, se il canale fonico-uditivo sia il canale necessario per lo sviluppo del linguaggio umano o se questo canale avrebbe potuto essere sostituito da altri (Cimatti 1998: 190-195).

[11] Si consideri, a titolo di esempio, che 20 db è la quantità di rumore di fondo presente di notte in una zona rurale.

[12] La capacità di discriminare in durata, anche molto importante, non figura nel campo uditivo, così come esso viene tradizionalmente rappresentato.

[13] Il termine e il concetto, mutuati dagli studi sulla percezione visiva, indicano il fenomeno per cui due stimoli (visivo o uditivo) che esibiscono contenuti spettrali diversi vengono percepiti come uguali.

[14] Prescindo qui tanto dai casi anomali quanto dai casi di voci educate, come quelle dei cantanti.

[15] Sul concetto di '*pitch* virtuale' cfr. Hess (1983: 73).

[16] Questi diagrammi presentano il tempo in ascissa, la frequenza in ordinata e l'energia nel maggiore o minore annerimento del tracciato. Nei casi qui considerati la banda di frequenza va da 0 a 8000 Hz. Le linee orizzontali rappresentano un intervallo di 500 Hz.

[17] "the learned schemas must obviously make use of primitive unlearned abilities (or else there would be nothing from which to build up the learned abilities)" (Bregman 1990: 602-603).

[18] La bibliografia sull'argomento sta crescendo vertiginosamente. 'tHart e altri (1990: 1-6) danno una presentazione vivacemente efficace del problema; Bolinger (1989) è una introduzione insostituibile; Bertinetto (1981) e Bertinetto & Magno Caldognetto (1993), con bibliografia, forniscono il quadro della situazione italiana, per la quale una rassegna bibliografica aggiornata è in Albano Leoni (in stampa).

[19] L'importanza della scrittura per la categorizzazione fonetica segmentale è intuita mirabilmente da Leopardi (in Gensini 1988: 49-54).

[20] In una prospettiva semiologica potrà essere interessante osservare che in prosodia le due facce del segno sono ancora più profondamente integrate che in altri livelli (cfr. a questo proposito Lo Piparo 1991).

[21] La parola *casa* può essere realizzata ['ka:sa], ['ka:za], ['ha:sa], ['kæ:sa], ['ka:s] ecc. senza che questo alteri il suo contenuto e senza che ciò sia indizio di particolari intenzioni comunicative del parlante (se non, in qualche caso, di una scelta stilistica).

[22] "the relation between intonational and attitudinal features may be one-to-many, possibly even many-to-many" ('t Hart *et al*. 1990: 111). Naturalmente questo vale per le variazioni pertinenti e intenzionali (anche se non è sempre facile distinguere tra queste e le variazioni casuali, in mancanza di una categorizzazione fine delle *attitudinal features*).

[23] "Ecchisietevòi, paladino di Francia?" viene sostituito da "Tàtta-tatatài tàtatàta-tatàta…" (Calvino 1959: 10).

[24] In italiano ad una affermazione come "ho mangiato una mela" si può risponde-

re "sì sì" con una intonazione particolare e si esprime in questo modo scetticismo, incredulità e il senso della risposta viene ad essere "non è vero, non ci credo".

[25] Il programma che adombro in queste conclusioni nasce dalle riflessioni di un gruppo di lavoro sul parlato, attivo presso l'università di Napoli, e in particolare dalle discussioni con Renata Savy, Dalia Gamal e Claudia Crocco, che su questa strada stanno lavorando.

Summary

The paper deals with the role of the hearer in speech communication with special regard for intonation.

After a short presentation of the problem and of its recent history in linguistic studies (section 1), a description of the caracteristics of hearing apparatus is presented and attention is paid to the synergy between speech production and speech perception and to the discussion about the role of primitive and inlearned auditory schemas (section 2-4).

In section 5 intonation is taken into account as major and powerfull communication tool, whose importance and efficiency are universally recognised, but whose study and description appear still difficult because of some of its semantics and pragmatics properties and because of the lack of a notation tradition (as letters provide for segmentals strings).

Section 6 provides some conclusions and some suggestions for investigating intonation from a linguistic point of view.

Bibliografia

Akmajian, A., R.A. Demers, A.K. Farmer & R.M. Harnish (1995^4), *Linguistics. An Introduction to Language and Communication*, Cambridge (Mass.), The MIT Press; Ital. transl. il Mulino, 1996.

Albano Leoni, Federico & Francesca M. Dovetto, (1996), "From Maine de Biran to the 'Motor Theory': A Note in the History of Phonetics", *Historiographia Linguistica* 23,3:347-364.

Albano Leoni, F. (in stampa), "Fonetica e fonologia", in Lavinio (in stampa).

Bertinetto, P.M. & E. Magno Caldognetto (1993), "Ritmo e intonazione", in Sobrero (1993: 141-192).

Bertinetto, P.M. (1981), *Le strutture prosodiche della lingua italiana. Accento, quantità, sillaba, giuntura, fondamenti metrici*, Firenze, Accademia della Crusca.

Bolinger, D. (1989), *Intonation and its uses*, London, Arnold.

Bregman, A.S. (1990), *Auditory Scene Analysis. The Perceptual Organization of Sound*, Cambridge (Ma) / London, The MIT Press (paperback 1994).

Calvino, I. (1959^3), *Il cavaliere inesistente*, Torino, Einaudi.

Casula, M.S. *et al.*, eds. (1998), *Linguistica e dialettologia. Studi in memoria di Luigi Rosiello*, Roma, Carocci

Chomsky, N. & M. Halle (1968), *The Sound Pattern of English,* New York, Harper and Row.

CIMATTI, F. (1998), *Mente e linguaggio negli animali. Introduzione alla zoosemiotica cognitiva*, Roma, Carocci.

CORRADI FIUMARA, G. (1990), *The Other Side of Language. A Philosophy of Listening*, London / New York, Routledge.

CRUTTENDEN, A. (1986), *Intonation*, Cambridge, Cambridge University Press.

DE MAURO, T. (1994), “Intelligenti pauca”, in De Mauro (1994: 63-74).

DE MAURO, T. (1994), *Capire le parole*, Roma / Bari, Laterza

DELGUTTE, B. (1997), “Auditory Neural Processing of Speech”, in Hardcastle & Laver (1997:507-538).

DOVETTO, F.M. (1998), “Produzione e ricezione del linguaggio negli studi italiani della seconda metà del Settecento”, *Lingua e Stile* 33:231-266.

ECO, U. (1979), *Lector in fabula. La cooperazione interpretativa nei testi narrativi*, Milano, Bompiani.

FORMIGARI, L. & D. GAMBARARA, eds. (1995), *Historical Roots of Linguistic Theories*, Amsterdam / Philadelphia, Benjamins.

GENSINI, S. (1993[1994]), “Epicureismo e naturalismo nella filosofia del linguaggio fra umanesimo e illuminismo: prime linee di ricerca”, *Annali della Facoltà di Magistero dell’Università di Cagliari*, n.s., 16:55-119.

GENSINI, S., ed. (1988), Giacomo Leopardi, *La varietà delle lingue*, Firenze, La Nuova Italia.

GERNSBACHER, M.A., ed. (1994), *Handbook of Psycholinguistics*, Academic Press, San Diego-London.

GESSINGER, J. (1994), *Auge und Ohr. Studien zur Erforschung der Sprache am Menschen 1700-1850*, Berlin / New York, de Gruyter.

GOLDSMITH, J.A., ed. (1995), *The Handbook of Phonological Theory*, Cambridge (Ma) / Oxford, Blackwell.

GOODMAN, J.C. & H.C. NUSBAUM, eds. (1994), *The Development of Speech Perception*, Cambridge (Ma) / London, The MIT Press.

HARDCASTLE, W.J. & J. LAVER, eds. (1997), *The Handbook of Phonetic Sciences*, Cambridge (Ma) / Oxford , Blackwell.

HAWKINS, S. (1999), “Auditory Capacities and Phonological Development: Animal, Baby and Foreign Listeners”, in Pickett (1999:183-198).

HESS, W. (1983), *Pitch Determination of Speech Signals. Algorithms and Devices*, Berlin / Heidelberg / New York / Tokyo, Springer.

JAKOBSON, R. & L.R. WAUGH (1979), *The Sound Shape of Language*, Ital. transl. il Saggiatore, 1984.

JAKOBSON, R. & M. HALLE (1956), *Fundamentals of Language*, The Hague, Mouton.

KENSTOWICZ, M. (1994), *Phonology in Generative Grammar*, Cambridge (Ma) / Oxford, Blackwell.

LASPIA, P. (1997), *L’articolazione linguistica. Origini biologiche di una metafora,* Roma, Nuova Italia Scientifica..

LAVER, J. (1991), *The Gift of Speech. Papers in the Analysis of Speech and Voice*, Edinburgh, Edinburgh University Press.

LAVINIO, C., ed. (in stampa), *La linguistica italiana alle soglie del 2000 (1987-1998)*, Roma, Bulzoni.

LENNEBERG, E.H. (1967), *Biological Foundations of Language*. With Appendices by Noam Chomsky and Otto Marx, New York, Wiley (repr. Malabar, Fla, Krieger, 1984)

Liberman, A.M. & I.G. Mattingly (1985), "The Motor Theory of Speech Perception Revised, *Cognition*, 21:1-36.

Liberman, A.M. (1996), *Speech: A Special Code*, Cambridge (Ma) / London, The Mit Press.

Liberman, A.M., F. S. Cooper, K. S. Harris & P.F. MacNeilage (1963), "A Motor Theory of Speech Perception", *Proceedings of the Speech Communication Seminar* 2: 1-12, Stockholm, Speech Transmission Laboratory, Royal Institute of Technology.

Lieberman, P. & S.E. Blumstein (1988), *Speech Physiology, Speech Perception and Acoustic Phonetics*, Cambridge/New York, Cambridge University Press.

Lieberman, P. (1975), *On the Origin of Language*, New York, Macmillan (rist. 1987), Ital. transl. Boringhieri, 1980.

Lo Piparo, F. (1991), "Le signe linguistique est-il à deux faces? Saussure et la topologie", *CFS,* 45: 213-221.

Loi Corvetto, I. (1992), "Significati e usi di 'lettre' e 'son' nell'"Encyclopédie", *Lingua e Stile* 27,3: 363-377.

Loi Corvetto, I. (1995), "Il processo della percezione dei suoni nell'*Encyclopédie*", *Lingua e Stile* 30,1:83-93.

Loi Corvetto, I. (1998), "Il processo uditivo in Étienne Bonnot de Condillac", in Casula *et al*. (1998:75-94).

Malmberg, B. (1967), "Réflexions sur les traits distinctifs et le classement des phonèmes, in *To Honor Roman Jakobson* (1967:1247-1251), The Hague, Mouton.

Marchese, M.P., ed. (1995), Ferdinand de Saussure, *Phonétique. Il manoscritto di Harvard Houghton Library bMS Fr 266 (8)*, Padova, Unipress.

McQueen, J.M. & A. Cutler (1997), "Cognitive Processes in Speech Perception", in Hardcastle & Laver (1997: 566-585).

Mehler, J. & E. Dupoux (1992), *Appena nato. Che cosa vede, sente, capisce un bambino sin dai primi giorni di vita*, Milano, Mondadori, French transl. Odile Jacob, 1990.

Miller, J.L. & P.D. Eimas (1994), "Observations on Speech Perception, Its Development, and the Search for a Mechanism", in Goodman & Nusbaum (1994: 37-55).

Moore, B. C. J., ed. (1986), *Frequency Selectivity in Hearing*, London, Academic Press

Moore, B.C.J. (1988^3), *An Introduction to the Psychology of Hearing*, London, Academic Press.

Moore, B.C.J. (1997), "Aspects of Auditory Processing Related to Speech Perception", in Hardcastle & Laver (1997:539-565).

Nespor, M. (1993), *Fonologia*, Bologna, il Mulino.

Nusman, H.C. & J.C. Goodman (1994), "Learning to Hear Speech as Spoken Language", in Goodman & Nusbaum (1994:299-338).

Orsolini, M., ed. (2000), *Il suono delle parole. Percezione e conoscenza del linguaggio nei bambini*, Firenze, La Nuova Italia.

Pennisi, A. (1994), *Le lingue mute*, Roma, NIS.

Pickett, J.M., ed. (1999), *The Acoustics of Speech Communication. Fundamentals, Speech perception Theory, and Technology*, Needham Heights (Ma), Allyn & Bacon.

PINKER, S. (1994), *The Language Instinct*, Ital. transl. Mondadori, 1997.

ROSEN, S. & A. FOURCIN (1986), "Frequency selectivity and the perception of speech", in Moore (1986:373-487).

RYALLS, J. (1996), *A Basic Introduction to Speech Perception*, San Diego / London, Singular Publishing Group.

SAUSSURE, F. de (1922 [1916]), *Cours de linguistique générale*, Paris, Payot, Ital. transl. Laterza, 1968.

SIMONE, R. (1992), "Il corpo del linguaggio. Il paradigma dell'arbitrarietà e il paradigma della sostanza", in ID. (1992: 37-59).

SIMONE, R. (1995^6), *Fondamenti di linguistica*, Roma / Bari, Laterza.

SIMONE, R. (1995a), "The language User in Saussure (and after)", in Formigari & Gambarara, (1995:233-249).

SIMONE, R., ed. (1992), *Il sogno di Saussure*, Roma / Bari, Laterza

SOBRERO, A.M., ed. (1993), *Introduzione all'italiano contemporaneo*, vol. I, *Le strutture*, Bari / Roma, Laterza.

T 'HART, J., R. COLLIER & A. COHEN (1990), *A perceptual study of intonation. An experimental-phonetic approach to speech melody*, Cambridge, CUP.

TANI, I. (2000), *L'albero della mente. Sensi, pensiero, linguaggio in Herder*, Roma, Carocci.

UGUZZONI, A. (1990 a), "Uno sguardo al fenomeno della percezione categorica, *Lingua e stile* 25,1: 119-133.

UGUZZONI, A. (1990 b), "Dagli studi sulla percezione fonica: dati, questioni, proposte", *Studi Italiani di Linguistica Teorica ed Applicata*, 19: 3-21.

ZWICKER, E. (1961), "Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)", *JASA* 33, 2: 248.

The role of rhythmic and distributional cues in speech recognition

Valentina Caniparoli

The purpose of this study is to estimate the role of rhythmic cues (alternation of stressed and unstressed syllables) and of distributional cues (internal, external and of frequency) in the pre-lexical process of connected speech segmentation, through an experiment in artificial language learning.

In this study we analyse an artificial language, consisting in a limited set of meaningless words, created according to the syllabic-distributional and, in one case, metrical features of the Italian language.

Two versions of this artificial language were created, a rhythmic one, whose words present the most frequently occurring rhythmic pattern in Italian (the paroxytone rhythmic pattern), and a long one, whose words are only composed of long syllables.

The experiment, carried out on the rhythmic version (RV) and long version (LV), was has been divided were into two phases: the 'learning phase' and the 'test phase'.

Native speakers of Italian and French have been involved in both phases. In the former phase the subjects listened to the artificial language, consisting in a sequence of words automatically linked together. In this phase the subjects' task was to try to recognize and memorize the words of the artificial language.

In the latter phase, they were asked to take an alternate response test, in which the correct answer had to be chosen from two syllabic sequences. The subjects' task was to choose which of the two belonged to the language they had previously listened to.

In this test the words of the artificial language were compared with 'non-words', that is, with sequences created with the same syllables as those of the artificial language, and representing different parts of the words. In the test phase our purpose was above all to see if the subjects, involved in the experiment on the rhythmic version and on the long version, recognized the words of the artificial language, that is, if they have segmented the speech chain correctly and whether they were more influenced by rhythmic or by distributional cues during the process of segmentation.

1. Introduction

Our study is closely related to the debate concerning the problem of speech segmentation and recognition during a communicative performance.

When the linguistic performance is oral, a basic problem is

(ricevuto nel giugno 2001)

raised: the segmentation of the speech chain in its constituent parts has to occur before the decoding of the message.

It is known, in fact, that speech is arranged in a sequential order, that it consists of a sound continuum and, finally, that it is variable.

Psycholinguistic studies (for two reviews cf. Mattys 1997; Cutler *et al.* 1997) have suggested several models of speech acquisition and recognition. Some of them suppose that the process of recognition starts with a pre-lexical phase in which a preliminary segmentation of the sound continuum occurs, that is, boundaries are put between the words. In this phase, which precedes the actual lexical process, linguistic rhythm and prosodic features in general are considered to be very important, since they act as indicators of boundaries. According to the authors of these models (Grosjean & Gee 1987), listeners make use of rhythmic cues. They pay particular attention to the alternation of stressed and unstressed syllables, preferring the stressed ones and exploiting them in order to segment connected speech, that is to say in order to put boundaries between those units they will have to recognize later according to the linguistic code.

In order to estimate the hypothesis of the influence of rhythmic cues in the pre-lexical process, we carried out (as you will see in detail later) experiments in the learning of so-called ‘artificial’ languages, created according to the syllabic, distributional and metrical features of English (Saffran *et al.* 1996) and French. (Banel *et al.* in preparation). In the light of these experimental studies and using their same criteria, we made our research, whose purpose is to discover the role of rhythm and of syllabic-distributional features during the pre-lexical process of connected speech segmentation in the Italian language.

2. Rhythmic and distributional cues

2.1. Rhythmic cues

Rhythm is a prosodic, or supra-segmental, phenomenon shared by every natural language, consisting in the more or less regular alternation, along the temporal axis, of strong and weak, i.e. stressed and unstressed syllables. This alternation occurs at each level of the prosodic hierarchy (Nespor 1993: 235) and causes the rhythmic pattern.

Its main acoustic correlates are the variations of duration and of intensity which cause the relative prominence of some segments of the speech chain compared to others, whereas the relationships between loudness and fundamental frequency are not yet clear.

Having defined linguistic rhythm as a phenomenon of alternation, it is necessary to recognize rhythmic stress as the regulatory element and emphasize it as being a culminative element, rhythmically distributed, hierarchical and not assimilable (Hayes 1995: 24-26).

The units that rhythmic stress emphasizes are precisely stressed syllables which, as you will see later, are considered as 'stability points' within the speech chain and less subject to variation.

The features of stressed syllables are physical prominence, phonemic stability and perceptive distinctiveness (cf. Mattys 1997: 319). Physical prominence is related to duration, pitch and loudness; phonemic stability means that they are less subject to variation; perceptive distinctiveness implies that stressed syllables are seldom incorrectly recognized compared to unstressed syllables (Browman 1978; Cole & Jakimik 1978, 1980; Bond & Garnes 1980) and more easily recognized than unstressed syllables in the presence of noise (Kozhevnikov & Chistovich 1966) and in those words drawn off from their context of connected speech (Lieberman 1965). Finally, in stressed syllables the recognition of phonemes or groups of phonemes is faster than in unstressed ones (Cutler & Foss 1977; for the Italian language Tabossi *et al.* 1995).

The importance, and the perceptive significance, of stressed syllables has also been confirmed by several studies carried out on English and French concerning the role of these syllables in the process of segmentation, and of the introduction of word boundaries. A metrical segmentation strategy (MSS, Metrical Segmentation Strategy; cf. Cutler 1990) has in fact been proved productive in these languages.

In English, rhythm is determined by the alternation of strong and weak positions, i.e. stressed and unstressed syllables, whose main acoustic correlates are the variations of loudness and of fundamental frequency. Although the position of the stress is free, in the sense that it has a distinctive value from the lexical point of view, listeners make use of a metrical segmentation strategy in relation to strong syllables, putting a boundary immediately before them. Evidence is given by the following examples. A study by Cutler & Carter (1987) carried out on the "Corpus of English Conversation", a corpus of spontaneous speech drawn from a sample of conversations in British English, showed that most lexical words are composed of strong monosyllables (59,40%), and polysyllables containing the strong syllable at the beginning (28,20%). In this same context, the study by Cutler & Butterfield (1992) showed that, in those cases where the perception of spontaneous speech was difficult, listeners tended to put a boundary before a strong syllable and to take it away before a weak one (e.g.: case I: *by loose analogy* > *by Luce and Allergy*; case II: *how big is it* > *how bigoted*). Finally, Cutler & Norris (1988) discovered that the recognition of a monosyllabic target word, which is meaningless (e.g. *mint*) in a bisyllabic sequence, is faster in the presence of a strong-weak rhythmic pattern (e.g. *min-tef*) than in that of a strong-strong one (*min-tayf*). The reason for this is that, in the former case, listeners perceived the sequence as one word (*mintef*), whereas in the latter case they perceived it as the union of two monosyllabic words, exactly because they put a boundary between the two strong syllables (*min* and *tayf*).

With regard to the French language, rhythmic cues are determined by the alternation of long and short syllables, as in French the main acoustic correlates of stress are the lengthening of duration and the falling movement of the fundamental frequency in the final syllable of words. Unlike English and Italian, where stress has a distinctive lexical value, in French it has a fixed position within the word, that is, on the final syllable. Some studies by Banel & Bacri (1994, 1997) showed that French listeners use a metrical segmentation strategy according to the final syllable of the word, that is, they put a boundary immediately after it. In other words, in order to segment connected speech correctly they take advantage of the high predictability of French stress.

What is left to see is whether listeners make use of a metrical segmentation strategy also in Italian, which is a language with a distinctive lexical stress and whose acoustic and perceptive correlates are the prevalent variation of duration and of loudness. A study by Mancini & Voghera (1994), carried out on the "Lessico di frequenza dell'Italiano Parlato" (LIP, De Mauro *et al.* 1993), showed that in Italian the prevalent rhythmic pattern of bisyllabic and trisyllabic words (occurring for 31,06% and 17,60% respectively in the whole corpus) is paroxytone (93,3% of bisyllabic and 81,1% of trisyllabic words). Further research into this subject has not been performed; the purpose of the study which is presented below is therefore to see if the paroxytone rhythmic pattern is perceived as the most familiar and if it can make the process of segmentation easier.

2.2. Distributional cues

The distributional cues of a given language are characterized by three types of feature: internal, external and of frequency.

The 'internal distributional features' relate to the degree of correlation between groups of sounds within syllables [1]; an example is given by the high frequency of the CV syllabic type (57,71%) compared to the CCVCC syllabic type (0,002%) in spoken Italian (Mancini & Voghera 1994:70).

The 'external distributional features' relate to the probability of concatenation of given syllables, one following another. For instance, in Italian there is a higher probability that the syllable *so*, followed by the syllable *no*, forms a word (the form *sono* '(they) are', in speech, has a frequency equal to 2224) rather than when followed by the syllable *ro*, with which it does not form a word in Italian. Therefore, while in the first case a word is formed, in the second a lexical boundary should be put between the two syllables.

The 'frequency of occurrence' [2] of syllables can be related to the familiarity the listener has with the syllables themselves.

According to Hayes & Clark (1970), external distributional features must be considered as the most important. In fact, during the learning phase of a language, which can simply consist in an initial listening to it, the ability to segment and recognize words can come from their observation. A high correlation between one syllable and the following is connected with an internal position within a given word, whereas a low correlation is connected with a boundary. As regards other types of distributional cues, as already said, they are language-specific and not directly observable by a listener who approaches a foreign language for the first time.

3. Experiment of artificial language learning

As previously mentioned, experiments in artificial language learning have been carried out. The languages have been created according to the syllabic, distributional and metrical features of English (Saffran *et al.* 1996) and of French (Banel *et al.* in preparation), but before describing the experiments, it is necessary to give a definition of artificial language and explain its functionality.

The artificial language of our experiment consists of a limited set of meaningless words, created by linking automatically together syllables uttered in isolation and constant in their fundamental frequency, loudness and duration.

This language was created in order to face the need to remove all semantic cues, so that conditions preliminary to language acquisition could be re-created in adult people.

The experiments were divided into two phases. The first one, the learning phase, consisted in linking artificial words automatically together and in putting them in a random order. The words were repeated as many times as they could in a speech sequence approximately 15 minutes long, during which the subjects listened to the 'new language' and had to try to recognize and memorize its words. The second phase consisted in an alternate response test, in which the subjects had to choose which syllabic sequence, within a pair, was a word of the language they had listened to, or was more similar to that word.

The basic purpose of this kind of experiment is to discover if adult people, whose task is to learn a new language, when lacking semantic cues, use more the distributional features or the prosodic ones in order to recognize where one word ends and another begins.

Here it is a brief summary of the results of the previous experiments. As far as the learning of the 'English' artificial language (Saffran *et al.* 1996) is concerned, it was discovered that listeners productively used the distributional features (internal, external and of frequency) in connected speech segmentation and in artificial word learning.

As regards the learning of the 'French' artificial language (Banel *et al.* in preparation), it was discovered, confirming what we expected, that the listeners systematically used the long, i.e. stressed, syllables in order to put boundaries and therefore to recognize the words of the artificial language. Nevertheless, it is important to underline that they scarcely exploited the internal distributional features and those relative to frequency, since the syllabic type (CVC) and the syllables chosen for the creation of the artificial language are rarely used in French. The authors chose such syllables because, in French, most CV syllables are meaningful words.

3.1. Materials and methods

The artificial language of our experiment, as of the previous ones, was created according to the most frequent syllabic, distributional (internal, external and of frequency) and in one case, metrical features in spoken Italian (cf. Mancini & Voghera 1994).

The syllabic corpus used for the experiment was composed of 18 syllables, CV type, which represented the stressed open syllables of the most frequent lexical forms in spoken Italian. Each syllable, uttered in isolation by a male voice and constant in its pitch and loudness, was created in a long and short version, whose duration were equal to those of a stressed and an unstressed syllable. Only variations in duration were performed as it has been proved that this parameter is the most reliable acoustic correlate of Italian stress both in production and in perception (Bertinetto 1981).

As regards stressed and unstressed syllables, we chose the following values, drawn from a sample of speech uttered by a regional TV news reader: 180 ms, average duration of unstressed open syllables; 250 ms, average duration of stressed open syllables (Passaro 2000).

Using the above syllabic corpus, 8 'artificial words' were created, 2 trisyllabic and 6 bisyllabic; this division was made in order to respect the relation, occurring in spoken Italian, between syllabic length and frequency of occurrence of a word.

The words were created, by linking syllables automatically together, in two versions: a rhythmic one, in which the words presented an alternation of stressed and unstressed, i.e. long and short syllables; and a non-rhythmic one, in which they lacked this alternation, that is, they were only composed of long syllables. The paroxytone rhythmic pattern was chosen both for trisyllabic words (short-long-short syllable) and for bisyllabic ones (long-short syllable) since, as already said, it is the most frequent in spoken Italian.

For each word of the artificial language, such as *faledo*, 4 types of 'non-words' were created, using the 18 syllables of the corpus.

The first type was made up of syllables that never followed one another in the words of the artificial language (e.g. *radove*). The second type was made up of parts of two different words that could follow one another, for example, the final syllable of a word was linked with the initial syllable of the following word (e.g. *dokele*). The third type consisted in the union between the initial part of a word and the initial part of another (e.g. *faleke*). The fourth type consisted in the union between the final part of a word and the final part of another (e.g. *soledo*).

Like the words, also the non-words are meaningless and have been created in two versions, rhythmic (paroxytone rhythmic pattern) and non-rhythmic.

In table 1 all the words and non-words of the artificial language are listed clearly.

Table 1

WORDS		**NON-WORDS**			
		Type 1	**Type 2**	**Type 3**	**Type 4**
1	**faledo**	*radove*	**dokele**	faleke	*soledo*
2	**rovela**	*kelada*	**ladave**	rovesi	*divela*
3	**dako**	*roko*	**kora**	dafa	*loko*
4	**kepi**	*fapi*	**pide**	kede	*dopi*
5	**tʃadi**	*ledi*	**disi**	tʃaro	*bedi*
6	**sibe**	*debe*	**befa**	sida	*labe*
7	**deso**	*tʃaso*	**soro**	dera	*piso*
8	**ralo**	*silo*	**loʃa**	raʃa	*kolo*

The experiment of artificial language learning was carried out in two versions: the former with neutral metrical pattern (LV = Long Version), that is, with no rhythmic pattern and composed of words, each only containing long syllables, the latter with the Italian metrical pattern (RV = Rhythmic Version), i.e. composed of the artificial words in the paroxytone rhythmic version. Each version was divided into the 'learning phase', in which groups of subjects listened to the language, consisting in a sequence of words (automatically linked together, each repeated 160 times and put in a random order), and had to recognize and memorize them. In the 'test phase' the subjects answered an alternate response test, in which they were required to choose which of the two sequences was a word of the language they listened to or was more similar to that word. In this last phase, the 'words' were compared with the corresponding 4 types of 'non-words', and the 'non-words' among themselves. In this phase these criteria were followed, first of all, in order to see if the subjects were able to recognize the words of the language, secondly, to see, in case of errors, with which type of non-words they confused the words. Finally, in choosing between the non-words, it was noted when they preferred a particular type. The main purpose of the alternate response test was to try to understand how the subjects segmented the speech continuum of the artificial language and, once the words were recognized, which part of them, initial or final, they memorized more easily. This explains the sense of the non-words. As a matter of fact, if the subjects prefer the first type, this could mean that they made a random choice, since this type is an impossible sequence, i.e. it could never occur in the learning phase. If they easily choose the second type, this could mean a wrong segmentation of the acoustic signal which links with one or other parts of the following words. If they choose the third or fourth one, this could show on which part of the word (initial in the former type, final in the latter) the attention of the listener mostly focused.

The experiment was carried out on 24 native speakers of Italian, equally distributed between men and women and with a high-medium level of education. They were divided into 2 experimental groups, each containing 12 subjects. One group went through the two phases of the experiment on the Long Version (LV), the other through the two phases of the experiment on the Rhythmic Version (RV). After this, in order to estimate the validity of the experiment, we considered it useful to involve 20 native speakers [3] of French in the learning of the Italian artificial language. They were equally distributed between the LV and the RV experimental group. This comparison was made in order to estimate possible differences in the attitude of French speakers towards an artificial language whose features are different from those of their own language.

Nevertheless, it is important to underline that there were differences in the amount of cues the subjects had at their disposal. Above all, the native speakers of Italian could use every kind of distributional cues, internal, external and of frequency, since the syllables of the artificial language were chosen from the most frequent in Italian. Moreover, the subjects involved in the rhythmic version could also use the metrical cue typical of the most frequent rhythmic pattern in Italian, such as the paroxytone (Mancini & Voghera 1994). The native speakers of French, on the contrary, could only exploit the external distributional cues, and only some of them could also use the Italian rhythmic cues, to which then they were little accustomed.

3.2. Results

The response of the subjects to the experiment were divided according to their native language (Italian and French) and to the version of the artificial language, whose learning they went through (Rhythmic Version and Long Version, i.e. with neutral metrical pattern). After this, for each group we estimated in percentage terms how many times the words of the language were recognized correctly. Then we estimated the errors, that is, how many times the words were not recognized and with which types of the corresponding non–words they were confused. As regards the data concerning exclusively the non–words, they were divided according to how many times the non–words were compared with each other within the lists of items and then their values in percentage terms were estimated. Finally, for each experiment, on the results of the two versions, we carried out statistical estimations on averages, according to the *two-way t Student's test*, in order to find out possible significant differences between the responses of the two groups (RV and LV) due to the different experimental conditions (rhythmic or non - rhythmic version learning).

4. Analysis of the results and conclusions

4.1. Learning and its estimation in Italian subjects

The analysis of the data showed that both experimental groups learnt the words of the artificial language in a significantly higher percentage than chance. The group of the rhythmic version (RV) recognized the words in 80% of cases, the group of the non-rhythmic or long version (LV) in 74% of cases. Nevertheless, even if in the RV group the percentage of recognized words is slightly higher than in the LV group, such a difference must be considered as statistically insignificant (*stat t* –1,44; *critical t* 2,36).

As far as the estimation of errors is concerned, neither RV nor LV subjects chose a particular type of non–word. As for the pairs of non–words, the RV group did not prefer a particular type of non–word; whereas the LV group mostly chose the second and third type, but also in this case the differences between the responses of the two groups are not statistically significant.

In table 2 and 3 the results obtained are given in percentage.

Table 2. Recognised words in percentage (%)

RV Group Rhythmic Version Learning	LV Group Long Version Learning
80	74

Table 3. Compared non-words

(%)	Non-words Type 1	Non-words Type 2	Non-words Type 3	Non-words Type 4
RV Group	48	49	51	52
LV Group	47	56	56	42

4.2. Learning and its estimation in French subjects

The analysis of the data showed that the attitude of French subjects towards an artificial language completely differs from that of the Italians.

In fact, we estimated that in both groups the percentage of subjects who learnt the artificial words is totally random: in the rhythmic version group (RV), the recognition of the words occurred in 55% of cases, in the long version group (LV) in 64% of cases. However, though these percentages are totally random, their difference is statistically significant (*stat t* –2,65; *critical t* 2,36). This means that the subjects involved in the rhythmic version did not recognize and memorize the words, on the contrary, they had many difficulties when segmenting the artificial language, created according to Italian syllabic, distributional and metrical features. On the contrary, even if its percentage of recognized words is not very high (64%), the LV group, i.e. the non-rhythmic version group, significantly differs from the RV group.

As far as the estimation of errors is concerned, the RV subjects mostly showed preference for the second type of non-word (61%), that is, for the one that links the final syllable of a word with the initial syllable of another. The LV group uniformly distributed the errors between each type of non-words, when these were combined with the words during the test phase. From the comparison between the number of times in which the RV group (61%) and the LV group (40%) confused the second type of non-words with the words, a value resulted, which is very close to the significance (*stat t* –2,22; *critical t* 2,36) which could be caused by the different experimental conditions.

In table 4 and 5 the results obtained are given in percentage.

Table 4. Recognised words in percentage (%)

RV Group Rhythmic Version Learning	LV Group Long Version Learning
55	64

Table 5. Compared non-words

(%)	Non-words Type 1	Non-words Type 2	Non-words Type 3	Non-words Type 4
RV Group	46	61	47	46
LV Group	46	50	53	52

These tables show again that, unlike the LV group, the RV one preferred the second type of non–word. Nevertheless, from the statistical comparison of the data, no significant difference in the choice made by the two groups between the types of non–words resulted.

4.3. Conclusions

In total, from the results obtained and the statistical estimations performed, no particularly significant difference between the two versions of the experiment, carried out on Italian listeners, can be deduced. The listeners involved in the rhythmic version, in fact, segmented the connected speech correctly and recognized the words in 80% of cases. Therefore, we can say that they satisfyingly carried out their task, using all the cues they had at their disposal.

On the other hand, the absence of rhythmic cues did not lessen or create difficulties for the subjects involved in the learning of the long version, i.e. with neutral metrical pattern. In fact also these listeners learnt the artificial language, that is, segmented and recognized the words correctly.

This homogeneous behaviour of the two groups (RV and LV) can be explained by considering the quantity of cues they had in common, that is, the distributional cues (internal, external and of frequency). As previously said, the native speakers of Italian in both groups (RV and LV) could make use of the internal, external distributional features and of those relative to frequency of the artificial language, exactly because these were created according to those of Italian.

Both syllabic type (CV) and syllables probably had resulted very familiar to the listeners since the beginning of learning, as they were chosen from the most frequent in spoken Italian. Besides, during the first phase the listeners were able to determine the probability of concatenation of syllables, one following another, also thanks to the limited number of words of the language (8 words) and to their high frequency of occurrence in the learning phase (each word occurred

160 times). Finally, they may have also been helped by the duration chosen for the syllables, which was determined according to the values of duration of the stressed (250 ms) and unstressed syllables (180 ms) of spoken Italian, as the difference between the two values is not so great.

As a matter of fact, the subjects involved in the rhythmic version probably easily recognized the rhythmic pattern of the language, which corresponded to the most frequent in spoken Italian, and used it in order to segment the connected speech. The alternation of paroxytone bisyllables and paroxytone trisyllables, which partly broke the rhythmic regularity of the language, did not make learning difficult for them.

On the other hand, the subjects involved in the non-rhythmic version who listened to words only composed of long syllables, evidently perceived a list of isolated syllables only for an initial short period. Then they felt the need to link groups of syllables together, perhaps also giving this sequence a more familiar rhythmic pattern, since the duration of the syllables was not such as to make possible the sole recognition of monosyllabic words.

Therefore, rhythmic cues, even if they increased the word recognition rates, were not the determining factor in the process of connected speech segmentation, compared to distributional cues. Besides, from the choice of the types of non-word, i.e. syllabic sequences reproducing different parts of the words of the language, we were not able to determine which part of the word the subjects memorized more easily and used in the process of segmentation.

As already said for the Italian groups, in order to explain the results obtained in the French ones (RV and LV), we considered the amount of cues the listeners had at their disposal and the one they had in common. The cues they shared were only the external distributional ones, the other two (internal and of frequency), in fact, though they were present in the language, did not result productive in the French listeners. Moreover, only one group could also exploit the rhythmic cues typical of the Italian language (the paroxytone rhythmic pattern).

The analysis of the results also allowed us to discover that in French listeners rhythmic cues could even cause a worsening of the results. As a matter of fact, not only did they fail to recognize the words of the artificial language, but during the test phase they mostly showed preference for the second type of non-word, that is, for the one which links the end of a word with the beginning of the following. This phenomenon of wrong segmentation shows how much the subjects were influenced, during the pre-lexical process of speech segmentation, i.e. in the introduction of boundaries, by the rhythmic pattern of their own language, where stress is fixed and boundary is expected to occur soon after a stressed syllable. As regards distributional cues, the two French experimental groups could only make use of the external distributional ones, that is, of the possibility to determine the probability of concatenation of the following syllables during the learning phase. The other two distributional cues (internal and frequency), in fact, though they were present in the language, were not productive in French subjects since they have familiarity neither with the sounds typical of Italian syllables considered nor with their frequency of occurrence.

In conclusion, we discovered that in French groups the role of distributional cues was not significant, whereas the presence of rhythmic cues did not make the process of segmentation easier and even hindered the process of language learning, since the subjects were strongly influenced by their own strategy of metrical segmentation and upset by the rhythmic irregularity of the artificial language.

As far as the behaviour of Italian subjects is concerned, we cannot surely say that in this context they applied a strategy of metrical segmentation based on the prominence of the last but one syllable. First of all, in this experiment the influence of distributional cues, which in part annulled the efficacy of rhythmic cues, resulted determining. Secondly, the distinctive lexical function of Italian stress must be emphasized. As a matter of fact, even if in spoken language a rhythmic pattern is more frequent than others, (i.e. the paroxytone used in this experiment) the position of the stress cannot always be predicted. This happens in French where it has an horistic function, i.e. a function of delimitation, (Trubeckoj 1939). In conclusion, our study confirmed the hypothesis that rhythmic cues start up language-specific mechanisms. It also showed that Italian listeners partially use these cues during the process of segmentation as they are accustomed to variable rhythmic patterns, whereas French listeners make use of rhythmic cues, in any context, according to the strategy of metrical segmentation productive in their own language.

Address of the author: CIRASS - Università degli Studi di Napoli "Federico II", Via Porta di Massa 1, 80133 Napoli - e-mail: caniparoli@cirass.unina.it

Notes

[1] In this context we will only talk about the correlations between groups of sounds within syllables, but it is necessary to underline that such a degree of correlation can also be found in units longer than syllables, such as morphemes and words.

[2] As regards spoken Italian, research has been made into the frequency of occurrence of words in speech, but not of syllables (for writing see Batinti 1993). On the contrary, in this context we considered it useful to carry out, as it will be shown later, a study on the most frequent syllables in spoken Italian, from its most frequent words.

[3] The experiment on French subjects was possible thanks to the collaboration with Professor Uli H. Frauenfelder, Professor of Psycholinguistics at the Faculty of Psychology and Science of Education of the University of Geneva. We would like to thank him for the help given.

References

ALTMANN, Gerry T.M., ed. (1990), *Cognitive models of speech processing: Psycholinguistic and computational perspectives*, Cambridge, MA: MIT Press.

BANEL, Marie-H. & Nicole BACRI (1994), "On metrical patterns and lexical parsing in French", *Speech Communication* 15: 115-126.

BANEL, Marie-H. & Nicole BACRI (1997), "Reconnaissance de la parole et indices de segmentation métriques et phonotactiques", *L'Année Psychologique* 97: 77-112.

BANEL, Marie-H., Uli H. FRAUENFELDER & Pierre PERRUCHET (in preparation), "Contribution des indices métriques à l'apprentissage d'un langage artificiel".

BATINTI, Antonio (1993), *Il sistema fonologico dell'italiano*, Perugia, Guerra.

BERTINETTO, Pier Marco (1981), *Strutture prosodiche dell'italiano*, Firenze, Accademia della Crusca.

BOND, Z. S. & Sara GARNES (1980), "Misperception of fluent speech", in Cole (1980: 115-132).

BROWNAM, Catherine P. (1978), "Tip of the tongue and slip of the ear: implications for language processing", *UCLA Working papers in fonetics* 42.

COLE, Ronald A. & Jola JAKIMIK (1978), "Understanding speech: how words are heard", in Underwood (1978:67-116).

COLE, Ronald A. & Jola JAKIMIK (1980), "How are syllables used to recognize words?", *Journal of Acoustical Society of America* 67: 965-970.

COLE, Ronald A., ed. (1980), *Perception and production of fluent speech*, Hillsdale, NJ: Erlbaum.

CUTLER, Anne & Donald J. FOSS (1977), "On the role of sentence stress in sentence processing", *Language and Speech* 20: 1-10.

CUTLER, Anne & D. M. CARTER (1987), "The predominance of strong syllables in the English vocabulary", *Computer Speech and Language* 2: 133-142.

CUTLER, Anne & Dennis G. NORRIS (1988), "The role of strong syllables in segmentation for lexical access", *Journal of Experimental Psychology: Human Perception and Performance* 14: 113-121.

CUTLER, Anne & Sally BUTTERFIELD (1992), "Rhythmic cues to speech segmentation: Evidence from juncture misperception", *Journal of Memory and Language* 31: 218-236.

CUTLER, Anne (1990), "Exploiting prosodic probabilities in speech segmentation", in Altmann (1990: 105-121).

CUTLER, Anne, Delphine DAHAN & Wilma VAN DONSELAAR (1997), "Prosody in the Comprehension of Spoken Language: A Literature Review", *Language and Speech* 40, 2: 141-201.

DE MAURO, Tullio, Federico MANCINI, Massimo VEDOVELLI & Miriam VOGHERA (1993), *Lessico di frequenza dell'italiano parlato*, Milano, Etaslibri.

GROSJEAN, François & James P. GEE (1987), "Prosodic structure and spoken word recognition", *Cognition* 25: 135-155.

HAYES, Bruce (1995), *Metrical Stress Theory*, Chicago, The University of Chicago Press.

HAYES, John R. & Herbert H. CLARK (1970), "Experiments in the segmentation of an artificial speech analog", in Hayes (1970), 221-234.

HAYES, John R., ed. (1970), *Cognition and the development of language*, New York, Wiley.

KOZHEVNIKOV, Valerij A. & Ludmilla A. CHISTOVICH (1966), *Speech: Articulation and perception*, Washington, D.C., No. Joint Publication Research Service: 30543.

LIEBERMAN, Philip (1965), "On the acoustic basis of perception of stress by linguists", *Word* 21: 40-54.

MANCINI, Federico & Miriam VOGHERA (1994), "Lunghezza, tipi di sillabe e accento in italiano", *Archivio Glottologico Italiano* 79,1: 51-77.

MATTYS, Sven L. (1997), "The use of time during lexical processing and segmentation: A rewiev", *Psychonomic Bullettin & Rewiev* 4: 310-329.

NESPOR, Marina (1993), *Fonologia*, Bologna, il Mulino.

PASSARO, Gianluca (2000), "Stabilità dei nuclei sillabici in sillaba aperta ed in sillaba chiusa", *Atti del XXVIII Convegno Nazionale AIA*, Trani: 295-298.

SAFFRAN, Jenny R., Elissa L. NEWPORT & Richard N. ASLIN (1996), "Word segmentation: the role of distributional cues", *The Journal of Memory and Language* 35: 606-621.

TABOSSI, Patrizia, Cristina BURANI & Donia SCOTT (1995), "Word identification in fluent speech", *Journal of Memory and Language* 34: 440-467.

TRUBECKOJ, Nikolaj S. (1939), *Grundzüge der Phonologie,* Travaux du Cercle Linguistique de Prague VII; Ital. transl. Einaudi, 1971.

UNDERWOOD, G., ed. (1978), *Strategies of information processing*, London, Academic Press.

The role of literacy in the recognition of phonological units

Olga M. Manfrellotti

In this study, we analysed the relation existing between literacy and the clear representation of specific phonological structures during tasks requiring competence of phonological processes. The first question to deal with is whether the division of the spoken *continuum* into discrete units depends on the mastery of an alphabetic writing system. The second question is whether this knowledge is the display of a linguistic, or rather metalinguistic, competence, which allows us to consider language and to analyse uttered words as phonemes.

The survey was based on two experiments, respectively inquiring into phonemes and syllables in 12 sequences of laboratory speech and 12 sequences of natural speech. The survey involved two groups of Italian-speaking adults: 16 students/workers attending an evening school, experiencing great difficulty in reading and writing tasks, and 16 graduates, whose professions required both ability in and frequent use of orthographical competence.

We considered it interesting to make a qualitative analysis on this basis. Only the test for phonemic surveying was considered. Our aim was to understand if the different target phonemes analysed were equally recognized or if probably different acoustic and articulatory features made it more or less easy to recognize some phonemes rather than others.

Therefore, it is clear how important it is, during a survey, to consider different variables. These variables are extralinguistic conditions, such as literacy and illiteracy and, much more important, linguistic, phonological and lexical conditions, such as the two different kinds of material, or the consonant and vowel targets.

1. Introduction

The aim of this study is to make a contribution, through an experimental survey, to the discussion on the relation existing between literacy, on one side, and perception and recognition of specific phonological structures, on the other. The question to deal with is if the possession of abilities, such as reading and writing, has a role in the process of phonological segmentation and categorization of speech.

Dividing the spoken *continuum* into discrete units and recognizing these units are process of human mind, having no counterpart in physical reality. The question psycholinguistics has long asked is if

(ricevuto nel giugno 2001)

this process is based, in part at least, on the mastery of an alphabetic writing system. The second question is whether this knowledge is the display of a linguistic, or rather metalinguistic, competence, which allows us to consider language and to analyse uttered words as phonemes.

We cannot completely share the methodology that assumes the speaker's phonological competence *a priori*, on the basis of explicit judgement, which is exclusively required to subjects having orthographical competences. [1]

The discussion on the subject was developed most of all in psycholinguistics. There are three lines of research in this discipline:

1. analysis on the comparison between literate and illiterate adults in phonological tasks (Morais *et al*. 1979).

2. analysis on readers of non-alphabetic writing (Read *et al*. 1986).

3. analysis on the comparison between children having normal reading abilities and dyslexic children (Laenderl *et al*. 1996).

Results coming from the three studies are coherent beyond any doubt. Illiterate people and readers of non-alphabetic writing present a low percentage of correct responses in phonological tasks. Dyslexic children (considering that dyslexia is an illness involving reading and text understanding abilities) show a high percentage of correct responses in tasks of phonemic manipulation, in which a phonological representation, not influenced by orthography, is essential. There could arise the doubt that consciousness of phonemes is not developed spontaneously, but that it is a part of the speaker's 'linguistic baggage' which derives from the experience gained through literacy.

There is another aspect to stress in order to complete a bibliographical survey. In fact, it would be a great mistake to consider phonological knowledge as something homogeneous, because numerous and irrefutable data (Morais *et al*. 1986; Treiman & Zukowsky 1991; Eimas 1999) indicate a reality on "dissociated" levels. If the absence of reading and writing education does not seem even to allow the development of an analysis into phonetical units, this does not occur for units such as rhymes and syllables. In the absence of different evidence, such an indication could be explained by the fact that, in oral cultures, poets are however capable of using assonance, alliteration and rhyme relations in their poetical works and they can build up rhythmic sequences based on the number of syllables within the line.

Therefore the hypothesis worth considering is that the basic code, used in the strategies of segmentation and of speech perception and in the representation of lexical items, is the syllabic one. Further strategies, such as the one based on phonetic units, are ‘restricted’ and acquired later.

A starting point, for the Italian language, whithin the project dealing mostly, if not exclusively, with other languages, is the research by Albano Leoni *et al.* (1997). Results obtained in the phonemic survey from the two groups of people analysed (semi-illiterate/literate), on three kinds of material (natural/laboratory/structured), show a relative difficulty of completion for everyone, but a significant difference for literate people under laboratory conditions.

The hypothesis presented is that there are two active competences in communication: the linguistic competence and the metalinguistic one, the latter concerning the way in which we observe and consider language. This competence would not be available for all speakers in the same way, but defined by different experiences, among which there is contact with the learning of orthography. The string of discrete elements would be a part of this competence and it would be available only for literate people. This study is a constant point of reference for the research here presented, concerning both methodology and the choice of materials for the test.

2. Materials and methods

The data on which this study was conducted come from the cross analysis of two different tests, phoneme monitoring and syllable monitoring.

For both tasks we used the same phonic sequences, belonging to two distinct and well-defined groups:

24 sequences of natural material (12 experimental sequences and 12 fillers) taken from a recording of regional TV news. They are meaningful sentences, but characterised by a sort of prosodic, syntactic and often semantic incompleteness. Moreover, a necessary condition for the choice of materials was the single occurrence of the target event (phoneme or syllable) within every sequence.

Two list are given (respectively with phonemic and syllabic targets in bold characters and in phonemic transcription) with the 12 experimental sequences. In fact, fillers had no influence either on statistical analysis or on qualitative analysis.

Table 1

	A. Natural material	**A. Natural material**
nm1	…non accetterà **/p/**atti con i fuoriusciti	…non accetterà /**pa**/tti con i fuoriusciti
nm2	…basato su due grandi **/p/**oli	…basato su due grandi /**po**/li
nm3	…per ora ritenuto a**/d/**atto allo scopo	…per ora ritenuto a/**da**/tto allo scopo
nm4	…rinviato a **/d/**opo il venti giugno	…rinviato a /**do**/po il venti giugno
nm5	…vorrebbe tagliare **/s/**ubito il cordone ombelicale	…vorrebbe tagliare /**su**/bito il cordone ombelicale
nm6	…accade da più di un **/s/**ecolo	…accade da più di un /**se**/colo
nm7	…sembra cogliere la **/v/**oglia di andare ad uno scontro	…sembra cogliere la /**vo**/glia di andare ad uno scontro
nm8	…in cambio di posti di la**/v/**oro	…in cambio di posti di la**/vo/**ro
nm9	…anche questo ha portato alla sconf**/i/**tta	…anche questo ha portato alla scon**/fi/**tta
nm10	…per entrare a reg**/i/**me	…per entrare a re/**dʒi**/me
nm11	…fino ai pattisti di S**/e/**gni	…fino ai pattisti di /**se**/gni
nm12	…la tornata di dom**/e/**nica scorsa	…la tornata di do/**me**/nica scorsa

24 sequences of laboratory material made up of words read in acoustically controlled conditions and built up as the experiment requires.

The following list are given here (we simply refer to the experimental sequences):

Table 2

	B. Laboratory material	**B. Laboratory material**
lm1	Ha mangiato del **/p/**ane a tavola	Ha mangiato del **/pa/**ne a tavola
lm2	La notizia è nella seconda **/p/**agina del giornale	La notizia è nella seconda **/pa/**gina del giornale
lm3	Puoi avere fi**/d/**ucia in me	Puoi avere fi**/du/**cia in me
lm4	Silvia è una **/d/**onna meravigliosa	Silvia è una **/do/**nna meravigliosa
lm5	C'è grande incertezza al con**/s/**iglio comunale di Napoli	C'è grande incertezza al con**/si/**glio comunale di Napoli
lm6	Lavorano fuori **/s/**ede a Milano	Lavorano fuori **/se/**de a Milano
lm7	È stata una **/v/**era delusione	È stata una **/ve/**ra delusione
lm8	I fantini trattano i ca**/v/**alli con molta dolcezza	I fantini trattano i ca**/va/**lli con molta dolcezza
lm9	L'avvocato deve garant**/i/**re per Paolo	L'avvocato deve garan**/ti/**re per Paolo
lm10	Non può sal**/i/**re le scale	Non può sa**/li/**re le scale
lm11	Non mi hai più dato la ric**/e/**tta	Non mi hai più dato la ri**/tʃe/**tta
lm12	La notizia l'ha data il minist**/e/**ro oggi	La notizia l'ha data il minis**/tʃe/**ro oggi

The test was submitted to two groups each of 16 people. The first group was made up of 10 men and 6 women, aged between 20 and 55. They were students of a state evening school in the suburbs of Naples, attending the school for a course leading to the school-leaving certificate of the Italian middle school. They had all obtained the elementary school-leaving certificate, but the time elapsing between the current time and the last time they went to school ranged from 10 to 40 years according to their age. In spite of this, people belonging to this first group cannot be defined illiterate. They have, however, a great difficulty in reading and writing. In fact, we made certain from their teachers' comments, that they were the weakest from an educational point of view.

The second group was made up of 8 men and 8 women, aged between 23 and 50. The sixteen subjects were all graduates and worked as teachers, or university and medical researchers requiring ability and frequent use of reading and writing skills for their jobs.

The two groups of people, who were first tested on phoneme monitoring and then, after a week, on syllable monitoring, performed the following test.

Sitting in front of a computer, the person hears a bip warning him/her of the beginning of the experiment. After a pause of 2000/ms, a message arrives, informing the person of the target event (phoneme or syllable), then another pause of 500/ms and finally the phonic sequence. As soon as the person recognizes the target, the person has to push the spacebar on the computer keyboard, the sequence stops and another one immediately begins as before.

3. Results

3.1. Interrelation among variables

The conditions of literacy/semi-illiteracy relative to subjects, the laboratory/natural kind of materials, the phoneme/syllable recognition for targets are to be considered three independent variables, that is to say, controlled by the experimenter. Reaction times and errors are, instead, to be considered two dependent variables. The values of the variables are not established beforehand, but depend on the test performed by each subject.

Therefore, we will consider the interaction among variables from the more general to the more detailed level and we will firstly refer to reaction times and then to the number of errors.

After an initial analysis, the mean recognition times concerning phonemes and syllables shows a higher rapidity in the recognition of syllables when compared to phonemes. The difference of 244/ms in reaction times for the two tests is in fact significant from a statistical point of view:

Table 3

	Phonemes	**Syllables**
Half-illiterate/literate	1626 ms.	1382 ms.

More precisely, in the condition of semi-illiteracy, the improvement from phoneme recognition to syllable recognition is much more evident, with a difference of 300 ms (approximately), when compared with the 200 ms (approximately) used by graduates. Moreover, literate subjects are quicker in recognising both phonemes and syllables, but with a more significant advantage in the first test.

Table 4

	Literate	**Semi-illiterate**
Phonemes	1569	1683
Syllables	1373	1391

Concerning the different nature of the materials used, we could note that in phoneme monitoring, involving exclusively the laboratory material, there occurred a significant improvement of literate subjects when compared with semi-illiterate subjects.

Concerning errors analysis, we noted, in general, a better performance in recognising syllables compared with phonemes, both by graduates and by students/workers. Students/workers reveal, however, a more remarkable improvement when compared to graduates, together with reaction times. We also noted a significant difference between laboratory conditions and natural conditions as for errors, with an advantage for laboratory conditions in the phoneme monitoring test for literate subjects.

The following conclusions can be drawn from the above data. If we suppose in the speaker's mind two ways of lexical completion, a phonic icon with a low resolution (cfr. Albano Leoni *et al*. 1997) and what we can define an orthographical lexicon, the prototype which can be drawn presents a two-way relation between these two systems. We want to confirm that interference could only be possible if people have an orthographical lexical structure acquired by the learning of reading and writing. A crucial point is to establish when this interaction could occur.

The hypothesis supported in this study is that the appropriate form of lexical representation is not the primary step in decoding the acoustic message. The recognition of natural speech would occur prevalently because of cognitive resources available for each speaker (syntax, semantics, context). When these resources are not sufficient, in cases of misunderstanding, speakers turn to metalinguistic resources. According to what has been stated up to now, these resources cannot be activated for everybody in the same way. Speakers who are able to use them because they have a good orthographical competence, will turn to the support given by the phonemic code (in the case of use of alphabetical writing). People who cannot use them, instead, will use a syllabic code that, as we have already said, seems more easily accessible for all subjects. Perception and recognition of phonemes and syllables can be seen, in this way, in terms of a balanced interaction between linguistic and metalinguistic activities. On the contrary, given the difficulty of establishing a two-way relation among acoustic manifestations of natural speech and the appropriate representation of linguistic units, it would be difficult even to imagine these linguistic units as directly and immediately involved in the understanding of natural speech.

The results stressed seem to lead to the same direction. It should be emphasized that the general difficulty in recognizing phonemes when compared with syllables for both groups of people confirms the greater naturalness of syllables in metalinguistic tasks (Morais *et al*. 1986; Treiman & Zukowsky 1991; Eimas 1999). The emphasis on the gap between the two groups in the availability of a metalanguage, which becomes more and more perfect in proportion to the level of literacy, can be explained by two considerations. The former is the remarkable improvement of illiterate people during the passage from the phonemic recognition task to the syllabic one when compared with literate people. The latter consideration is the greater difference presented in phoneme monitoring rather than syllable between graduates and students/workers. In the phoneme test, which is common to this study and to that by Albano Leoni *et al.* (1997), an interesting agreement of the data is to be noted. In this test we can observe a significant improvement for literate people in laboratory conditions. An equally significant improvement is not to be seen for semi-illiterates, neither in the natural condition of the same task nor, for both kinds of materials, in the syllabic recognition, relative to the most recent test only.

These data would indicate a strong activation of the cognitive burden for both groups of people in recognizing natural speech, which would not leave any space for the activation of the phonemic code. In laboratory speech where the required cognitive burden is smaller, there would be the possibility to use a phonemic representation, even if only for the subjects able to use it. If the task requires a syllabic recognition, differences are smaller, even where the cognitive burden for the recognition of the stimulus is less urgent because the use of the (syllabic) code itself is possible for all subjects.

3.2. Qualitative analysis

A qualitative analysis was also made on the target phonemes of experimental materials. The intention was to understand if, and to what extent, some specific phonemes were recognized better than others, and if the fact that a word belongs to a certain part of the discourse rather than to another can have influenced the recognition of the phonemes. To give an example, a table is shown, in which, for each word containing a target, the average of reaction times and the total number of errors are reported without any distinction concerning either subjects or kind of materials, but with distinction of their grammatical class. It did not seem appropriate to make further considerations because the number of elements of the four classes is not perfectly balanced in order to allow a significant analysis.

Table 5

Names	Times	Mist.	Adj.	Times	Mist.	Verbs	Times	Mist.	Adverbs	Times	Mist.
pane	1496,8	4									
pagina	1471,5	6									
fiducia	1559	12									
donna	1522,7	8									
consiglio	1866,10										
sede	2115	4									
			vera	1505,9	7						
cavalli	1461,13										
						garantire	1471,9	14			
						salire	1517,9	13			
ricetta	1148,5	15									
ministero	1425,6	13									
patti	1636,4	7									
poli	1405,1	8									
			adatto	1491	13						
									dopo	1674	9
									subito	2160	9
secolo	2111	6									
voglia	1483	6									
lavoro	978	28									
sconfitta	1344,11	12									
regime	1315,611	19									
Segni	1278,625										
Domenica	1378	24									

Concerning the kind of phonemes, we kept apart the two conditions of literacy vs. illiteracy of the subjects and two different groups were formed.

In the former case, the six kinds of phonemes are separated whereas the variable relative to the kind of material is unified. Below are the tables of reaction times and errors made concerning the two groups considered:

Table 6. Semi-illiterate

	times		**errors**
p	1526,049	**p**	23
d	1577,29	**d**	33
s	2092,825	**s**	24
v	1458,08	**v**	39
i	1434,524	**i**	43
e	1359,769	**e**	51

Table 7. Literate

	times		**errors**
p	1473,726	**p**	2
d	1558,782	**d**	9
s	2055,254	**s**	5
v	1454,286	**v**	15
i	1390,551	**i**	15
e	1276,095	**e**	22

The first difference we want to underline is between consonants and vowels.

Concerning vowels, reaction times are much lower when compared to consonants, both for semi-illiterate and for literate subjects. This does not mean automatically that vowels are recognized better than consonants because, in the case of errors, we observe a completely opposite situation.

For both groups, in fact, vowels are also the targets on which the greatest number of errors concentrates. A trade-off is to be observed in the recognition of vowels, according to which the subjects are quicker, but pay less attention to their task. The contrary happens for consonants.

Concerning consonants, among the four analysed, the consonant

/v/ seems to show a situation similar to that of vowels: a low reaction time and a high number of errors, both for literate and semi-illiterate people. However, a deeper analysis on single target sequences with their relative stimuli casts new light on the situation noted above.

Below, a table is shown, summarizing times and errors of each sequence for both groups of people.

Table 8

		Illiterate		**Literate**	
		Averages	**Errors**	**Averages**	**Errors**
LM1	**p**	1532,583	4	1461,188	0
LM2	**p**	1517,6	6	1425,5	0
LM3	**d**	1611,333	10	1506,714	2
LM4	**d**	1611,333	7	1434,133	1
LM5	**s**	1822	9	1910,867	1
LM6	**s**	2192,167	4	2037,938	0
LM7	**v**	1552,444	7	1459,375	0
LM8	**v**	1449,75	12	1474,2	1
LM9	**i**	1565	11	1378,923	3
LM10	**i**	1586,2	11	1449,786	2
LM11	**e**	1076	12	1221,154	3
LM12	**e**	1610,833	10	1240,538	3
NM1	**p**	1742,444	7	1530,375	0
NM2	**p**	1331,9	6	1478,429	2
NM3	**d**	1350	10	1632	3
NM4	**d**	1662,6	6	1685,462	3
NM5	**s**	2183,2	6	2136,846	3
NM6	**s**	2074,636	5	2147,4	1
NM7	**v**	1486,3	6	1481,438	0
NM8	**v**	909	14	1047	14
NM9	**i**	1315	9	1373,231	3
NM10	**i**	1291	12	1340,222	7
NM11	**e**	1236	13	1321,25	8
NM12	**e**	xxxx	16	1378	8

The sequence nm8 (natural material), containing one of the four stimuli /v/, is not recognised 14 times out of 16 by both groups of subjects. On the contrary, in case of recognition, the reaction times are very low. It is evident that this failed recognition is a feature linked to the stimulus and not to the kind of target. This assumption is supported by the fact that the second element /v/ of the couple of natural material (nm7) is recognized at its best by literate subjects and with few errors compared to the average of illiterate people. Through a deeper analysis of the acoustic spectrum, the target /v/ in nm8 in the word *lavoro* ‘job’ is performed more as a labiodental approximant and, for this reason, much closer to vowels in its phonic substance. [2]

The question concerning the other three kinds of consonants is different. The unvoiced occlusive shows reaction times which are lower when compared to the voiced occlusive and the unvoiced fricative, but differently from what happened in the comparison between the two groups of consonants and of vowels, it shows a clear advantage of recognition even when errors are concerned.

In the latter case, phonemes are grouped without any discrimination between the presence or absence of voice sonority. Three groups of phonemes (occlusive-fricative-vowels) were taken into consideration. The reaction times and number of errors of these phonemes remain separate according to the nature of material (natural *vs.* laboratory). Two tables are presented below, summarizing respectively the reaction times and errors of semi-illiterate and literate subjects.

Table 9. Semi-illiterate

		Times			**Errors**
p/d	lm	1560,459	**p/d**	lm	27
	nm	1535,057		nm	29
s/v	lm	1838,469	**s/v**	lm	32
	nm	1858,606		nm	31
i/e	lm	1486,25	**i/e**	lm	44
	nm	1291,214		nm	50

Table 10. Literate

		Times			**Errors**
p/d	lm	1455,623	**p/d**	lm	3
	nm	1576,982		nm	8
s/v	lm	1721,5	**s/v**	lm	2
	nm	1864,935		nm	18
i/e	lm	1325	**i/e**	lm	11
	nm	1355,474		nm	26

The tendency is one-way both for natural material and for laboratory material: short times and remarkable errors for the subgroup i~e; a marked advantage of recognition for the subgroup p~d when compared to the subgroup s~v both concerning reaction times and errors. Finally, we want to stress that the only detectable difference between literate and semi-illiterate subjects during the recognition test is a better performance of the former group on laboratory material in all the three subgroups of phonemes. These results are in compliance with the tendency noted in the statistical analysis discussed in the preceding paragraph.

So, there are two conclusions to draw.

1. Concerning the different behaviour of consonants and vowels in the recognition strategy, the following hypothesis is proposed. The poor attention paid to vowels by the subjects, which is evident from the number of errors, could depend on the fact that vowels physically act as a support for consonants. If they are taken off from a sequence, they would not cause particular difficulties in decoding the sequence itself. So, paying greater attention to the surveying of consonants would be part of the recognition mechanism. Besides, vocalic tones tend to be confused much more than specific phonetic features of consonants, making the identification of vowels themselves more difficult to perform. Vowels, in fact, are often exposed to reduction, in spite of the phonetic context in which they are found. It rarely happens, instead, that events such as a shift towards fricative or voiced consonants occur outside specific sequences of phonemes.

2. Concerning the recognition of consonants, it was found that the unvoiced occlusive is the more easily recognized phoneme. If it is true that the addition of phonetic features increases the markedness of an element, the advantage on reaction times and errors concerning /p/ seems coherent with the situation that shows the unmarked parts as more natural. These parts are, in fact, more frequent, more spread within the various phonological systems in the world and the first to be acquired in linguistic development (the occlusive /p/ in fact lacks voice).

The preceding reflections would need to be, in any case, further verified during experiments that take into consideration a greater quantity of data and allow also a statistical analysis of the materials under examination.

Address of the author:

CIRASS - Università degli Studi di Napoli “Federico II”, Via Porta di Massa 1 - 80133 Napoli - manfrellotti@cirass.unina.it

Note

[1] For example, M. Nespor (1993:20-ff.): "…Concentrandosi sull'aspetto fisico del suono, la fonetica si distingue perciò dalla fonologia che…si concentra sull'aspetto mentale, cioè sul sistema che governa la competenza fonologica del parlante nativo. Se per esempio chiediamo a un parlante nativo dell'italiano qual è la composizione sonora di una parola come vento, molto probabilmente dirà che questa parola contiene cinque suoni, tre consonantici e due vocalici…"

[2] The non-fricative nature of the 'phone' [v] in the context VCV was taken into consideration, for example, in a work by Rispoli & Savy published in 1993. In this work, the 'phone' is defined, according to specific spectro-acoustic features (such as the lack of noise and the formantic structure), as a labiodental approximant.

Bibliographical References

ALBANO LEONI Federico, Francesco CUTUGNO & Alessandro LAUDANNA (1997), "L'attivazione di rappresentazioni fonetiche durante il riconos cimento del parlato: una risorsa metalinguistica?", in Benincà e altri (1997:35-52)

BENINCÀ Paola e altri, eds. (1997), *Fonologia e morfologia dell'italiano e dei dialetti d'Italia,* Roma, Bulzoni

BRADY S. A. & D. P. SHANWEILER, eds. (1991), *Phonological Process in Literacy,* Hillsdale, Lawrence Erlbaum

EIMAS Peter D. (1999), "Segmental and syllabic representation in the perception of speech by young infants", *Journal of Acoustical Society of America* 105:1901-1911

LAENDERL Karin, Uta FRITH & Heinz WIMMER (1996), "Intrusion of orthographic knowledge on phoneme awareness: Strong in normal readers, weack in dyslexic readers", *Applied Psycholinguistics* 17:1-14

MORAIS Josè, Paul BERTELSON, Luz CARY & Jesus ALEGRIA (1986), "Literacy training and speech segmentation", *Cognition* 24:45-64

MORAIS Josè, Luz CARY, Jesus ALEGRIA & Paul BERTELSON (1979), "Does awareness of speech as a sequence of phones arise spontaneously?", *Cognition* 7:323-331

NESPOR Marina (1993), *Fonologia*, Bologna, Il Mulino

PERETTI A.& Franco FERRERO, eds. (1993), *Atti del XXI Convegno Nazionale*, Padova, A.G.P.

READ Charles, Yun-Fei ZHANG, Hong-Yin NIE & Bao-Quing DING (1986), "The ability to manipulate speech sounds depends on knowing alphabetic writing", *Cognition* 24:31

RISPOLI Gianpiero & Renata SAVY (1993), "Alcune considerazioni spettroacustiche sulla [v] italiana", in Peretti A., Ferrero F. (1993:91-94)

TREIMAN Rebecca & Andrea ZUKOWSKY "Levels of Phonological Awareness", in Brady S. A. & Shanweiler D. P (1991:67-83)

Polysp: a polysystemic, phonetically-rich approach to speech understanding

Sarah Hawkins & Rachel Smith

We outline an approach to speech understanding, Polysp (for POLYsystemic SPeech understanding) that combines a richly-structured, polysystemic linguistic model derived from Firthian prosodic analysis and declarative phonology, with psychological and neuropsychological approaches to the organization of sensory experience into knowledge. We propose that the type of approach exemplified by Polysp promises a fruitful way of conceptualising how meaning is understood from spoken utterances, partly by ascribing an important role to all kinds of systematic fine phonetic detail available in the physical speech signal and by rejecting assumptions that the physical signal is analysed as early as possible into abstract linguistic units. Polysp provides a framework by which episodic multimodal sensory experience of speech can be simultaneously processed into different types of linguistic and non-linguistic knowledge at a variety of levels of abstraction, with the emphasis always on understanding meaning in order to interact with another person rather than on building a complete description of a given utterance at successive, obligatory stages of formal linguistic analysis. We discuss phonetic data consistent with these views.

1. Introduction

This paper explores the contribution of phonetic knowledge to how we understand words, and some implications for what makes a plausible model of spoken word understanding. We show that certain types of fine phonetic detail systematically reflect not just the phonemic content but the wider phonological and grammatical structure of the message; that while some systematic differences in phonetic fine detail are relatively localised in the speech signal, others stretch over several syllables; and that both types can make speech easier to understand. We make the case that one consequence of neglecting fine phonetic detail in models of spoken word recognition and understanding is that other processes and stages of analysis may be given inappropriate emphasis, and that this has happened in models which adopt the convenient fiction that the phoneme is the basic input unit to the lexicon. In consequence, no current phonetic or psycholinguistic theory accounts satisfactorily for how normal connected speech is understood.

Rivista di Linguistica, 13.1 (2001), p. 99-189 *(ricevuto nel giugno 2001)*

We turn then to discuss central properties needed to model the process of speech understanding. First, we propose a model, derived from Firthian prosodic analysis and declarative phonology, that provides a clear linguistic-phonetic structure onto which incoming sensory speech information might be mapped. Then we set out critical characteristics of human listeners that may define the way they make sense of richly informative sensory signals. Amongst these, we emphasize various forms of learning, and some current neuropsychological views about the nature of memory and the organisation of mental categories, both linguistic and non-linguistic. We suggest that phonetic categories are like all other mental categories: self-organising (emerging from the distribution of incoming sensory information in combination with pre-existing relevant knowledge), multimodal and distributed within the brain, dynamic, and context-sensitive (or relational) and therefore plastic, or labile. These two threads, linguistic and neuropsychological/neurophysiological, together form the theoretical approach we name Polysp. We conclude by discussing central properties needed to model how normal speech is understood, identifying at each stage an existing (usually computational) model which includes properties that we consider promising.

2. The dominance of the phoneme in models of speech perception and spoken word recognition

2.1. Overview

It is well known that there are interdependencies between grammatical, prosodic and segmental parameters in speech. Yet the linguistic concept of the phoneme as the basic unit of sound contrast has dominated the thinking of phoneticians and psychologists over the last 50 years to such an extent that it is central to most theories. Either the phoneme (or an ill-defined 'phonetic segment' that is treated functionally as a phoneme) is taken as axiomatic, or it is effectively given a crucial role even while acknowledging that it has limitations and that other units may be more fundamental. Thus, phonetic models of speech perception usually focus on how phonemes are distinguished from one another, while psychological models of spoken word recognition use the phoneme as a crucial unit in the early stages of the process of recognition, and indeed usually as the input unit. One consequence is that phoneticians, speech scientists and psycholinguists all tend to assume that the sensory signal is transformed into an abstract form early in the process of word recognition, and certainly before lexical access.

Phoneme labels are valuable for certain types of linguistic and phonetic description, but we question their value in modelling speech understanding under normal conditions, as will become clear below. However, the purpose of this section is not to reiterate arguments against the phoneme as a necessary unit of speech perception (see for example Marslen-Wilson & Warren 1994; Hawkins 1995; Coleman 1998; Nguyen & Hawkins 1999; Warren 1999:169*ff*; Hawkins & Nguyen in press) but to argue that over-emphasis on phonemes has deflected attention from the presence of other types of linguistic information in the speech signal and, in consequence, distorted the relative importance of various processes in models of how speech is understood.

2.2. Information conveyed by unstructured strings of phonemes or allophones

The phoneme is an abstraction without physical reality. Its close relative, the allophone, can be loosely interpreted as having physical reality, but as is well known, a phoneme string cannot be uniquely related to a string of lexical items, and allophones cannot be related to the right phonemes, nor even to the right phoneme slots, independently of the linguistic structure in which they occur. For example, the phoneme string /katsaɪz/ signals *cat's eyes* (or *cats' eyes*) and *cat size* but in natural speech is distinguished by a number of durational and spectral differences predictable from the linguistic structure. The /s/ will be relatively longer when it is the onset of the second syllable and, depending on the accent, there can be other differences such as in the quality and degree of diphthongization of the second nucleus (in *eyes* / *size*). These allophonic differences presumably help the listener find the right meaning.

Similarly, in Standard Southern British English (SSBE), *Carter Knight* and *car tonight* have identical phoneme sequences, and in this case they can be pronounced very similarly. However, in at least one London accent, they are usually differentiated by the pattern of glottal stops: *Carter Knight* would be [kɑʔənaɪʔ] whereas *car tonight* would be [kɑtənaɪʔ], because in this accent glottal stops cannot substitute for [t] word-initially. All the glottal stops in this pair of phrases happen to be allophones of /t/, but glottal stops do not always signal /t/ in this accent. For instance, the phrase *hand it over*, said as [handɪtəʊvə] or [handɪʔəʊvə] in SSBE, also has (at least) two glottal stops in the London accent, [ʔandɪʔaʊvə]. But while the second one still signals /t/, the first is an allophone of neither /t/ nor /h/, since this accent lacks /h/; instead, it signifies (optionally) that the vowel is utterance-initial. In other words, allophones like glottal stop do not map simply onto phonemes, so there is still room for confusion unless a larger structure than the phoneme-sized segment is used to access meaning.

Similar arguments and illustrations can be made for the way an utterance's phonetic structure can indicate its grammatical structure. For example, in English and many other languages, the phoneme structure of function words is much more limited than that of content words, and function words are subject to rather different connected speech processes with surrounding words. For example, whereas many English function words begin with /ð/, no content words do, and word-initial /ð/ conditions distinctive connected speech processes after word-final /n/ that are not found for other /n/-fricative sequences. Thus *ban that*, phonemically /banðat/, is commonly pronounced [ban̪ːat] in SSBE and other accents of English, but other /n/-fricative sequences must retain their fricative, although they may lose or almost lose the /n/ completely as long as the preceding vowel is nasalized. So *ban thatch* can be [ban̪θatʃ] or [bãⁿ̪θatʃ] or similar, and *ban zips* can be [banzɪps], [bãⁿzɪps] or even [bãzɪps]; but neither would be understood if, following the rules for /nð/ in function words, they were [ban̪ːatʃ] and [banːɪps] respectively.

Note that these transcriptions oversimplify, and in particular do not do justice to the subtleties of timing, vowel quality, and other variation that tend to co-occur with the more obvious allophonic differences, creating a coherent sound that systematically reflects much of the linguistic structure of the intended utterance. Examples of such subtle systematic variation can be found in Section 3.2 below, while Kelly & Local (1989), Manuel *et al.* (1992), Manuel (1995), and Ogden (1999) offer more detailed treatments of particular examples.

What these examples do illustrate is that, by itself, an allophone string is effectively about as abstract and uninformative as a phoneme string unless given an explicit context which, for connected speech, includes grammar as well as syllable and word structure. That being so, although replacing phoneme strings with allophone strings appears to add desirable phonetic detail to the input to models of word recognition, it is unlikely to solve the problems introduced by the use of phoneme strings. What is needed is an input to the lexical identification process that preserves all the linguistic information inherent in the speech signal, and an output that is also sensitive to the linguistic structure—lexical and grammatical—so that speech can be understood correctly as soon as it is heard. Although experiments show that both meanings of homophonic single words can be simultaneously activated even when an appropriate context is provided (Swinney 1979), it seems unlikely that this will standardly happen in a normal spoken exchange. For example, it is unlikely that the apparent phoneme structure of [ban̪ːat] means it is understood first as *ban Nat* and only later corrected to *ban that*, or even that both meanings are simultaneously activated, for the sorts of phrases we are talking about are not homophonic if the fine phonetic detail is attended to. In other words, we suggest that systematic differences in fine phonetic detail provide one sort of disambiguating context that constrains which meanings are accessed, much as has been demonstrated in a number of experiments for other types of context (see Simpson 1994 for a review). Moreover, because fine phonetic detail often signals grammatical structure, it co-occurs with non-phonetic types of disambiguating context, and the two together can presumably provide even stronger constraints on which meaning is accessed. Information of this type has not normally been considered the domain of standard phonetic theories of speech perception, nor of the phonetic input to most psycholinguistic models of lexical access.

2.3. Consequences for phonetic and psycholinguistic models of a focus on phonemes

The focus of early phonetic and word recognition research on abstract, idealised and unstructured linguistic units like phonemes as the primary unit of perception is understandable and indeed defensible, but it has had at least two biasing consequences on the development of theory. It has encouraged (1) ‘short-domainism’ and (2) the introduction of separate and often arbitrary processes to explain how speech can be understood despite the impoverished information provided by unstructured phoneme strings.

‘Short-domainism’ is exemplified by much phonetic speech perception research, which has typically focussed on simple sound contrasts in highly controlled phonetic environments, if only to keep the size of investigations manageable. Experiments exploring the perceptual correlates of consonants, for example, often use only one vowel, while those on vowels normally restrict the consonantal context. There is surprisingly little literature on the perception (or production) of unstressed syllables, and relatively few experiments examine the perception of phoneme identity in more complex environments—

when they do, results are often quite different (e.g. Kewley-Port & Zheng 1999). Most research on prosody (apart from stress) is conducted in isolation from that on segmental identity, and results of experiments that examine influences on segmental perception of domains of more than one or two syllables have tended not to be integrated into the most influential theories of perception. They may even be seen as part of some other process, such as vocal-tract normalisation (e.g. Ladefoged & Broadbent 1957) or how we recognise words as opposed to phonemes (e.g. Warren 1970, 1984).

Thus, word recognition and sound recognition (of features or phonemes) are often axiomatically distinct processes in many of the most influential theories arising from the various branches of speech science. For example, the Motor Theory (Liberman & Mattingly 1985) distinguishes between acoustic-phonetic trading relations, phonetic context effects, and 'higher-order' effects of lexical knowledge (cf. Repp 1982; pers. comm. 1989), although they can alternatively be seen as due to the same underlying process of using multiple cues to make complex decisions, though often operating on different temporal domains. As the above discussion of information conveyed by allophones suggests, we maintain that sound recognition and word recognition should not be axiomatically separate, because knowing about words and larger structures makes it easier to interpret allophonic detail and *vice versa*. (This position provides a viable context for explaining the classical intelligibility experiment by Pickett & Pollack 1963). It will become clear later in this paper that, for us, the appeal of direct realist theories of speech perception (Fowler 1986; Best 1994, 1995) would gain significantly if the focus were on identifying all units of linguistic structure from all aspects of the highly complex vocal tract behaviour for speech, rather than on recognition of individual phonemes from particular details of movement.

The second problem of a principal focus on phonemes is exemplified by much of the debate on psychological models of spoken word recognition over the last 20 or more years, for they typically introduce a linguistically impoverished phoneme-like input to the lexicon. Even theoreticians who acknowledge the difficulties of identifying phonetic units of speech perception nevertheless usually do most of their actual work with discrete, abstract units corresponding to phonemes (or to distinctive features that have clear temporal boundaries and are grouped into units that correspond to phonemes), if only because to do otherwise is so complicated that it would effectively mean abandoning their own work and moving into another discipline.

Unfortunately, emphasis on an unstructured phoneme string as the basis for lexical activation or search brings with it the implicit assumptions that the input is structured into discrete units capable of distinguishing lexical meaning but is linguistically incomplete in other respects. Theoretical debate thus focuses on such issues as whether various hypothesized processes are autonomous or interactive (that is, on whether an earlier process is affected by feedback from later processes), and on whether inhibitory as well as excitatory processes are necessary to produce the desired output, as theorists struggle to build in enough supplementary sources of information to compensate for the impoverished information available from an unstructured phoneme string or its like. Comparison of the architecture and performance of computational models will not necessarily offer a resolution to these issues, since the same effects can result from quite different models, depending on the detailed assumptions and definitions adopted (e.g. Norris 1992). However, for a recent example of this type of debate, see Norris' *et al.* (2000) discussion of their model Merge, and the accompanying commentaries.

Merge, a development of Shortlist (Norris 1994), provides an interesting example of how early abstraction of linear strings of phonetic units demands invocation of particular, equally abstract processes to compensate for this simplification. Merge is presented as a model of phonemic decision making. It is unusual in having phonemes represented in two places and subject to quite different processes. The motivation for this is to distinguish initial sensory information (called prelexical information) from decisions about phoneme identity that are made from all the available information, lexical and sensory/prelexical. Sensory information is coded in terms of probabilities, with the probability for each prelexical phoneme being independent of that of the others, while the phoneme decision stage is categorial and subject to influence from other phoneme decisions.

The properties of the prelexical phoneme stage are introduced to preserve information about the sensory input so that a wrong decision about one phoneme need not jeopardise word recognition—in other words, to allow error correction. But neither simple feature bundles nor phonemes reflect all the systematic linguistic information available in the speech signal that could facilitate error correction or even avoid the need for it in the first place. Additionally, to account for certain experimental findings such as 'compensation for coarticulation', the prelexical stage includes abstract knowledge of transitional phoneme probabilities. Other types of linguistic knowledge are represented later, lexically or post-lexically. Here we have an example of invocation of a distinction between types of abstract knowledge that is forced on the model because it assumes phonemic prelexical representations.[1]

Although psycholinguists normally represent higher-order structure as linguistic knowledge, distinct from the sensory signal, some of that structure is directly available from the sensory signal, and we suggest it should not be considered as any more different from the signal than is an abstract prelexical phoneme or feature representation. Information about linguistic structure conveyed by connected speech processes such as those illustrated in Section 2.2 above seem to us to be no more 'higher-order knowledge' than is an attribution of feature or phoneme identity from spectro-temporal properties of the signal. Psycholinguistic processes that capture these higher-order regularities direct from the signal thus seem desirable, if only for reasons of model economy. Merge takes a welcome step in this direction, though we believe it would be more effective if it took a more radical approach to the prelexical representation.

In short, we see the incorporation of pre-lexical knowledge of transitional probabilities as a sort of metaphor for just one type of linguistic structure that is available in the speech signal but absent in unstructured phoneme strings, probabilistically encoded or not. Without a wider temporal focus to allow systematic covariation of fine phonetic detail to be detected, and even allowing for transitional probabilities, correcting errors by reference to probabilities of phonemes in undifferentiated strings is unlikely to reflect the way error-correction takes place in normal listening situations. We suggest that the claims made for Merge and indeed all similar models of spoken word recognition could be more convincing if the focus were widened from individual phonemes to a range of different linguistic units with varying temporal domains, and if detailed phonetic information were preserved as late as possible in the model. Furthermore, the output as well as the input must be defined in enough detail to provide a grammar of the mapping between speech and meaning, so that it takes into account the systematic relationship between phonetic detail and phonological and grammatical status.

2.4. Summary

In summary, because the phoneme has dominated thinking in both speech science and psycholinguistic research on spoken word recognition, at the expense of other types of phonological and grammatical structure, much of the systematic variation in speech that indicates linguistic structure has been ignored. Short-domain spectral-temporal events that relate most directly to phoneme identity have dominated perceptual research in speech science, together with a tendency to separate segmental and prosodic information in thinking and in research. Partly for practical reasons, this local focus has either been adopted in many computational models of spoken word recognition and lexical access, or else it has strongly influenced them. Despite long-standing acknowledgement of the potential importance of systematic fine phonetic detail (e.g. Elman & McClelland 1986; Pols 1986) the information conveyed by the speech stream has been assumed to be more impoverished than it really is, with consequent biases in the mental processes proposed to recognize and understand spoken words.

3. Properties of the speech signal

We have rejected the traditional view that speech is understood by being organised by listeners as well as linguists into independent prosodic and segmental strands and discrete, pure forms of abstract units. Instead, we suggest that the perceptual correlates of linguistic units are typically complex, often spread over relatively long sections of the signal, simultaneously contribute to more than one linguistic unit, and don't cluster into discrete bundles in time. We believe that this complexity is a crucial determinant of how we understand speech. It is crucial because it reflects vocal-tract dynamics. And vocal-tract dynamics provide perceptual coherence, which is a central property of natural speech, and possibly the main thing that sets it apart from most synthetic speech.

These views form the springboard for a more detailed consideration of properties of the speech signal that we believe are fundamental, and thus that need to be included in a phonetically-valid model of spoken word understanding. Some of these properties are well understood and well accepted; others are less well understood and in some cases controversial. The section has seven parts. Section 3.1 deals with aspects of what makes a signal robust, natural-sounding, and interpretable as speech; we call this 'perceptual coherence'. Section 3.2 develops the point introduced in Section 2.2 that certain types of grammatical and phonological information are systematically distinguished by fine phonetic detail in the speech signal. Section 3.3 presents a polysystemic declarative linguistic-phonetic model capable of describing this type of information. We suggest that it encapsulates the type of information a listener looks for—and finds—in the speech stream, and that it may be a reasonable metaphor for how that information is represented in the brain during the process of understanding normal connected speech. Section 3.4 discusses the temporal distribution of information about phoneme-sized phonetic segments in the speech signal, distinguishing between short and long domains of influence, and exploring relationships between linguistic informativeness and the temporal extent of the domain of influence of a given phonetic segment. Sections 3.5 and 3.6 discuss the role of rhythm and time respectively in speech understanding. And Section 3.7 formally notes that phonetic categories are like other linguistic categories, and indeed all categories, in being fundamentally relational, or contrastive, in nature. Although the main focus is on systematic acoustic-phonetic properties of speech, results from experiments assessing the perceptual relevance of these properties are referred to when available.

3.1. Perceptual coherence

In natural speech, there is a tight relationship between vocal tract behaviour and acoustics, and this relationship provides the signal with perceptual coherence. A speech signal is perceptually coherent when it appears to come from a single talker because its properties reflect the detailed vocal-tract dynamics. Hawkins (1995) used the term acoustic coherence, but changed it to perceptual coherence to place the emphasis on the listener rather than the talker and to reflect the fact that multimodal information can contribute perceptual coherence (Fowler & Dekle 1991; Faulkner & Rosen 1999), or, conversely, incoherence: it is well known that when visual information conflicts with information in the acoustic signal, the resulting incoherence can radically change perception (McGurk & Macdonald 1976; Massaro 1998). However, the present discussion focuses on the acoustic signal because it is the most important medium for speech.

The concept of perceptual coherence is part hypothesis, part factually-based, in that we do not know exactly what properties make speech perceptually coherent, but we do know from many different types of work that small perturbations can change its perceived coherence (e.g. Huggins 1972a, b; Darwin & Gardner 1985). To be heard as speech, time-varying acoustic properties must bear the right relationships to one another. When they do, the perceptual system groups them together into an internally coherent auditory stream (Bregman 1990) or, in some views, into a more abstract entity (cf.

Remez *et al.* 1994, Remez *et al.* 1998). A wide range of acoustic properties seems to contribute to perceptual coherence. The influence of some, such as patterns of formant frequencies, is widely acknowledged (cf. Remez *et al.* 1981). Others are known to be important but their contribution is not always well understood. Examples are the amplitude envelope which governs some segmental distinctions (e.g. Rosen & Howell 1987) and also perceptions of rhythm and of 'integration' between stop bursts and following vowels (van Tasell *et al.* 1987); and correlations between the mode of glottal excitation and the behaviour of the upper articulators, especially at abrupt segment boundaries (Gobl 1988; Pierrehumbert & Talkin 1992; Ní Chasaide & Gobl 1993; Stevens 1998). This partial list of properties that can make a signal perceptually coherent presumably includes both general acoustic consequences of vocal-tract behaviour, such as the complex events at the boundaries of vowels and obstruents, and coarticulatory patterns of the particular language and accent being spoken.

Perceptual coherence is a fundamentally dynamic concept, as the examples above show, and it interacts with knowledge. An example of the dynamic nature of perceptual coherence interacting with general knowledge comes from experiments using 'silent-center' syllables. Silent-center CVC syllables leave intact formant transitions and the duration of the vocalic steady state, but replace the excitation during the steady state with a silence of the same duration, thus preserving only the vowels' durational and dynamic spectral information. Listeners identify the vowels of such syllables with remarkable accuracy (Strange *et al.* 1983; Jenkins *et al.* 1983; Parker & Diehl 1984; Strange 1989). More interesting is that listeners tend to hear such syllables as interrupted by a hiccup or glottal stop, suggesting that the drive to attain perceptual coherence introduces percepts of real-world experiences inherent to vocal-tract behaviour (even while speaking) but not necessarily to normal English speech behaviour. The relevance of direct realist theories (J. Gibson 1966; E. Gibson 1991; Fowler 1986; Best 1994, 1995) to the concept of perceptual coherence is obvious.

It is tempting to try to separate 'general vocal-tract dynamics' from 'knowledge' of language-specific vocal-tract dynamics like coarticulatory patterns (e.g. Perkell 1986; Manuel 1990; Magen 1997; Beddor & Krakow 1999), but this does not seem to be a valid distinction for listeners. How is a listener to know what is general and what is language-specific? Certainly, infants seem to know by 18 weeks or so of age which facial configuration is associated with particular vowel sounds (Kuhl & Meltzoff 1982) but in our view, a perceptually-coherent signal conveys far more subtle distinctions than the fairly obvious ones between the acoustic or visual properties of /i/ and /a/. A mature ability to detect and use perceptual coherence must depend on cumulative experience over several years.

If general and language-specific differences cannot be separated in practice, is it worth separating them on logical grounds? We believe that it is not. From the listener's point of view, the vocal-tract dynamics of his or her language(s) are systematic and must be known about; it is immaterial that they represent only a subset of the possible range (cf. Keating's (1985) similar conclusion for speech production). Furthermore, sounds the vocal tract can make, even ones that all or almost all humans do make, may not be perceived as speech when the relevant experience and consequent knowledge are missing, or when expectations bias the interpretation. For example, when native speakers of non-click languages hear a click language for the first time, they may not hear the clicks as part of the acoustic stream from the speaker's mouth, let alone as a structured part of his or her language. This is especially likely to happen if the context provides an opportunity for a more familiar interpretation. Anecdotally, on hearing Bushman spoken by a man striding rapidly across a desert at the outset of the movie *The Gods Must Be Crazy*, the first author, who was expecting to hear only English, first 'heard' speech plus twigs breaking as they were stepped on, then, realising there were no twigs, immediately 'heard' speech plus a disappointingly crackly sound track to the movie. It took a palpable amount of time, possibly several seconds, to realise that the clicks were integral to the speech, and thus that the language being spoken was probably genuine Bushman, even though she was aware of the existence of click languages and had heard them spoken before. (In her defence, she was very tired and not trying very hard.) Similarly, many Westerners hear dissociations between two parts of Mongolian chanting produced simultaneously by a single person. For example, one mode of Mongolian chant raises the amplitude of a single high-frequency harmonic significantly higher than that of surrounding harmonics. Westerners often hear this mode as having two sources: a deep chant from one person and a whistle from a second source. Though chanting may not be speech, it is closely related to it, and there are parallels with speech. For example, native speakers of English are typically amazed when they hear a baby making an aryepiglottic trill, because its extremely low pitch seems lower than any baby could make; but speakers of languages in which these trills are part of the speech repertoire (John Esling, pers. comm.) are presumably much less surprised.

We conclude that perceptual coherence is part properties of the physical signal and part mental construct. It is not useful to try to separate the two, because the coherent sensory experience of any physical signal is also a mental construct. In other words, just as phonemes are not in the signal, but may be constructed from it, so is all experience of words and their meanings; different expectations, different experiences or training, and different tasks, can influence how listeners respond to structured stimuli. These points have been made in detail by a number of people working in different fields of perception and taking a variety of approaches. Representative summaries can be found, for example, in Remez (2001), Moore (1997), and Warren (1999).

To sum up, perceptual coherence is the perceptual 'glue' of speech (cf. Remez & Rubin 1992). It is rooted in the sensory signal but relies on knowledge; the two are not distinct in this respect, but feed each other. It underlies the robustness of natural speech and determines why it sounds natural, and, conversely, it may be the key to understanding a number of the challenges in producing good synthetic speech, such as why synthetic speech can be very intelligible in good conditions but tends not to be robust in difficult listening conditions, and why speech synthesized by concatenating chunks of natural speech that preserve natural segment boundaries can give an impression of sounding natural even if it is not very intelligible (see Ogden *et al.* 2000).

Although we can identify a number of good candidate contributors of perceptual coherence, we do not yet know exactly what it is about the phonetics of a particular unit or linguistic structure that causes it to be perceived as a coherent unit. Because coherence seems to result from complex relationships between physical and language-systemic constraints experienced in a particular context, we need a linguistic model that allows clear predictions about the phonetic properties associated with the different phonological and grammatical structures that exist in a particular language. It needs to systematize the complexity introduced by the diverse linguistic factors that influence speech production and acoustics. The next section gives examples of some of these linguistic factors, while the following one describes the type of model we use to try to systematise them.

3.2. Systematic phonetic variation

A speech signal will not sound as if the talker is using a consistent accent and style of speech unless all the systematic phonetic details are right. This requires producing often small distinctions that reflect different combinations of linguistic properties. Speech that lacks these properties will lack perceptual coherence to one degree or another, and in consequence will be harder to understand.

There are many well-known examples of systematic phonetic variation, some of which have been mentioned above. This section does not provide a complete review, but gives just two less-known examples of systematic variation in fine phonetic detail. Other cases are discussed in Section 3.4.2 below. The first example in this section reflects complex interplay between metrical-phonological and morphological influences and is found in both clear and casual speech. The second comes from casual speech. It can be described as a simple phonetic process of partial assimilation, but it nevertheless results in subtle phonetic marking of a grammatical distinction. Both these examples have potential perceptual salience but would be neglected in most models of speech understanding.

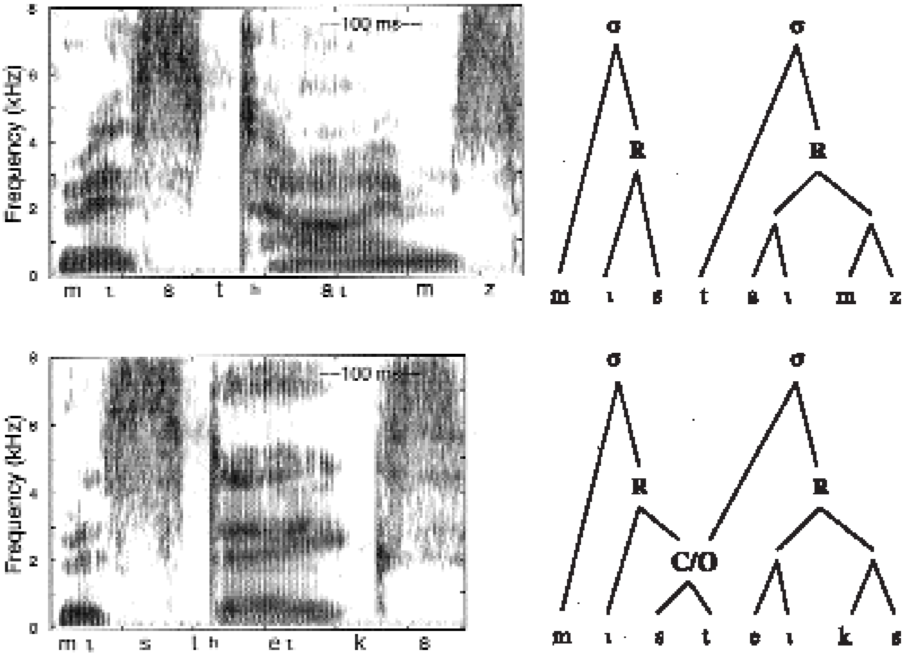

Figure 1. Left: spectrograms of the words *mistimes* (top) and *mistakes* (bottom) spoken by a British English woman in the sentence *I'd be surprised if Tess* ____ *it* with main stress on *Tess*. Right: syllabic structures of each word.

The first example compares the words *mistakes* and *mistimes*, whose spectrograms are shown at the left of Figure 1. The beginnings of these two words are phonetically different in a number of ways, although the first four phonemes are the same. The /t/ of *mistimes* is aspirated and has a longer closure, whereas the one in *mistakes* is not aspirated and has a shorter closure. The /s/ of *mistimes* is shorter, and its /m/ and /ɪ/ are longer, which is heard as a rhythmic difference: the first syllable of *mistimes* has a heavier beat than that of *mistakes*.

These acoustic-phonetic differences in the words' first four phonemes arise because their morphological structure differs. The word *mistimes* contains the morphemes *mis+time*, which each have a separate meaning; and the meaning of *mistimes* is straightforwardly related to the meaning of each of the two morphemes. But the meaning of *mistakes* is not obviously related to the meaning of its constituent morphemes, and the word is best regarded as monomorphemic. This difference in the productivity of *mis-* is reflected phonologically in the syllable structure, shown on the right of Figure 1. When *mis-* is non-productive, the /st/ cluster is ambisyllabic (*mistakes*), but the morpheme boundary in *mistimes* prevents /st/ from being ambisyllabic. Hence, in *mistimes*, /s/ is the coda of syllable 1, and /t/ is the onset of syllable 2. In contrast, the /s/ and /t/ in *mistakes* form both the coda of syllable 1 and the onset of syllable 2. In an onset /st/, the /t/ is always relatively unaspirated in English (cf. *step, stop, start*). The durational differences in the /m/ and the /ɪ/ arise because the morphologically-conditioned differences in syllable structure result in *mist* being a phonologically heavy syllable whereas *mis* is phonologically light, while both syllables are metrically weak. Thus the morphological differences between the words are reflected in structural phonological differences; and these in turn have implications for the phonetic detail of the utterances, despite the segmental similarities between the words. Complex though these influences are, listeners seem to keep track of them and use them, in that unless the various properties have the right relationships to one another, they will be heard as unnatural, and we predict they would be harder to understand.

The second example concerns vestigial acoustic indications of the presence of a /z/ in a heavily-assimilated /zʃ/ context. The two spectrograms at the top of Figure 2 show the first two syllables from *Who sharpened the meat cleaver?* (left) and *Who's sharpened the meat cleaver?* (right) respectively, spoken by the same person. Notice that the /z/ of *who's* is heavily coarticulated with the /ʃ/ of *sharpened*, so that there seems to be little or no change in the quality of the friction

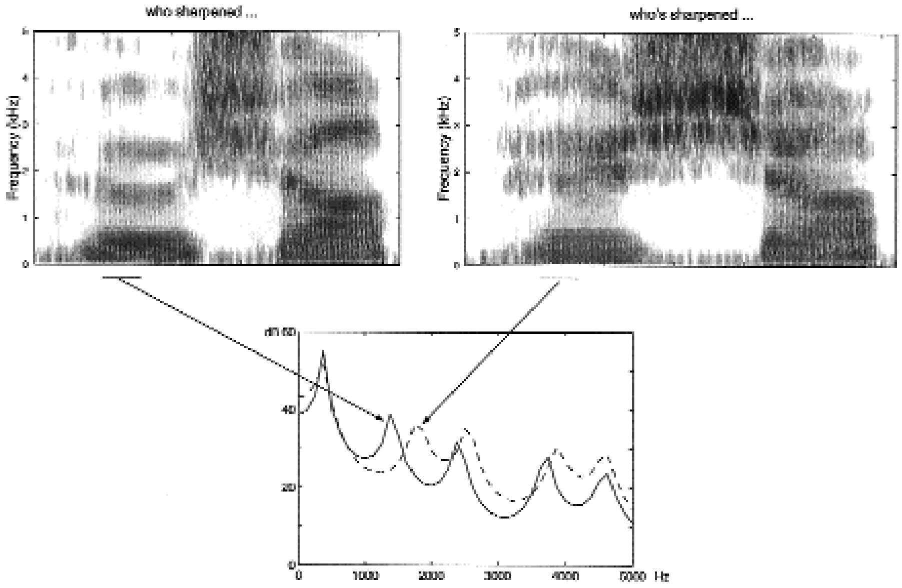

Figure 2. Top: spectrograms of the first two syllables from *Who sharpened the meat cleaver*? (left) and *Who's sharpened the meat cleaver?* (right). Bottom: 50-ms lpc spectra (18-pole autocorrelation, Hanning window) of the first part of the vowel in *who* and *who's*, as indicated by the arrows: solid line spectrum from *who*; dashed line spectrum from *who's*. The horizontal lines under the spectrograms indicate the 50-ms portions of the signal over which the spectra were made.

that represents the two fricatives. (It is, of course, longer in the *who's* than the *who* version, which provides additional phonetic evidence that there is an auxiliary verb in this structure). A conventional analysis would say that the /z/ is fully assimilated to the place of articulation of the following fricative. However, the assimilation is not complete, because the two /u/ vowels before the fricatives are very different, in ways consistent with alveolar versus palatal-alveolar articulations. The panel at the bottom of Figure 2 shows lpc spectra from 50-ms windows at the beginning of each vowel, as indicated by the connecting lines to the two spectra. Both F2 and F3 are considerably higher in frequency in *who's* than in *who*. That is, an 'underlying /z/' engenders higher F2 and F3 frequencies in the preceding /u/ and, of course, in the /h/ preceding the /u/.

Although we have not tested the perceptual power of this particular example of systematic phonetic variation, we know the difference in the vowels is perceptible, and are confident that listeners would use it. Both classical and recent experiments (Repp 1982;

Strange 1989; Duffy & Pisoni 1992; Pisoni 1997a; Kwong & Stevens 1999) suggest that most systematically varying properties of speech will enhance perception in at least some circumstances, perhaps especially in adverse listening conditions or when the cognitive load is high (Hawkins & Slater 1994; Tunley 1999; Heid & Hawkins 1999). Slightly raised F2 and F3 frequencies may not have a strong effect by themselves, and they are not always present in such sequences; but when they are there, and especially when they co-vary with other cues such as duration of the fricative, they could offer good information about the grammatical structure of the utterance.

3.3. The linguistic model for Polysp: polysystemic and non-segmental

If we are right in suggesting that the properties of speech discussed in the previous two sections significantly affect our understanding of connected speech, then we need a model that can capture them. Hence we regard each 'phonetic segment' as best described in terms of all of its structural properties, rather than solely or mainly in traditional phonetic terms. Acknowledging these structural properties allows much phonetic variation to be seen as systematic, and therefore informative.

3.3.1. Antecedents and principles from Firthian prosodic analysis

The linguistic model we propose as most useful for this purpose was developed from Firthian prosodic analysis. Its principles are described by (amongst others) Local (1992) and Ogden & Local (1994), who are at pains to distinguish it from autosegmental phonology (Ogden & Local 1994); instantiations have been used as the theoretical model for synthesizing speech that sounds natural and is robust in adverse listening conditions, namely YorkTalk (Coleman 1994; Local & Ogden 1997) and ProSynth (Ogden *et al.* 2000). Related grammatical models have been used in psycholinguistics e.g. Jurafsky (1996). We outline here only the most important properties of this class of models for our present purposes; fuller treatments are available in the references cited and elsewhere.

The most important properties of this model for us are that it is (1) declarative, (2) nonsegmental and (3) polysystemic. The most important of these is the third—that it is polysystemic. The other two properties more or less follow from this, but we deal with them first because they provide a familiar basic structure from which to begin to discuss polysystemicity.

(1) Because the model is declarative, it attempts to provide a single invariable phonological and grammatical description of any given word or phrase. Differences in phonetic forms of words or phrases are accounted for by differences in the hierarchical structural description associated with each one, as partially illustrated in Figure 1 for *mistimes* and *mistakes*.

(2) The model is explicitly and primarily nonsegmental in that traditionally segmental distinctions are seen as differences in the features placed on particular nodes in the structural hierarchy, and they thus exist in paradigmatic and syntagmatic relationships with one another. Each feature is represented at the highest possible point in the hierarchical phonological structure. When a phonological node possesses a particular feature, that feature affects every phonetic segment within its domain. For example, lip rounding due to a particular rounded vowel can be represented high in the structural hierarchy of the phrase in question, so that an appropriate duration of anticipatory rounding is achieved. How lip rounding is actually realised in each acoustic segment in this domain depends on general principles specific to lip-rounding in the particular language, and on the other properties of the structure—that is, on the full specification of the segments in question.

(3) The model is polysystemic in that language is seen as a set of interacting systems rather than one single system. We have already discussed some examples of polysystemicity in English: distinctions between content and function words in the connected speech processes they take part in (Section 2.2), and distinctions due to morphological differences between words (Section 3.2). Likewise, the nominal and verbal systems of a language can be viewed as separate, and auxiliaries can be systematically distinguished from other verbs, thereby making explicit the fact that they obey different phonetic constraints (Ogden 1999). For example, auxiliaries can be much more phonetically reduced than other verbs: compare *I've seen it* with *I have it.* Another pervasive distinction in English is that between the Germanic and Latinate lexical stress systems. These can be viewed as having different phonological systems and structures, different accentual patterns and so on, which explains why superficially similar polysyllabic words like *unnatural* and *unknown* have different rhythmic patterns from *innate*: *un-* is a Germanic prefix, one of whose properties makes the /n/ in *unnatural* and *unknown* long; in contrast, *in-* is a Latinate prefix, whose properties make the /n/ short. Latinate and Germanic prefixes may both precede Latinate stems, but only Germanic prefixes can precede a Germanic stem. Thus a

dual-system contrast combines to make three (but not four) systems. The picture is further complicated by the fact that some prefixes, such as *dis-*, can behave as if they are Latinate or Germanic (cf. *disperse*, *disburse*), while others are superficially similar but arise for different reasons. For example, *mistakes* and *mistimes* are both Germanic throughout, and the contrast between them, illustrated in Figure 1, is due to differences in morphological productivity resulting in differences in ambisyllabicity. These patterns are well known in lexical phonology (cf. Halle & Mohanan 1985) but accounted for there by using derivational rules, whereas our model captures the differences declaratively.

3.3.2. Some attributes of the model

The above properties mean that the model has different systems within linked grammatical and prosodic hierarchies. The prosodic hierarchy, for example, extends from intonational phrases to syllabic constituents, with phonological information distributed across the entire structure. Phonological contrasts can be expressed over any domain in the structure e.g. over the domain of the phrase, syllable, syllable onset, syllable rhyme, and so on. Links with grammatical structure are made at appropriate nodes in the hierarchy through shared features. So, for example, the phonetic properties consistent with a voiced dental fricative in the onset of a weak syllable signal that the syllable is a function word, and provide a link between phonetic-prosodic structure and grammatical structure; they also indicate that the word is likely to be monosyllabic so that the next syllable onset probably signals a new word. The phonological categories and features are abstract: for instance, a contrast bearing the feature label [+ plosive] has different phonetic realisations depending on where it occurs in the phonological structure. The phonetic interpretation of the feature could be entirely different if it occurs in a syllable onset or a syllable coda. Within this system, the /p/ of *potato* has more properties in common with the /t/ of *toboggan* than it does with the /p/ of *cap*.

With reference to the discussion in Section 2.2, there is nothing to stop strings of phonetic segments being identified and referred to in the traditional way, but phonemes are not a part of the structure, and allophones bring with them their full systemic description. Thus, the structural differences between the two glottal stops of [ʔandɪʔaʊʌə] are immediately apparent and unproblematic. Similarly, the differing representations of /t/ in [kɑʔənaɪʔ] and [kɑtənaɪʔ] bring with them their structural motivation.

Fundamental to this approach is that each unit of linguistic-phonetic analysis is seen IN RELATION TO its neighbours, and to other units above and below it in the prosodic hierarchy. The relational nature of speech sound categories is likewise fundamental to the long- and short-term phonetic properties discussed in Section 3.4. It is also relevant to memory, and hence to the brain's representation of linguistic categories, as discussed in Sections 4.2 and 4.4. In short, (1) a phonetic segment cannot be properly described independently of the larger structure(s) within which it occurs, and (2) each phonetic segment is definable only in relational terms.

For our purposes, the polysystemic attributes of the model are more important than the others. As we see below (Section 4.2 *ff*) there is neuropsychological evidence to encourage the view that some linguistic subsystems are functionally separate in speech processing. If some are functionally separate, it is worth pursuing the idea that they all are. Thus, although superficially the rhythmic system of English is chaotic, it can be analysed in terms of distinct subsystems, and experienced listeners are presumably able to relate new constructions to one or other of these subsystems.

We are less convinced that the model must be declarative but one reason to favour a declarative over a derivational model is that a polysystemic model requires a declarative structure, because otherwise the links between systems may be lost. Moreover, the declarative model seems broadly compatible with the types of functional groupings of neural activity described below. This does not mean to say that rule-derivations cannot be equally well represented, but what, after all, is a rule, if not the formal description of a general pattern?

Another advantage of a declarative structure is that it can be seen as providing an explicit way of modelling different degrees of casualness onto the same underlying form. This may be easier to envisage from the point of view of speech production or synthesis: when parameter values are changed to affect the tempo of speech synthesized by this model, there are concomitant spectral changes that affect perceived segmental quality (providing that the parameterization is properly implemented). A similar inverse mapping can be envisaged for understanding speech, if only as a metaphor for what really happens.

Lastly, this type of linguistic model in principle allows extension to discourse and interactions such as real conversations, which can be analysed as other systems subject to the same general constraints. To develop this point would go far beyond the scope of our present work, but the fact that it should be possible increases the appeal of the model.

3.3.3. Relationship between Firthian linguistic analysis and speech understanding

This model is linguistic. We do not claim that it encapsulates the conceptual structure that the brain constructs in order to understand spoken messages. However, we do suggest (1) that it can help guide the search for system within the acoustic-phonetic variation of speech, (2) that something like it must be used if the linguistic information available in speech is to be exploited by models of speech understanding, (3) that its principles are broadly compatible with current views of how the brain processes sensory information, and that it may therefore represent a more reasonable metaphor for how the brain constructs meaning from speech than those offered by more standard models that assume that the information in the sensory signal is relatively impoverished.

For convenience, then, we call this model Polysp (for POLYsystemic SPeech understanding), and do not try to distinguish what is linguistic from what might be psychological. We stress that we prefer the term speech understanding to speech perception, because we are trying to address how speech is understood, rather than only how it is given phonological form, the latter being the endpoint of most phonetic theories of speech perception.

3.4. The temporal distribution of information about segmental identity

We have proposed that properties of speech that make the signal perceptually coherent and that systematically reflect linguistic structure all contribute crucially to whether it is understood easily. We turn now to consider how acoustic cues to phonetic segments are distributed over stretches of speech. Although we have just proposed that a polysystemic nonsegmental linguistic model should underlie models of speech understanding, this is one of those times when it is nevertheless convenient to refer to phoneme-sized phonetic segments as entities in their own right, because phoneme names are familiar and simple.

We draw a fundamental distinction between information that happens relatively fast and information that extends over longer time intervals. This distinction is perhaps more conceptual than actual, inasmuch as it represents not a simple binary division, but rather two points spaced rather far apart along a continuum. However, we believe that the distinction has important implications for modelling speech understanding, and so is a fiction worth supporting.

3.4.1. Short-domain segmental information

In our view, the best systematisation of short-time events that we have at present is Stevens' and colleagues' work on relational acoustic or auditory invariants, which developed from a search for invariant acoustic or auditory correlates of phonological distinctive features. Ideally, these correlates are invariant relational properties described in terms of differences between successive short-term spectra. Stevens (1998) offers the fullest treatment of this work, using a slightly different style of presentation from his earlier work, summaries of which can be found in Stevens (1983, 1989, 1995) and Stevens *et al.* (1992). Stevens distinguishes three types of features: landmarks, which describe degree of constriction and are hence articulator-free; and two types of articulator-specific feature, one representing place of articulation, and the other representing additional properties like voicing and nasality.

Landmarks in the signal mainly describe manner of articulation via degree of constriction, although in practice, 'pure' manner features may be inextricably associated with excitation type. Landmarks exploit quantal relationships, and most are found in short (10-40 ms) sections representing either local maxima or minima in degree of spectral change in that region. The successive spectra that represent landmarks for consonantal closures and releases fall on either side of an abrupt acoustic boundary. For vowels, they fall where the vocal tract is maximally open between two consonants, and hence approximately the centre of a vowel steady state, marked acoustically by a local maximum in F1 frequency or by relatively high amplitudes at low frequency. So landmarks distinguish between the syllable nucleus and its boundaries.

These short-term events in the acoustic signal thus reflect the way the signal changes as the vocal tract moves at critical points in the speech stream. They are relational, reflecting the fact that even short-term events are interpreted with reference to their context. Notice, however, that landmarks for vowels may extend over longer time domains than landmarks at abrupt segment boundaries involving consonants, at least for relatively slow speech. Longer durations will be necessary for identification of the phonological class of certain vowels (cf. Hillenbrand *et al.* 2000). This is one example of the relative nature of the distinction we draw between short-term and long-term information in the signal.

Articulator-specific features also tend to be short-term and occur in the vicinity of the landmarks, but, unlike landmarks, they can involve several different acoustic properties and can extend for longer, up to about 100 ms.

Since landmarks distinguish between the syllable nucleus and its margins, the phonological features they include (consonantal, sonorant, continuant, syllabic, lateral, and strident), (a) mark parts of the signal that have status relatively high in our model of linguistic structure, including, vitally, those contributing to rhythm, and (b) singly or in combination with one another, correspond fairly closely to categories identified as robust features by Zue (1985). (The main difference between Stevens and Zue is that whereas robust features are seen as relatively steady state, Stevens' landmarks (especially the consonantal ones) are typically dynamic, and define points in time at which particular critical events occur.) Both these attributes of landmarks seem particularly satisfactory from the point of view of offering evidence that they are central to understanding speech. Although lateral and strident features may be considered as rather different from the others, and not applicable at high levels, in fact there are some independent reasons why they could reasonably be modelled relatively high in structure. Some of these are discussed in Section 3.4.2. below.

Opinions differ as to the validity of Stevens' arguments (cf. the commentaries following Stevens 1989). In our view, relational invariants for features tell part of the story, though not all of it since the documented spectral patterns work better in some contexts than in others. It may be that, when they work less well, other properties of the signal need to be considered. Some of these may be as local and short-domain as Stevens' proposed relational invariants, but in other cases, information over rather longer domains may be relatively more informative. For example, auditory enhancement theory represents a systematic attempt to show how a variety of acoustic properties, some local, and one of at least a syllable in duration, may combine to produce robust percepts of phonological voicing, not by enhancing the abstract feature of voicing itself, but by enhancing one or more intermediate (abstract) perceptual properties that together contribute to the overall perception of voicing (Kingston & Diehl 1994). The details are both controversial (cf. Nearey 1995, 1997) and not fully worked out, but it is sufficient for our present purpose simply to agree that, by and large, a great number of perceptual cues to segment identity provide their information in less than about 40 ms; and that, although that information is of variable quality, there is often much justification for regarding it as the 'primary' cue. That is, short-domain acoustic properties like those Stevens identifies typically convey reliable linguistic information.

Not all short-domain cues are as short as those Stevens identifies, but all appear to be relational, and the time-scale over which they operate is itself relational. An early indication of this is Miller & Liberman's (1979) demonstration that the same synthetic formant transitions can cue /b/ or /w/ depending on the duration of the following vowel, which is interpreted as indicating the rate of speech (cf. also Miller & Baer 1983; Schwab *et al.* 1981). There are, of course, limits to the range of variation possible, and by and large these limits are determined by the surrounding context: each acoustic property is interpreted in relation to other acoustic properties in the signal. Such limits parallel those in the spectral domain. For instance, coarticulated vowels in context are identified no worse than isolated vowels and sometimes better, although the steady states of different coarticulated vowels are not as distinctive in F1-F2 space as those of isolated vowels (Strange *et al.* 1976; Strange *et al.* 1979; Gottfried & Strange 1980; Macchi 1980; Assmann *et al.* 1982). Thus, understanding linguistic meaning appears to be very much dependent on the 'Gestalt' conveyed by the whole signal, rather than on the gradual accumulation of information from a sequence of quasi-independent cues.

3.4.2. Long-domain segmental information

Long-domain segmental information can be defined in terms of time, but is perhaps more sensibly defined in terms of syllables, since syllables vary so much in duration. We define long-domain perceptual information as extending for at least a syllable, or, somewhat arbitrarily, for about 100 ms or more. However, some long-domain information lasts much longer.

Some of the information conveyed over long domains is well established. For example, vowel-to-vowel coarticulation (Öhman 1966; Recasens 1989; Manuel 1990; Magen 1997; Beddor & Yavuz 1995), lip rounding (Benguerel & Cowan 1974), and nasalization (Clumeck 1976; Bell-Berti 1993; Krakow 1993, 1999; Solé 1995). However, with some notable exceptions (e.g. Alfonso & Baer 1982; Fowler & Smith 1986; Beddor *et al.* 1986; Warren & Marslen-Wilson 1987; Krakow *et al.* 1988; Marslen-Wilson & Warren 1994; Beddor & Krakow 1999) there has been comparatively little research on the perceptual power of such cues, and little attempt to integrate them explicitly into either psycholinguistic or even phonetic models of speech perception. Although activation models can in principle accommodate long-domain effects with no trouble, most psycholinguistic and computational models neglect them in practice, because they assume a discrete and abstract input unit, usually phonemic. Amongst phonetic models, gestural models (Liberman & Mattingly 1985; Fowler 1986; Best 1994, 1995) could be said to be predicated on principles of long-domain influences. However, perhaps because gestures are axiomatic in this class of models, and related to 'linear', phoneme-like segments independent of other linguistic structure, the relative influence of long-domain and short-domain information has received no more attention in gestural than in auditory phonetic models. In particular, not even gestural models of speech perception can account for the class of long-domain perceptual cues known as resonance effects. As will become clear, we believe that resonance effects have great theoretical significance (possibly far greater than their contribution to speech understanding) and for this reason, and the fact that these data are comparatively new, we discuss them in some detail.

In many accents of British English, syllable-onset /l/ is realised as a clear palatalised segment that contrasts with a relatively dark syllable-onset /r/. Kelly & Local (1986) observed that these clear and dark resonances of /l/ and /r/ not only colour the entire syllable of which they are a part, so that the /i/ of *Henry* is darker than the /i/ of *Henley*, but can also affect vowels in neighbouring syllables. Subsequent work has supported their observations with acoustic (Hawkins & Slater 1994; Tunley 1999; Heid & Hawkins 2000) and EMA measurements (West 1999a) of a relatively wide range of carefully-controlled sentences, such as *We heard that it could be a mirror* and *We heard that it could be a miller*. Formant frequencies are lower in parts of the utterance when the last word is *mirror* than when it is *miller*.

Collectively, these studies show that the realisation of these liquid resonances is affected by a wide range of factors about which there is still much to be learned. They are fiendishly difficult to study in controlled experiments, partly because the effects are subtle, but mainly because, naturally, they interact with other linguistic-phonetic variables to produce complicated effects. For example, for independent reasons, Hawkins & Slater (1994) and Tunley (1999) both suggested that unstressed syllables will be more subject to resonance effects than will stressed syllables, but Tunley (1999) also found that vowel height affects them, and this is not straightforwardly separable from syllable stress. Moreover, Heid & Hawkins (2000) conjectured that acoustic consequences of severe vowel reduction might obscure liquid resonance effects.

Heid & Hawkins (2000) tried to disentangle some of the effects of metrical and segmental structure on the spread of anticipatory resonance effects in the speech of one British English male. They found resonance effects in some utterances for up to five syllables before the conditioning /r/ or /l/; this represents a temporal range of half to one second, depending on segmental and metrical structure. These durations are far longer than current models of speech understanding deal with. As expected, they found both segmental and metrical influences on the appearance of these effects, but not always in the expected patterns. For example, it was expected that stressed syllables and intervening lingual (especially velar) consonants would show weaker effects than unstressed syllables and labial consonants. The observed patterns showed that these expectations, though broadly correct, were too simplistic. Intriguingly, some unstressed syllables that showed anticipatory resonance effects preceded a stressed syllable that did not show the effect; in other words, the resonance appears to 'pass through' a stressed syllable to colour earlier unstressed syllables. For example, *it* showed anticipatory /r/-resonance effects in *We 'heard that it 'could be a \mirror* when *could* was stressed but not when it was unstressed (*We 'heard that it could be a \mirror*), even though stressed *could* did not itself show evidence of /r/-resonance effects. This stress-dependent difference may have been connected with the fact that the vowel in *it* was longer and less reduced before stressed than before unstressed *could*. Similarly, segment type (velars *versus* bilabials) appears to have a local influence on the appearance of resonance effects, but need not block their anticipatory spread to yet earlier syllables.

Though much more work is needed before we can fully describe the realisation of resonance effects, listeners know about them and can use them. In a forced choice test, West (1999b) showed that when regions of speech surrounding an /r/ or /l/ are replaced by noise, differences (resonance effects) in the remaining, more remote, regions of the utterance are sufficiently distinctive to allow listeners to tell whether the excised word contained /l/ or /r/. Perceptual tests in our laboratory have compared the intelligibility of synthetic speech with and without resonance effects when it is heard in naturally-fluctuating noise from a cafeteria. When resonance effects are included, phoneme intelligibility increases by about 10-15% (Hawkins & Slater 1994; Tunley 1999).

Synthetic speech that includes these resonance effects does not usually sound very different from synthetic speech that omits them, and many listeners say they can hear no difference even though their intelligibility scores indicate that they can. We thus conjecture that resonance effects have a subtle effect on the perceptual coherence of the signal, sufficient to increase intelligibility in adverse listening conditions but not to change any but the most detailed phonetic transcription of the utterance. We do not know whether they are perceptually salient in all listening conditions or only in adverse circumstances, but since speakers produce them and listeners use them to understand connected speech, models of speech understanding must be able to account for integration of subtle information relating to phonetic segments over several hundreds of milliseconds.

3.4.3. Relationships between informativeness and domain of influence

Every phonetic segment is probably cued to one extent or another by both long- and short-domain acoustic properties, but, generally speaking, short-domain events within the framework outlined above tend to be highly informative about some aspect of segmental identity, whereas information that takes longer to unfold is likely to carry weaker information, time unit for time unit. However, like the fiction of short-domain *versus* long-domain cues, the correlation between the duration of a cue and its informativeness is imperfect.

Standard examples of short-domain, highly informative events are acoustic cues associated with stop identity that are found in the vicinity of the acoustic segment boundaries, and traditionally described in terms of formant transition patterns and the shape of the burst spectrum, although alternative descriptions such as Stevens' may make more auditory sense. Examples of more long-domain, less informative acoustic cues include information about place of articulation of a coda obstruent that results from coarticulatory effects distributed throughout the rhyme of a syllable (e.g. Warren & Marslen-Wilson 1987), information about the voicing of a coda obstruent that is present in an onset /l/ in the same syllable (Hawkins & Nguyen 2000, 2001, in press), and the resonance effects discussed in Section 3.4.2, which may extend over several syllables.

In contrast, however, there are weak short-domain cues (an example is first formant cutback for the voicing of English stops) and very powerful, more long-domain cues such as vowel duration in cueing the phonological voicing of coda obstruents in English and many other languages. Other properties, such as those indicating vowel identity, also take time to build up, and it is not clear whether they should be classed as short- or long-domain. For example, at normal rates of speech, all diphthongs, and vowels like /a/ in some consonantal contexts, can take 100 ms or more before they are identified with certainty. Similarly, reliable information about manner of articulation, including excitation source and certain distinctions that rely on relative duration (e.g. between affricates and stop-fricative sequences as in *grey chip, great ship*, etc), by its nature extends over relatively long durations. Stevens' work is interesting in this context because his model looks for highly informative short-domain acoustic events to identify features such as nasality that can in fact spread (often weakly) over rather long domains.

These complex interrelationships between domain size, domain duration, and quality of linguistic information pose problems for theoretical descriptions that rely on phoneme-sized phonetic segments that are unstructured, or in which the various types of structure are separated into rather independent strands. They are more easily accommodated in a theory that accords less status to phoneme-sized segments, and more to time-varying relationships within a systematic, rich linguistic structure.

In summary, we adopt the position that the information the signal provides includes the immediate effect of clear, usually short-term, events (which may resemble Stevens' relational invariants), and also the cumulative effect of information that is distributed more widely across the signal. The distributed information can be very clear, as for strident fricatives and aspects of vowel quality and nasalization, or it can be rather weaker, as in the coarticulatory influences discussed with Figure 2 and in this section. As long as weak information points consistently to the same linguistic structure(s) across time, it seems reasonable to suppose that it could have a strong influence on speech understanding (Hawkins 1995; Hawkins & Warren 1994).

Taken together with the type of structured, nonsegmental, polysystemic model that is Polysp, it follows that we assume that there is no pre-defined order in which a listener must make decisions in order to identify words or understand meaning: decisions are made in parallel, and can be at any level of linguistic analysis. For example, a strong syllable can be identified with high certainty even when its component phones are only partially identified; a strident fricative can be identified with high certainty when there is only low certainty about which syllable it is a member of, and whether that syllable is stressed. These types of information offer islands of reliability at different levels within the linguistic structure. All available evidence is, as it were, fitted together until the best linguistic analysis is found. Hence the importance to perception of weak but systematically-varying information as well as of powerful information, because it adds perceptual coherence.

3.5. Speech rhythm

Although both segmental timing and phonological relationships must be organised in particular ways to produce the percept of natural-sounding rhythm, neither segmental durations nor relational phonological contrasts 'are' rhythm, nor do they fully explain its role in perception (e.g. Buxton 1983; Beckman 1986; Couper-Kuhlen 1986; Ladd 1996; Tajima & Port in press; Zellner Keller & Keller forthcoming). The potential units and/or focal points in the signal that are rhythmically important have been investigated for years and some of the most interesting experimental demonstrations relevant to our purposes are in fact some of the earliest (e.g. Huggins 1972b; Kozhevnikov & Chistovich 1965; Lindblom & Rapp 1973).

One important issue for models of speech understanding is that local changes in timing can have effects on perceived rhythm and perceptual grouping that are distributed over a domain of several syllables or the entire utterance. Thus Huggins (1972a), using natural speech to investigate just noticeable differences (JNDs) for segment durations, found that altering segmental durations sometimes caused the perceived distribution of stress in the experimental utterances to change. For instance, when the duration of initial /l/ was altered in the phrase *more* ***l****awful than it was*, primary stress was reported to shift between *lawful* (when /l/ was long) and *was* (when /l/ was short). The duration of a segment thus affected both the overall rhythm of the utterance and the perceived prominence of the word *was*, five syllables away from the segment that had been altered. Distributed effects of local changes can also be seen in cases where phonetic detail affects perceived syllabicity and, as a result, word identity. For example, Manuel *et al*. (1992) analysed casually-spoken natural tokens of the words *support* and *sport*, and found that the oral gestures for the consonants /s/ and /p/ in *support* may be timed so as to be contiguous, leading to potential confusion with *sport*. The acoustic consequences of glottal events, however, differed in their timing between *support* and *sport*. They report that these subtle timing differences affected word identification, and by implication also the overall rhythm in the sense of the number of syllables of the perceived word.

In the study cited above, Huggins (1972a) also found that JNDs for segment durations tended to be smaller when subjects reported that they had been attending to the rhythm of the sentence rather than the component phonemes. These observations suggest that the relationship between segmental detail and higher-order organisation is not one-way: the presence of higher-order patterning seems to refine listeners' ability to detect fine-grained properties at the level of segments.

These and similar data have a natural interpretation in the context of a declarative linguistic model, in which rhythm is not associated with any one particular linguistic unit, but is implicit throughout the hierarchy by nature of the relational contrasts that exist at all levels. These relationships collectively determine segmental-level timing (Zellner Keller & Keller forthcoming). So, for example, Ogden *et al.* (2000) argue that in speech synthesis, if all the phonological relationships are stated correctly and the entire structure receives a suitable phonetic interpretation, the resulting utterance is heard as having an appropriate rhythm.

3.6. Representation of time in speech understanding

With the exception of the simple durational parameter of VOT and effects due to rate of speech, the temporal properties of speech have received less attention than spectral properties in research on speech perception, and the time domain is relatively underplayed in theories of spoken word understanding. Models frequently assume, for example, that the signal is sampled at small, constant intervals of 5-10 ms (e.g. McClelland & Elman 1986; Johnson 1997; Nearey 1997), and some of the more interesting neural models represent time as gross quantal changes (e.g. Abu-Bakar & Chater 1995).

Yet the temporal structure of speech offers a rich source of information for listeners. Low-frequency information may normally be processed in the time domain rather than the frequency domain, so speech frequencies that have rates slower than a few hundred Hertz may be perceptible from the temporal structure of the speech signal alone (Greenberg 1996; Moore 1997; Faulkner & Rosen 1999). At these relatively slow rates, amplitude variation reflects manner of articulation and hence syllabic units, and also aspects of segmental identity such as stop release *versus* affricate *versus* fricative onsets, and presumably sonorant consonants *versus* vowels. Other candidates for temporal processing are fundamental frequency and low first formant frequencies.

Nevertheless, when there is a choice, phoneticians typically emphasise the spectral in preference to the temporal aspects of perceptually important acoustic-phonetic properties, although temporal aspects are acknowledged to be important and heavy use of illustrative spectrograms does not hide them. Even exceptions like Stevens' use of frequency and amplitude changes in successive short-time spectra to describe perceptually crucial events (Section 3.4.1) make relatively restricted use of time.

One reason for this preference for the spectral domain may be the common assumption of early abstract representation of the physical signal in terms of a string of feature bundles or phonemes, which simplifies the modelling of the mental representation of speech time to that of sequencing abstract linguistic units. Such simple representations of phonological form may be encouraged by a further tendency to assume that lexical and even sentence meaning remains unchanged no matter how fast the speech: invariance of form and meaning if not of the physical signal. In reality, an important aspect of meaning in normal conversation includes so-called paralinguistic information like overall rate and rate changes, but even if we ignore this, acoustic-phonetic models of speech perception must be able to account for variations in rate of speech, if only because of the segmental reorganisation that typically accompanies them. Time as rate is therefore crucial.

Another obstacle to representation of time in speech perception is that rather little is known about how temporal and spectral properties are integrated in speech processing. Auditory models like AIM (Patterson http://) that address this issue are themselves in development, so, in addition to the significant practical challenges of combining work from different disciplines, it is tempting to pursue phonetic and psycholinguistic research independently of psychoacoustic research, and especially of particular models, for they may change.

However, no one doubts that the time domain is a key defining attribute of speech and its perception. Rhythm is clearly crucial, and systematic variation in the time domain appears to underlie the perception of speech as a coherent signal. As noted three decades ago by Huggins (1972b:1280), in speech synthesis "*spectral* detail can sometimes be dispensed with, provided that *temporal* detail is intact". Huggins cites the example of Cohen *et al.* (1962), who synthesized "surprisingly good speech" from spectrally gross building blocks, by carefully reproducing the temporal properties within and between the blocks. More recently Shannon *et al.* (1995), using a technique that preserved amplitude information but virtually abolished spectral variation, showed that listeners can recognize short utterances and even nonsense syllables almost perfectly from the temporal envelopes of spoken utterances (cf. van Tasell *et al.* 1987). Sine-wave versions of utterances may be perceived as speech if they exhibit temporal patterning characteristic of normal spoken utterances; but if they consist only of sequences of isolated vowels, they are not perceived as speech (Remez *et al*. 1981).

The fundamental question, then, concerns what might be the appropriate window or windows of temporal analysis. Overall rhythm and intonation provide good evidence that we need a long time window in addition to a short one, but there has been little suggestion that we need a time window longer than two or three segments for the perception of segmental identity, even though models of speech production use windows longer than that to model, for example, spread of lip rounding. However, one consequence of our view that phoneme-sized segments should be interpreted in their appropriate structure is that we immediately have a long-domain window, because temporal factors contributing to overall rhythm are distributed throughout the structure and are not separate from segmental identity. Direct empirical support for this point comes from findings that lingual articulation varies systematically with prosodic structure, in a way that emphasizes the differences between consonants and vowels at prosodic boundaries. Consonant articulations are often ‘strengthened’ at the edge of a new prosodic domain, and such strengthening is greater for larger domains than for smaller ones, i.e. for domains higher up in the prosodic hierarchy (Fougeron & Keating 1997; Keating *et al.* in press).

In addition to our views on perceptual coherence, two sources of empirical evidence indicate that segments need a long as well as a short time window. The most compelling empirical evidence is the long-domain resonance data discussed in Section 3.4.2. The other evidence comes from a word-spotting experiment by Smith & Hawkins (2000). In word-spotting experiments, the task is to respond as fast as possible when a real word is heard within a longer nonsense sequence, such as *oath* in *puzoath*. Smith & Hawkins (2000) independently varied the syllable stress and the presence or absence of a word boundary in utterances like *puzoath* to make four variants of each. These manipulations produced allophonic differences in the first (nonsense) syllables reminiscent of those in *mistimes* and *mistakes* (Figure 1). As predicted, listeners spotted the embedded words faster and more accurately when the sequence had been spoken as if it had a word boundary, and the segmental allophonic variation had a stronger influence than syllable stress. These data affirm the perceptual relevance of the linguistic structures used in Polysp, and suggest that systematic segmental variation in whole syllables can combine to facilitate word recognition. In other words, optimal perception of segmental information (which, after all, can be seen as the medium that 'carries' prosodic information) may require analysis windows that span at least one or two syllables and sometimes much more. On independent grounds, others have proposed processing buffers and attentive processes that last 100-300 ms (e.g. Silipo *et al.* 1999; Grossberg 2000a); we think some such processes may last even longer.

3.7. The relational status of linguistic categories

This final part of the discussion of properties of the speech signal largely reiterates individual points made above, but they merit being drawn together in one place. One advantage of Polysp is that it emphasises the relational nature of linguistic categories at the phonetic as well as the phonological level: many linguistic categories can be identified within levels in the model, but none can be properly described independently of its place within the larger structure of the utterance. It is obviously possible and for some purposes entirely justifiable to discuss a single type of category—syllables, and their stress, for example—independently of other categories in linguistic structure, but a complete description of their phonetic realisation is forced to take account of the complete, rich linguistic structure in which they are realised. Phonetic variation is thus simultaneously highlighted and constrained—or systematized—by position in structure, and no one level of analysis can be taken as more important than any other. In particular, it is possible to include phoneme labels in a separate layer, but abstract, context-free phonemes cannot be represented because each node in linguistic structure can only be properly described with reference to its place in the entire structure—i.e. with respect to its context. Thus the gain in using phoneme labels is the provision of a convenient and familiar label, but no more. To be useful at the acoustic-phonetic level, one must know about their context. We believe that this system, in which 'units' are functionally inseparable from 'context', comes close to reflecting current understanding of acoustic-phonetic and perceptual reality, as discussed below.

4. Properties of listeners

One purpose of the previous section was to highlight some of the important properties of speech that have tended to be sidelined in models of speech understanding. Similarly, the purpose of this section is to highlight some aspects of cognitive behaviour that we believe should play an important part in models of speech understanding. We arrived at the ideas put forward here over a number of years, as a consequence of trying to explain relationships between phonetics, word forms and meaning, in ways that take into account a wide range of findings from language acquisition, language breakdown in the aphasias, and episodic memory, together with exemplar-based theories of speech perception. We were delighted that, far from being hopelessly speculative, there is good support for our conclusions in recent neuroscientific literature. We are therefore encouraged to make them a central part of our theoretical position.

4.1. Learning: 'adaptation' and the plasticity of phonetic categories

Listeners typically adjust rapidly to different speakers, rates of speech, accents and other sources of variability in the speech signal. Their responses in laboratory tasks also evince a fast-acting sensitivity to statistical, temporal and other properties of stimulus sets. The available data suggest that listeners have detailed memory for specific perceptual experiences and are accordingly capable of continuous learning, and that speech sound categories are labile. These facts present a challenge for models of spoken word recognition and speech understanding in general.

4.1.1. Adaptation to speakers, rates of speech, and regional accents

For a long time, the widespread conception of adjustments to different speakers, rates of speech, and so on was that they involved 'normalization' which eliminated some of the undesirable inter- and intra-speaker variability in the signal, allowing an abstract linguistic code to be derived. The normalization approach was not, however, a monolith: Goldinger (1997), Pisoni (1997b), and Johnson *et al.* (1999) give examples of alternative earlier approaches, such as Richard Semon's theory of memory in the early 20th century. Since voice and rate properties have been shown to be retained in memory and to affect phonetic processing (see Pisoni 1997b and Goldinger 1997 for reviews), the term 'adaptation' seems more appropriate than normalization.

Multiple talkers and rates of speech can make some speech tasks harder, but they can also add useful information. The presence of many voices in training materials facilitates the learning of foreign segmental contrasts (Pisoni *et al.* 1994). Multiple-talker word lists are recalled more accurately than single-talker lists at slow rates of presentation (although less accurately at fast rates, perhaps due to distraction: Martin *et al.* 1989; Mullennix *et al.* 1989; Goldinger *et al.* 1991). Presumably the presence of many voices in conversations should enhance communication, since turn-taking between speakers is part of the dynamics of the conversation and changes in the speaking voice convey information.

Although adaptation to all aspects of a new regional accent presumably takes several months, adaptation to a new voice seems to be almost immediate. Listeners appear to use not only 'intrinsic' factors like formant spacing (Ladefoged & Broadbent 1957; Remez *et al.* 1997; Fellowes *et al.* 1997), but also diverse multimodal 'extrinsic' cues including breath sounds (Whalen & Sheffert 1997) and seen or imagined information about the gender of a speaker (Johnson *et al.* 1999). Perceptual adaptation also appears to be reflected in individuals' own speech production in a rather immediate way. Goldinger (2000) showed that when speakers read words aloud before and after hearing a list containing those words, their productions after exposure to the list were judged (by third parties) to be better imitations of the tokens that they had heard.

4.1.2. Adaptation to properties of stimulus sets

The extraordinary sensitivity of listeners to the detailed properties of stimulus sets is demonstrated directly in the vast literature on boundary shifts in categorical perception and other experiments involving phoneme identification, as well as indirectly by the type of experimental controls typically built into such experiments to minimise unwanted influences on establishing a boundary between two phonemic categories (cf. Repp 1984; Repp & Liberman 1987). Many experiments on categorical perception show category-boundary shifts as the distribution of the stimulus set changes, and shifts can be induced in the voicing boundary by randomising stimuli so that different talkers are heard in the same session, or in separate blocks (e.g. Johnson 1990; Green *et al.* 1997). These psychophysical context effects on category boundaries are well known, and are described in more detail in Section 5.4 below.

4.1.3. Adaptation as a form of learning

The type of training that listeners undergo can change their responses to acoustic stimuli. In the real world of phonetics teaching, we have long known that this is true, for much time in practical phonetics is spent on learning to perceive distinctions that are phonemic in someone else's language. Evidence from infant speech perception indicates that what we are probably doing is relearning what we apparently could do as babies (for a review, see Pickett 1999:ch 13). There are also a number of laboratory demonstrations of this process. Recent examples conducted in the context of the perceptual magnet effect and the internal structure of sound categories (Grieser & Kuhl 1989; Kuhl 1992; Iverson & Kuhl 1995, 1996; Kuhl & Iverson 1995) include Guenther *et al.* (1999), who found that categorization training made listeners less sensitive to differences between non-speech stimuli, while discrimination learning increased sensitivity, and Barrett (1997), who found similar results for music and speech stimuli. This work is described in more detail in Section 4.4.2 below. Learning is also involved in more commonplace experiences such as adaptation to a new regional accent. We have recently observed that listeners learn to use very subtle acoustic properties of the stimulus set and change criteria rather fast when the situation demands it (Hawkins & Nguyen 2001).

4.1.4. The theoretical significance of adaptation effects

Adaptation phenomena challenge conceptions of speech sound categories in a number of ways. One of these concerns the specificity of information remembered: how does a stimulus in a familiar voice map on to similar productions by the same or similar speakers? Then, given this specificity, how do listeners generalize even from limited experience with a speaker's voice so that novel tokens in that voice can be identified more easily than novel tokens in an unfamiliar voice, as shown for example by Nygaard *et al.* (1994) and Nygaard & Pisoni (1998)?

In addition, adaptation effects force us to recognize the lability of speech sound categories: that boundaries between categories are plastic, changing with factors such as surrounding phonetic context, talker, speaking rate, and all kinds of linguistic experience such as lexicality, frequency, and recency, as well as the detailed acoustic properties of the particular speech sound. This context-sensitivity is not simply a 'performance' factor that influences how the category is identified from the sensory signal: it indicates that the mental representation of linguistic-phonetic categories is fundamentally relational and multidimensional. Finally, although many factors that cause category boundary shifts are well known, the temporal evolution of the perceptual process which results in one categorization or another is poorly understood. Adaptation phenomena, reflecting as they do real speech situations, suggest that this dynamic aspect of categorization is particularly important. These conclusions contribute to our conceptualisation of memory, language acquisition, and the nature of linguistic categories, as discussed in the following sections.

4.2. Memory

Memory is obviously crucially involved in understanding speech, although its prominence in the literature on models of speech perception has waxed and waned over the years. Early work distinguishing between short-term 'auditory' and longer-term 'phonetic' memory modes was introduced to explain categorical perception (Fujisaki & Kawashima 1970; Pisoni 1973) and has influenced work in that area ever since, often by manipulation of interstimulus intervals to gain insights, for example, into discriminatory vs. classificatory capabilities (e.g. Werker & Tees 1984, Werker & Logan 1985). This basic distinction has parallels in the psychoacoustic literature, for example in Durlach & Braida's (1969) distinction between a sensory-trace mode and a context-coding mode, applied respectively to basic auditory sensitivity and phonetic labelling in work on perception of consonants and vowels (Macmillan *et al.* 1988). Note the explicit connection in the psychoacoustic model between phonetic categories and context.

These memory modes are undoubtedly important, but they focus on explaining one particular type of (possibly not very natural) speech perception, and there are probably other 'memories' capabable of holding other types of information, such as a 'working memory' for processing syntax, for example. When the focus of phonetic research moves from labelling phonemes to understanding a message, the importance of more wide-ranging connections between speech and memory becomes clearer.

At the risk of stating the obvious, memory encodes experiences. Hence the memory of an object or event is based on sensation: what it looked like, sounded like, what it felt like to touch or manipulate, what happened when it was manipulated in a particular way, what emotions were experienced at the time, and so on. Memories are thus complex, multimodal, and intensely personal: no two people have identical memories of the same event, not least because connections between the various modalities of a single memory must be deeply affected by what the individual is attending to at the time. Concepts related to words are formed from experiences in interaction with the individual in just the same way. MacWhinney (1999:218) eloquently describes multimodal, dynamic memories for the concept *banana*: the shape, colour, texture, smell, range of sensations while peeling a banana, taste, and possibly other sensory experiences associated with bananas may all be activated when we hear or see the word *banana*. He terms these sensations ‘affordances’, and describes a theory of ‘embodiment’ of which affordances (neural responses reflecting features of the object relevant to interacting with it) are one of four perspectival (cognitive) systems necessary to understanding sentence meaning, interpreted in the rich and realistic sense of implied as well as denotative meaning.

This interesting polysystemic approach is intrinsic to behaviourism as well as to branches of philosophy, but it did not play a major part in mainstream linguistics and experimental phonetics of the last 40 years. However, it was not completely neglected. For example, MacWhinney takes the term affordance from Gibsonian philosophy (J. Gibson 1966; E. Gibson 1991), which underpins the direct realist class of phonetic theories of speech perception (e.g. Fowler 1986; Best 1994, 1995). The approach is also compatible with exemplar-based theories of speech perception (e.g. Goldinger *et al.* 1991; Goldinger 1997, 2000; Johnson 1997).

Most pertinently, there is now a variety of neuropsychological evidence that memories and concepts are linked to the modalities in which they were experienced, and thus are complex and multimodal and can include a temporal aspect. Evidence from aphasia, and from functional brain imaging (fMRI, PET, and ERP) and neurophysiological (EEG, MEG) studies of normal people converges to show that, while some particular parts of the brain are closely associated with specific behaviours, no one part can be said to govern a particular behaviour. Some of this evidence and associated argument is quite old (e.g. Warrington & Shallice 1984; Warrington & McCarthy 1987), but the surge of relevant work in functional brain imaging is recent. Excellent recent reviews and discussions are provided by Coleman (1998) from the phonologist’s perspective and Pulvermüller (1999 and associated commentaries) from the psycholinguist’s perspective, so the main points need only be summarised here.

Memory for language, and hence language processing, including processing individual words, is distributed in both cortical hemispheres, as well as more primitive parts of the brain such as the limbic system and possibly the cerebellum. The limbic system is basic to the experience and expression of emotion, while the cerebellum is concerned with coordinated movement. While certain parts of the brain such as the visual cortex are involved with many words, there are differences in the distribution of brain activation associated with different categories of word. The evidence that 'vision' and 'action' words activate different parts of the brain is particularly convincing. This includes the gross grammatical distinction between nouns and verbs, and more fine-grained distinctions such as words for animals versus for small man-made objects like tools. Broadly speaking, words that involve action are associated with parts of the brain that are concerned with action and motor control, while words that do not generally involve action are more strongly associated with visual areas. There is also evidence for differences in brain processing between so-called content and function words. While both hemispheres are strongly involved in processing content words with physical referents like concrete nouns, the more abstract function words are more localised to the left perisylvian region, i.e. the region around the Sylvian fissure which separates the temporal from the frontal lobe, and that is traditionally associated with speech and language; it includes Broca's area and primary auditory cortex.

This brief summary indicates not only that memory is fundamental to models of speech understanding, but that the theory of memory adopted can deeply influence other aspects of the model. That memories are dynamic and structured according to the modes of experience they encompass lends particular force to the arguments in Section 4.1 above that the speech understanding process is fundamentally adaptive to the detailed circumstances in which speech is heard.

Further, this model of memory is compatible both with exemplar-based models of episodic memory and with our view of the mental organisation of speech represented by Polysp's declarative, polysystemic properties (Section 3.3). By implication, if the linguistic structure of Polysp is relevant to speech perception, then the mental representation of linguistic structure is grounded in tangible phonetic memories. These phonetic memories are crucially linked to other sensory memories which may not be linguistic. Coleman has likewise argued that the mental representation of words is "essentially phonetic, rather than symbolic-phonological." (ms: abstract), as discussed in Section 4.4 below.

Another consequence of this reasoning is that it offers the chance of laying to rest some of the inconsistencies of standard phonetic theory, especially those that attempt to separate 'levels' of phonetic analysis. If the representation of speech and language is based on distributed, multimodal, dynamically-organised memory traces, then the same sensory information can be implicated in more than one class of memory, or more than one use to which memories are put. Thus, as described above, systematic fine phonetic detail simultaneously contributes segmental (allophonic) and prosodic information (e.g. Fougeron & Keating 1997; Smith & Hawkins 2000; Keating *et al.* in press). Perhaps most significantly, we can re-evaluate the traditional separation of speaker and message—of voice quality from segmental from prosodic information—as normally undesirable. (Equally, separating them is unproblematic if it is done for practical reasons.) Fundamental frequency, for example, can feed functional groupings of brain cells that themselves combine with other such functional groupings to produce complex memories of both speaker and message. Once this is acknowledged, 'message' can be interpreted as far more than simple phonological form and lexical item: it can encompass the whole range of sociophonetics, together with phonetic markers that facilitate successful conversation, such as slowing rate and adding creak to indicate the end of a conversational turn. In short, most information available from the spoken signal is truly polysystemic. We have long known that this must be the case, and that some languages (for example tone languages like Mandarin and Burmese) are probably better analysed in this way. We lacked the independent empirical support for its psychological reality, which the neuropsychological evidence now provides. In sum, we suggest that this approach allows phonetics to take a central place within a broad theory of the communication of meaning, rather than being seen as an arbitrary and somewhat inconsequential carrier of meaning. This argument for phonetics parallels MacWhinney's (1999) approach to how we understand sentence meaning.

4.3. Acquisition

Our view of how babies learn to understand speech is more or less dictated by our position on broadly-defined adaptation effects and memory and the polysystemic, non-segmental principles of Polysp, although it is fairer to say that it was evidence from infant speech perception that influenced our thinking on the other areas, rather than *vice versa*. Following Jusczyk (1993, 1997) we assume that children exploit statistical properties of spoken language to 'bootstrap' their way into understanding speech. These distributional regularities are inherently context-sensitive and thus relational, and can arise at any level of formal linguistic analysis; Jusczyk's work has identified prosody and phonotactics (i.e. rhythm/intonation, and sequential sound dependencies) as early influences on learning to segment words in the first year of life (Jusczyk *et al.* 1999a, b). The process is gradual, and relies on picking out salient properties of the signal that make discernible patterns in particular contexts, building up mental categories from these features, and using those early categories to further systematize the structure of sound and meaning with reference to categories that are already established. This model is broadly compatible with Plaut & Kello's (1999) model in which phonetic input is mapped directly to meaning, and with a number of other models of how babies learn phonology (e.g. Suomi 1993). The approach is also compatible with models of how children learn grammar from distributional regularities in the input, without recourse to innate mental constructs that specifically concern grammatical relations (e.g. Rumelhart & McClelland 1986; Allen & Seidenberg 1999). Our general view has been beautifully put by Smith in the context of how children learn about nouns: " … word learning biases that constrain and propel learning in certain directions are themselves made out of general associative and attentional processes. Each new word learned by a child changes what that child knows about learning words—adding to, strengthening, weakening associations among linguistic contexts and attention to object properties. In the end, word learning looks special and predestined. But specialness and destiny are themselves made out of more ordinary stuff." (1999:301).

4.4. Linguistic categories

The foregoing discussion makes it clear that we see linguistic categories as (1) self-organising, (2) multimodal and distributed within the brain, (3) dynamic, and (4) context-sensitive (or relational) and therefore plastic, or labile. We suggest that speech sound categories have these same properties. A familiar way to conceptualise this is in terms of polysystemic hierarchies such as those Polysp provides.

4.4.1. Self-organising, multimodal and distributed

The assumption of self-organising categories is consistent with current computational models of acquisition such as those mentioned in Section 4.3 as well as those discussed in Section 5.5 below. Self-organising categories also allow continuous adaptation to new information, as proposed in Section 4.1. Categories 'emerge' as a result of the distributional regularities of the input experienced, modulated by attention and the task(s) at hand, because organization of sensory information into coherent groups of 'like' and 'unlike' is a fundamental property of how organisms interact with their environment.

Speech categories are multimodal and distributed within the brain because they rely on memory, which is inherently multimodal. The association of meaning with speech learned under normal communicative circumstances is likewise multimodal. Both Pulvermüller (1999) and Coleman (1998, ms) have proposed detailed arguments consistent with this idea.

Pulvermüller's (1999) model of how the brain encodes meaning for words postulates that word representations typically have two parts, a perisylvian part related to the word form, and thus to phonetics/phonology, and a part distributed mainly in other parts of the brain that represents its semantic word properties, and thus reflects the individual's experience with that word, as described in Section 4.2 and illustrated schematically in Figure 3.

Pulvermüller's model is based on associative learning and in particular the principles of complex, functional cell assemblies proposed by Hebb (1949), modified by more recent research in neurophysiology and functional brain imaging. Hebb's basic principles of functional groupings of cell assemblies have parallels with the coordinative structures proposed for control of skilled movement (Kelso *et al.* 1979; Saltzman & Kelso 1987; Saltzman & Munhall 1989) that are axiomatic in articulatory phonology (Browman & Goldstein 1989, 1990, 1992) and may be more familiar to linguists and phoneticians. (Coordinative structures are functional synergies between groups of articulators.) Hebb's functional cell assemblies are also broadly compatible with the affordances of direct realism, and MacWhinney's (1999) general theory of embodiment (Section 4.2). Moreover, Hebb's formulation includes Gestalt principles, which have been incorporated into some models of hearing, notably by Bregman (1990). Remez and colleagues have argued that Gestalt principles do not fit phonetic evidence, and that phonetic categories are inherently abstract, but, while we acknowledge the value of many of their arguments, we are not convinced that wholesale rejection of Gestalt principles is necessary (cf. also Fixmer 2001). There is considerable appeal in finding parallels in the organisation of categories in phonetics, grammar, and meaning, if at the same time they can be externally motivated, in this case by being linked to neurolinguistic data. It is thus worth looking at Pulvermüller's arguments in more detail.

In Pulvermüller's model, a cell assembly is essentially the func-

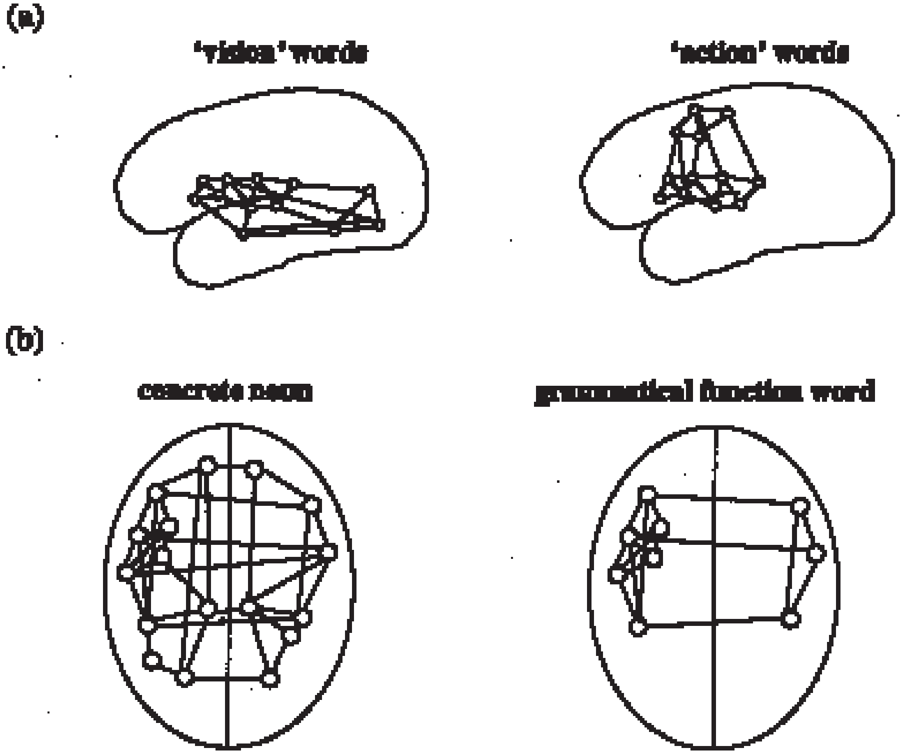

Figure 3. Schematic diagrams of the brain showing representations of the type of functional cell groupings that may represent different types of spoken words. (a) View of the left cerebral hemisphere showing distributions of activity during processing of vision and action words. (b) View from above (left hemisphere on the left), showing different degrees of lateralisation in the two hemispheres for content and function words. Adapted from Pulvermüller (1999).

tional physical substrate of the computational scientist's emergent category. All mental categories are emergent because all are groups of neurons that develop functional unity due to persistent stimulation of the same groups of cells. That is, cell assemblies form when similar sensory episodes are repeated sufficiently often. Once part of a cell assembly, connections between the cells that comprise the assembly become more easily activated, and may continue for longer in a self-perpetuating process of excitation called reverberation. Initial activation (ignition) of all cells in the assembly is more or less simultaneous, and is followed by reverberation, which defines the dynamic activity pattern of the cells. Thus an activated cell assembly is a dynamic, spatiotemporal pattern of neural firing. The particular pattern of activity can in turn activate other cell assemblies.

Pulvermüller suggests that ‘synfire chains’ (Abeles 1982, 1991) provide support for the neurological reality of Hebbian cell assemblies. Although regarded with some scepticism when first proposed, synfire chains now seem to be accepted (cf. Arbib 1995) as spatio-temporal patterns of cortical neuronal firing that create wavelike patterned sequences of activity. What characterizes a synfire chain is that a subgroup of neurons fires synchronously, activating a second subgroup of (at least as many) neurons that in turn fire synchronously, and so on. Their timing is precise to 1-2 ms. They form networks that appear to reverberate, and amongst their interesting properties for speech and language understanding are that a given neuron can take part in more than one subgroup, and that connections can skip subgroups in the chain to directly link, say, subgroups 1 and 3 in the chain (Pulvermüller 1999:256). Abeles (1991:258) points out that groups of synfire chains are equivalent to multilayer perceptrons (essentially feedforward neural networks) in the sense that the first stage of all the synfire chains can be seen as equivalent to the input layer of the multilayer perceptron, the second stage as a hidden layer, and so on. Although each synfire chain acts only as a feedforward mechanism, synfire chains can be formed in a network that has feedback connections.

Pulvermüller suggests that entire words or morpheme strings might be laid down as synfire chains. Synfire chains might form the neural basis for Coleman’s (ms) proposal that words are represented in the brain as trajectories through space-state networks that relate auditory states which vary through time. These networks include spectral information and deal well with rate changes, although more work is needed to establish how to capture the correlation between rate and spectral change.

The application of this work to speech and language is exciting, but the thinking behind it is highly speculative at present, and many details are under debate both by Pulvermüller and by others (see commentaries on his paper in the same journal volume). For example, instead of a separate cell assembly for each word, a number of words of the same semantic category, such as those with similar meanings, or action words vs. vision words, could have the same cell assembly, but would be distinguished by the firing patterns within it. Alternatively, different functional structures with overlapping sets of cells seem equally plausible: an example is the semantic attributes of *crocodile* versus *alligator*, for which the colour neurons would differ somewhat (and presumably, for experts, some other cells, such as those reflecting aspects of shape).

However, there seems general agreement that there is good evidence that functional groupings of brain cells associate modality-specific sensory memories for words with concepts—that is, with meaning—and that some sort of word form representation is localised to the perisylvian area, particularly in the left hemisphere. For example, so-called function words are represented more strongly than content words in the left perisylvian area, and less strongly elsewhere, and it is argued that this could be because many of them are less intimately connected with sensory experience than, say, concrete nouns. Likewise, cells in the left perisylvian cortex are activated when meaningful words are processed, but not when pseudo-words are processed. Our interpretation of this last finding is that phonemes, as abstract entities, may not be represented in this region, for if they were, then phonotactically-acceptable pseudo-words should activate them. However, some aspects of phonological form do seem to be implicated in this area (cf. Coleman 1998).

There is less agreement about how these principles work for grammar and sentence meaning. Relatively simple cell assemblies may combine into more complex functional assemblies, by various types of connections between cell assemblies and groups of cell assemblies, and by their activation dynamics. Such links between functional groups of cells potentially offer the flexibility and possibility for nuance that characterise language and speech. Extrapolating from these views, we suggest that complex connections between assemblies could be involved even in recognising isolated words: as noted earlier, cross-modal priming studies indicate initial activation of all possible forms of an ambiguous word, with a relatively fast decay of those that are contextually inappropriate (Swinney 1979). Although much work remains to be done before these processes will be understood for either isolated words or words in context, the relevance to the hierarchical structures of standard linguistic theory is obvious.

The possibility of increasingly complex links between simple cell assemblies is obviously also relevant to the hierarchical structures of polysystemic, nonsegmental phonology and thus of Polysp. This interpretation is consistent with Coleman's views, and although it is not developed by Pulvermüller, to do so would require only a small shift in focus. Both Coleman and Pulvermüller see phonological representation as emergent categories associated with lexical representation and anatomically localised in the perisylvian area. In

Pulvermüller's terms, these categories could be cell assemblies, or parts of cell assemblies for lexical items realised as synfire chains. Coleman suggests that semantic and all manner of phonetic information are bound together and accessed almost simultaneously when words are being processed; phonemes play no part. Pulvermüller has the same interpretation except that he favours strings of phoneme-sized units sensitive to their segmental context and neglects the relationship between the physical phonetic signal and prosodic and grammatical structure. If the context-sensitivity included information about rich (polysystemic) linguistic structure, then the two views would be in agreement.

Coleman (1998) further argues that there is just one phonological lexical representation for speech, and it is mainly auditory rather than motoric. Cross-modal integration of auditory and visual sensory information is thought to take place in the superior temporal lobes, but though the visual information is about movement, it is not of course the actual movement required to produce speech; sight of movement, and haptic or kinaesthetic sensations of one's own or others' movements, are represented in different, modality-specific regions of the brain. Coleman suggests that, to speak, an individual must translate the auditory lexical representation into motor gestures. This has a number of interesting implications, of which the two most important for our present purposes concern phoneme monitoring experiments and long-domain resonance effects.

Coleman (1998, ms) points out that phoneme monitoring is almost the only single-word perceptual task known to activate Broca's area, which is associated with articulatory encoding. He concludes that articulatory encoding is involved in phoneme monitoring but not in long-term memory for words. He also summarises work by Démonet and colleagues showing that activation of perisylvian cortex in phoneme monitoring is much later than in the normal course of lexical access. This work suggests that decomposing a nonsense word into constituent phonemes is a different and slower task from accessing a lexical entry from auditory input. Démonet suggests a sequential mode when decomposition into phonemes is necessary, but a probabilistic, nonexhaustive way of processing lexical items in semantic tasks. These data lend independent support to our position that phoneme identification is neither a necessary nor a normal part of understanding speech.

Coleman's conclusions that lexical representations do not involve memory for articulation offer a pleasing way of describing how resonance effects and other forms of long-domain coarticulation can arise.

If the word is translated into articulator movement only after the individual words have been slotted into place, consistent with many models of speech production—cf. Harley (1995) for a textbook account—then the necessary actions can be planned as a coherent, efficient sequence according to the demands of the particular accent and language. This does not explain what constrains the spread of resonance effects, nor why listeners are able to use them to facilitate perceptual tasks. If, moreover, Pulvermüller is right in his view (1999:321) that articulatory as well as acoustic and semantic aspects of each word are bound together into one functional unit, then this explanation of how long-domain resonances arise does not work. Much more work is needed before these questions can be answered.

However, the idea that long-domain 'units' of linguistic analysis could emerge from repeated exposure to consistent resonance effects fits comfortably within Pulvermüller's general Hebbian approach and can in principle be accommodated within the rich polysystemic structures of Polysp. Presumably, such coherent long-domain effects could operate in a similar way to longer prosodic units and might be represented high up in linguistic structure. In this case the complexity and flexibility offered by synfire chains or similar processes may be sufficient to account for both local phonetic information and long-domain resonance effects. Local information could be mediated by one type of detailed pattern-matching, while a more general picture is sought at a more global level of analysis in which local phonetic detail might be effectively treated as noise.

Another possibly contrary view to Coleman's (1998) is provided by Rizzolatti & Arbib (1998) and Arbib (2000), who show that monkey cortex includes 'mirror neurons' that discharge both when the monkey grasps or otherwise manipulates an object itself, and when it sees someone else grasp or manipulate the object. The mirror neurons appear to link the observer and the actor by matching similar actions regardless of who makes them. Rizzolatti & Arbib suggest that such a system could be fundamental to successful communication. Mirror neurons offer striking relevance to infants' early sensitivity to the connection between facial expression and speech sounds (Kuhl & Meltzoff 1982) and to motor theories of speech perception, especially Gibsonian approaches in which affordances are fundamental. Those mirror neurons found so far, however, appear to be for activities that can actually be seen; since most activity required for speech cannot be seen or otherwise directly experienced by the listener, it is a big—though not inconceivable—step to assume that mirror neurons or their like underlie perception of speech.

In summary, we suggest that all linguistic categories, including phonetic ones, are constructed by the individual by organising memories of his or her own sensory experiences into functional groups that reflect frequently associated factors. Memories are axiomatically modality-specific, but the linguistic categories they underlie are multimodal because they relate sensory linguistic-phonetic experience to experience of the referents of language. Successively more complex or more all-inclusive groupings represent abstraction. Consistent with Polysp's polysystemic properties, individual language-relevant memories can be part of a large number of different functional groups, so a given piece of phonetic information may contribute to several quite distinct percepts. According to this view, attributes of voice quality can contribute to phonetic information about prosody, segmental identity, and discourse structure while simultaneously contributing to the percept of the speaker's identity and affecting the listener's attitudes. For example, if the listener likes the speaker, then he or she will probably have a positive attitude towards other speakers who have similar voices, regional accents, rhythmic patterns and so on. Supportive data are provided, for example, by Johnson *et al.* (1999).

4.4.2. Dynamic, context-sensitive (relational) and therefore plastic

The case for dynamic and context-sensitive phonetic categories needs no reiteration, but comments in the recent literature suggest it is worth discussing the claim that plasticity (or lability) is a fundamental property of phonetic categories. We offer both a logical argument and empirical evidence. The logical argument for plasticity of phonetic categories is that all mental representations of categories are plastic and context-sensitive, and there is no reason to suppose phonetic categories should be different. Birds are often identified by their silhouette, yet the prototypical blackbird is expected to have a plumper silhouette when you identify it on a lawn covered in snow compared with one bathed in sunshine, because it will have fluffed its feathers out to keep warm. Thus, in identifying the species of a particular bird, the absolute criteria for shape will change according to the perceived ambient temperature, and therefore the relative importance of shape and other criterial attributes such as colouring and beak size may also change. Just so with speech: the analogy with stimulus range and trading relations between multiple, apparently redundant, acoustic cues is obvious. In other words, in terms of perception, phonetic categories are no less contrastive than phonological ones. To have a contrast, there must be a context. If the context changes, then the relative perceptual salience of the identifying physical attributes of the phonetic category will probably also change.

Figure 4 illustrates these principles. A 'stick person' without hair is normally taken to be male, as in the top left of the figure. Add hair, as at the top right, and the figure is conventionally taken to represent a female (especially if both figures appear together). But this convention only works in the context of stick people, for many men have long hair. Hair length is irrelevant in deciding the gender of the two people in the photograph at the bottom of the figure. Instead, one must focus on quite different attributes, such as bone and muscle structure of the face, arms and hands, possibly the type and distribution of tattoos, and so on. Thus a property that completely defines class membership in one context is completely unhelpful in another. This figure has a message for speech perception methodology, in particular the issue of what constitutes appropriate experimental control: when synthetic stimuli are stylized and strongly over-simplified, or natural utterances are cross-spliced to produce unnatural combinations, there is a real danger that the relative perceptual salience of particular acoustic properties may be misinterpreted because their necessary context is missing or distorted.

The empirical evidence for plasticity comes from a number of sources. One is the adaptation effects discussed in Section 4.1. Another, and amongst the most interesting, is Coleman's (1998) interpretation from neurological and other evidence that lexical representations are phonetic, for lexical representations are by definition contrastive. This is supported by work we have reported elsewhere that suggests that systematic fine acoustic-phonetic detail in natural speech facilitates lexical decisions. Hawkins & Nguyen (2000, in press) have shown that in several varieties of British English, longer and phonetically darker word-onset /l/ tends to cooccur with voiced coda obstruents, and that listeners can exploit this knowledge in lexical decision and lexical identification tasks. As noted above (Section 4.1.3), work in progress suggests that listeners may be very quick to assess whether this type of acoustic information is sufficiently systematic to use as a perceptual cue (Hawkins & Nguyen 2001). These data are consistent with dynamic models of perception in which new exemplars have a large effect on category boundaries, as discussed in Nguyen & Hawkins (1999).

Our final example concerns the perceptual magnet effect (PME: Kuhl 1992; Iverson & Kuhl 1995, 1996; Kuhl & Iverson 1995). Kuhl coined this term to describe experimental findings that were interpreted as showing that listeners' experience with the sounds of a par-

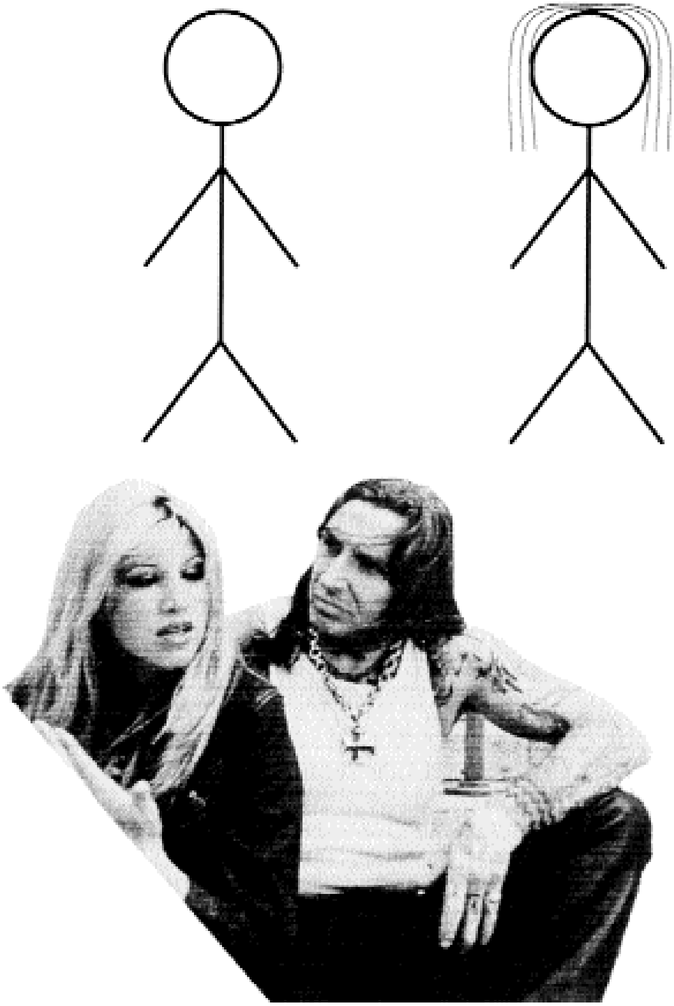

Figure 4. Top left: A stick figure of a person. It has no distinguishing characteristics and might be interpreted as representing a man or a woman. Top right: when long hair is added to the same stick figure, the standard interpretation is that the one on the left is male and the one on the right is female. Hair length is criterial of gender for stick people. The photograph below shows that hair length is irrelevant to classifying the gender of real people, at least when they are hippies, whereas attributes such as bone and muscle structure are important. The stick figures are analogous to highly controlled synthetic stimuli, and the photograph to natural speech.

ticular language causes poorer discrimination around the best exemplar of a phonetic category in that language than around a poor exemplar of the category (i.e. one that falls nearer the boundary with another phoneme). She interpreted this loss in discrimination as a reorganisation or distortion of psychoacoustic space, such that some sort of prototype develops, towards which physically similar stimuli are drawn as if by a magnet. For a number of reasons, theoretical and methodological, this interpretation has generated a lot of interest, with many researchers' views being strongly polarised for or against the PME. The aspect of this literature that is relevant to the present discussion is that much of the negative criticism of the PME, and of Kuhl's rather theory-neutral use of the term prototype, seems to be predicated on the assumption that the mental representation of phonemes (or phoneme-like phonetic segments) must be invariant. For example, one criticism of Kuhl's work made by Lotto *et al.* (1998) is that her proposed prototypes are pliable (their word, which we take to be the equivalent of our use of the word plastic). Lotto *et al.* (1998:3654) note that context can change the phonemic label a listener assigns to a given stimulus, and suggest that "category structure is a result of the input distribution itself and its relation to other categories", but they seem to expect that, for a phonetic prototype to exert a perceptual magnet effect, it must have a stable mental representation which cannot be easily manipulated. Neither we nor Kuhl dispute the first two points (e.g. Iverson & Kuhl 1995:560), but we do not think mental representations of phonetic categories must necessarily be acoustically stable. For example, Kuhl and Iverson (1995:146) suggest that prototypes might differ for gender.

One problem with interpreting these arguments is that the literature on the PME does not make it clear whether a so-called phonetic category is to be thought of as a phoneme. Our impression is that this issue has not been thought through: for example, the names of so-called phonetic categories are consistently given between slashes rather than square brackets, thus /i/ rather than [i], and most work has been done on isolated vowels. An isolated vowel, whose immediate phonetic context is silence, is arguably closest to the canonical quality of a phoneme, should one exist; but isolated vowels make a negligible contribution to real speech. In fact, experiments by Barrett (1997; Barrett & Hawkins in preparation) confirm that Kuhl's proposed phonetic prototypes must be context-sensitive. Barrett showed that the best exemplars of synthetic /u/ vowels in /u/, /lu/ and /ju/ syllables must have different F2 frequencies if they are to sound natural, and that each best exemplar syllable produces a magnet effect that is specific to that syllable. The same F2 frequency in one of the other two contexts does not produce a magnet effect, but instead acts like a nonprototype. Notice that Barrett's findings are consistent with our views on perceptual coherence, as well as plasticity of linguistic-phonetic categories.

In other experiments, both Barrett (1997) and Guenther *et al.* (1999) showed that the PME can be reversed: given an appropriate task, discrimination is increased around the best exemplar of a category and the magnet acts as a repellor rather than an attractor. For example, musicians who need to tune their instruments need to discriminate especially finely around the best (most in-tune) example of a particular note or chord. In contrast, speech understanding is presumably helped most by discriminating between but not within relevant categories.

In sum, no matter what theoretical status one wishes to give to the PME, the evidence is clear that the behavioural effect is demonstrable. As Guenther and Gjaja (1996) point out, logic, as well as neuroscientific and computational evidence, suggest that there is no need to postulate an actual prototypical representation in the brain, in that the same effect can be produced by representing memory of many similar instances and fewer more dissimilar ones, together with consideration of the demands of the task at hand. Equally, however, and central to our current argument, plasticity is fundamental to the concept.

4.4.3. Summary: The nature of phonetic categories

The neuropsychological research summarised above suggests that rich, context-sensitive structure may more nearly resemble the way words (and probably all aspects of language) are represented in the brain than the abstract, more independent strands that characterize much phonological-phonetic analysis. In other words, no category can be described independently of its context, and though the contexts can be combined to produce an apparently infinite range of finely-graded phonetic detail, the rules by which they are combined are relatively constrained. To extend the analogy of a blackbird seen in snow or sunshine: ornithologists learn a set of criteria necessary to identify it in either weather, from observation or explicit instruction; with experience, they develop a finely-tuned sense of how to weight the various criteria in different circumstances (contexts) so that the label *blackbird* is accurately used despite superficial differences. Although this knowledge may be gained faster with explicit instruction, what distinguishes the novice from the expert is that the expert has experience and knows what to attend to, which presumably allows rich and nuanced structuring of mental categories.

Phonetic categories, then, seem well conceptualised at present as part of a hierarchical structure of emergent categories, formed from repeated associations of meaning with patterns of experienced sound and other relevant sensations. To the extent that formal phonological and grammatical structure is represented in the brain, it is a by-product of this type of organisation, the union or intersection of related dynamic links. What is actually represented at the neural level depends on what tasks the individual habitually carries out; other aspects of linguistic analysis can be arrived at, but must be computed on the fly, possibly after some particular sensory input and its meaning have been linked. This view means that no two individuals will necessarily have identical phonetic categories.

5. Attributes of Polysp in relation to some recent models of perception

A number of recent models of speech understanding offer promising new approaches to some of the properties that we have identified as important, although no single extant model deals with them all. In this final section of the paper, we try to identify models whose strengths reflect views compatible with ours. This might help us, or others who agree with us, to combine the various strengths into a single, more comprehensive system. Some of the issues have been addressed in detail in the context of many different models, and we do not have space here for a full review of all the contributions that have been made. Instead, for several of the properties outlined in Sections 3 and 4, we discuss in some detail the one or two models that best address that property, making only brief reference to other related or contrasting approaches.

Most of the models we favour have in common that they are self-organising, in that the perceptually relevant units are not specified in advance but rather emerge dynamically in the course of processing information. In other respects, the various models we discuss differ computationally, and their processing assumptions may even contradict one another: something that is treated as a process in one model may be part of the architecture of another. These issues need not usually concern us, for the theoretical position we are developing here does not depend crucially on whether some important property is modelled as a process or built in to the architecture of a model. Indeed this is an issue to which phonetics may not in general contribute a great deal, inasmuch as the status of memory and concepts is ultimately a question for neuroscience and molecular biology, or alternatively for philosophy.

5.1. Modelling perception of temporally distributed information

We have proposed that a model of speech understanding needs a mechanism for identifying both short-domain and long-domain events in the speech signal. It must be sensitive both to weak, subtle information as well as to clear information, bearing in mind that there are complex dependencies between the informational value of phonetic properties and the time domains over which they occur. In Polysp, the aim is not necessarily to arrive at abstract, time-free linguistic representations as soon as possible, but rather to allow the temporal structure of the speech signal to play a determining role in the course of processing. Rate variation, for instance, should not be modelled as time-warping of a sequence of underlying elements, but rather as a potentially meaningful restructuring of the temporal organisation of an utterance, which may result in an apparent reorganisation of segmental quality.

Most computational psycholinguistic models are neural networks, and developing any kind of internal representation of time in neural networks presents its own challenges (see Protopapas 1999 for a review). Perhaps as a consequence, comparatively little attention has yet been paid in neural net models of spoken word understanding to the specific constraints imposed by the temporal structure of phonetic events.

Short-domain events are comparatively easy to handle computationally given a sufficiently fine temporal resolution, such as spectra sampled at 10 ms intervals (cf. Klatt 1979; Johnson 1997; McClelland & Elman 1986), though 5 ms might be more realistic. The identification and integration of information that becomes available over longer time scales poses more of a problem, and certain kinds of model, such as Simple Recurrent Networks (e.g. Elman 1990; Norris 1992), have great difficulty integrating information over long time scales. The reason is that relationships between non-adjacent acoustic segments are subject to what can be termed a degrees-of-freedom problem. Local coarticulation and local aspects of rate variation can be modelled fairly easily because there are few degrees of freedom: that is, a property could be assigned to the current segment, the preceding segment or the following segment. But with greater distance between a segment containing some acoustic property and the segment that gives rise to that property (e.g. 6 phonemes if the first vowel in *it's a berry* exhibits /r/-

colouring), the degrees of freedom required to correctly interpret that property increase substantially, at the expense of the model's explanatory power. Thus acoustic cues distributed over stretches of the signal as long or longer than a syllable may prove difficult to capture using recurrent neural networks (Nguyen & Hawkins 1999).

The difficulty with long-domain events can be attributed in part to the fact that word recognition is still often assumed to be a process which entails partitioning the signal into segments. A related problem is that most models take as the INPUT to speech processing either a series of static spectra (Klatt 1979; Johnson 1997) or auditory parameter values at successive time slices (e.g. Plaut & Kello 1999). Phonetic information is assumed simply to accumulate from one time slice to the next. Implicit in this view is the idea that speech processing proceeds uniformly (because the signal undergoes the same analysis process at each time step) and thus that processing responds only passively to the fact that the speech signal unfolds in time. Against this background, the integration of information forwards and backwards in time is bound to appear problematic.

In contrast, we suggest that processing itself is modulated or driven by the temporal nature of the speech signal, including its rhythm. That is, that the occurrence of particular acoustic properties, or the rate at which acoustic events occur, influences how subsequent information is processed. This is by no means an original idea: similar views are expressed by Lehiste (1972), by Whalen (2000), in Stevens' concept of landmarks (Section), and in Cutler & Norris's (1988) Metrical Segmentation Strategy, where strong syllables trigger lexical access attempts, and thus prosody is a major determinant of word segmentation.

The idea that temporal properties might drive speech processing has not been fully worked out, and hence is poorly implemented or unimplemented in computational models of speech understanding. However, recent developments in the class of computational models developed in the context of Adaptive Resonance Theory (ART; Grossberg 1986) are promising in this respect. They are PHONET (Boardman *et al.* 1999), ARTPHONE (Grossberg *et al.* 1997), and ARTWORD (Grossberg & Myers 2000).

PHONET is designed to model phonetic context effects such as the effects of speech rate on consonant perception in a synthetic continuum between /ba/ and /wa/, and is applicable to speech rate variation in general. In outline, the model features separate, parallel auditory streams, one of which responds to transient and the other to sustained properties of the speech signal (e.g. formant transitions versus steady states); they store their inputs in parallel working memories. The model neurons in the transient channel operate together to detect rapid events like bursts and formant transitions, preserving information about the frequency, rate and extent of each transition: each is sensitive to frequency changes in its own particular narrow band of frequencies, and is subject to lateral inhibition so that their combined response reflects the relative degree of excitation in each frequency band. Activation in the transient channel is increased by a faster rate of spectral change, and decreased by a broader frequency extent.

The activation in the transient stream can modify the processing rate in the sustained stream, because greater activation in the transient stream (e.g. by faster formant transitions) increases the processing rate in the sustained channel. This cross-stream interaction ensures that the ratio of activation levels between the transient and sustained channels is maintained across syllables spoken at different rates. The logic is that, if there were no interaction between the channels, then when a CV syllable was spoken fast, it could produce a greater response in the transient channel, but no accompanying faster processing in the sustained channel, so that the ratio of transition-to-steady state durations would be different at different rates. Instead, the processing rate in the sustained channel is speeded up when the transient channel indicates that transitions are faster. The output is a ratio of transient-channel activation to sustained-channel activation, which is assumed to be mapped onto phonetic feature categories via a self-organising feature map.

The consequence is an invariant perceptual response for invariant relative durations; in this case, a relatively rate-invariant representation in working memory of the /b/-/w/ contrast. Since learning in the ‘adaptive filters’ of Adaptive Resonance Theory encodes the *ratio* of activations across working memory, phonetic categories in turn come to encode the ratios of sustained to transient information: a phonetic category boundary is enshrined in relative terms as relationships between distinct ‘events’, taking the whole context into account.

This appealing solution to an old problem has been tested on highly controlled synthetic speech. In natural speech, the ratios of transitions to steady states is not constant, and so some changes might be required. This might not prove too difficult. For example, even in simple CV syllables, rate changes are normally accompanied by changes in vowel quality, with concomitant changes in transition trajectories. Dynamic spectral differences of this type could presumably be related to the type of spectral difference Stevens uses to distinguish onset /b/ and /w/ (how abruptly the amplitude and spectral tilt change in the vicinity of the opening). Presumably the final decision should be based on sensitivity to a complex mix of phonetic detail.

ARTPHONE and ARTWORD do not explicitly consider the integration of acoustic information over time. The assumption is that this has already been performed in a prior stage, to produce what are termed 'phonemic item representations'. That is, in ART, speech input activates 'items' which are composed of feature clusters. Items in turn activate 'list chunks' in short-term memory, corresponding to possible groupings of features, such as segments, syllables, and words. Disappointingly from our point of view, it is only this latter process, with phonemic items as its starting point, which is modelled in ARTPHONE and ARTWORD. Even so, because ART models operate in real time, phonetic temporal structure still plays a part in these models, as we outline below.

Adaptive Resonance Theory distinguishes the rate of external input and various kinds of internal processing rate. Particularly important, as the name suggests, is the notion of resonance. Resonance is a positive feedback loop between items in working memory and list chunks. It involves non-specific top-down inhibition, and specific top-down confirmation of expected items. When listeners perceive speech, a wave of resonant activity plays across working memory, binding the phonemic items into larger language units and raising them into the listener's conscious perception.

The time scale of conscious speech is not equal to the time scale of bottom-up processing (the resonance evolves more slowly than working memory activation) nor to the rate of external input. The resonant dynamics include resonance 'reset', which prevents a resonance from continuing indefinitely, and may be triggered either by bottom-up mismatch, or by the resonance self-terminating. Self-termination is called 'habituative collapse', and represents the gradual cessation of resonance as synaptic neurotransmitters habituate to the same stimulus, so that they gradually stop transferring excitation between working memory and stored representations, and resonance self-terminates. An example is the decay of resonance at the end of an utterance because there is no new stimulation. Grossberg *et al.* (1997) suggest that geminates are perceived as two consonants because one resonance self-terminates and another begins, and that this is why the closure has to be longer for intervocalic geminates than for stop clusters with different places of articulation. Within a narrow temporal window, it may also be possible for a resonance to transfer seamlessly from one, smaller list chunk to another, larger chunk. This can happen if new information arrives while the support for the smaller chunk is beginning to weaken due to habituation (i.e. as the resonance is winding down), but before that chunk's activation levels have fallen too low. Notice the congruence of these model processes with those of Pulvermüller (1999).

The nature of resonant dynamics in the ART models allows fine phonetic detail to play an important role in the grouping of items into list chunks that correspond to larger language units. The way this happens relates to a distinction drawn by Grossberg & Myers (2000) and Mattys (1997) between two kinds of phonetic evidence. One is informational*,* contributing to mapping of the acoustic stimulus into 'phonemic items', often over quite extended time periods; it is only implemented in PHONET. The second type, durational evidence, relates to the time of arrival of information such as silence, noise, and so on. Durational evidence can affect processing in a global way, because the time of arrival of a particular piece of information at a particular processing stage influences the competition between activated list chunks and the development of a resonance.

These ideas are illustrated in ARTPHONE by simulating the emergence of percepts of geminate stops versus two phonetically distinct stops. ARTWORD (Grossberg & Myers 2000) uses similar principles to simulate data on the contribution of silence and noise durations to percepts of *grey ship, great ship, grey chip* and *great chip* (Repp *et al*. 1978). Take, for example, the way in which increasing fricative noise duration produces a transition between percepts of *grey chip* and *great ship.* Both signals contain enough information for the perception of a stop-like sound, but the acoustic information groups differently, to yield a /tʃ/ percept in the former case, and a /t/ percept in the latter. The development of this grouping involves competition, which emerges at a slow enough rate to allow the initial competition between *grey* and *great* to be influenced by the later-occurring noise and the new competition it engenders between *great* and *chip*. When evidence for /t/ is strong, at low noise durations, the chunks corresponding to *grey* and *chip* competitively team against *great.* At longer noise durations, the excitation of /ʃ/ has more time to develop, and is proportionally greater, so *ship* out-competes *chip*, thereby allowing *great* to receive greater levels of activation than *grey.*

In general, we are strongly sympathetic to the way in which ART models allow the temporal properties of the speech signal to affect the way the signal is processed. This contrasts with the approach taken in many models, where analysis of the signal is identical at each time step and phonetic information merely accumulates over time. Its principles bear some resemblance to those of the FLMP (Massaro 1998), although the two classes of model use different mathematical procedures. The fact that phonetic categories come to encode relational information is also very much in tune with our views. We do not have space here to discuss the other appealing properties of ART, which include its account of learning, its relative neural plausibility, and its very wide application in areas that include visual perception, memory and attention.

One less appealing property of ART models is that phonemic representations are obligatory, which at first sight seems incompatible with Polysp. We suspect, though, that the problem is fairly superficial. Because the range of contrasts examined so far is small, the objects of the simulations could as well be termed allophones as phonemes. More importantly, information is not thrown away when phonemic items are identified, because fine phonetic detail is passed up to subsequent stages as durational evidence. At the least, these models evince a strong sensitivity to syllable structure. In consequence, we do not view the ART approach as incompatible with non-segmental phonological representations, especially since the ART constructs of items and list chunks are not fixed responses corresponding to particular linguistic levels, but instead represent attractors of various sizes (Grossberg *et al.* 1997). Indeed, in typical ART resonance, longer chunks (e.g. words) mask smaller chunks (e.g. phonemes), so the largest coherent unit constitutes the focus of attention (Luce *et al.* 2000). This might be one way of dealing with the perceptual cueing function of long-domain phonetic resonance effects such as those described in Section 3.4.2.

We suspect, therefore, that the difference between Grossberg's and our views of phonemes is more terminological than substantive. This issue could become clearer if ART models investigated natural rather than synthetic speech, because in natural speech (unlike the synthetic stimuli in the studies simulated) spectral detail co-varies along with durational factors, as outlined above for rate.

We would also welcome a focus on long-domain properties other than speech rate, such as nasality, lip rounding, vowel-to-vowel coarticulation, and, at least for English, resonances due to liquids. If such longer-domain resonance effects had a status in the model equivalent to a Hebbian cell assembly, then a word containing /r/ might have its own cell assembly, and also an associated cell assembly that represented the resonance effect, which develops because the word has been associated with the long-domain resonance effect many times. When a putative /r/-resonance effect is detected, it would presumably raise activation levels of words containing /r/ in anticipation that one of them will be heard. So, when a particular word containing /r/ is heard, its activation level rises dramatically because it is congruent with the longer-domain resonance effect within which it occurs. In a broad sense, modelling these effects could be a first step towards making ART-type models explicitly polysystemic.

5.2. Modelling the percept of rhythm

We have suggested that a speech understanding model needs to allow local and sometimes subtle timing changes to influence the overall perceived rhythm of an utterance and the prominence of units within it, in ways which may convey grammatical, affective or other meaning. Such a mechanism might be naturally incorporated into future ART systems, given the attention paid to timing in these models. However, we also think a model should take into account the more general fact that people have a SENSE of rhythm in speech and other domains, especially since this may be an important bootstrapping tool for prelinguistic infants (for an overview, see Jusczyk 1997:ch. 6). No current speech understanding models can adequately explain the percept of rhythmic structure, but there do exist general accounts outside of speech research. One of the most appealing, for models of speech understanding, is Dynamic Attending Theory.

Dynamic Attending Theory (Jones 1976; Large & Jones 1999) addresses the puzzle that there exist all manner of dynamic events in which a clear temporal structure is apparent, yet the main beats are not isochronous. One example is the sound of a horse's hoof beats as it gallops faster and faster. People appear to apprehend stable rhythmic structures in such events, and they can perceive as meaningful the fluctuations in the periodicities that compose them, much as we have argued with regard to rate variation in speech. Dynamic Attending Theory explains this behaviour in a broadly Gibsonian framework, by postulating 'attending rhythms,' or internal oscillations which are capable of entraining to external events and targeting attentional energy to expected points in time.

The mathematical formulation of dynamic attending theory employs two entities: external rhythms and internal (attending) rhythms. External rhythms involve a sequence of related environmental events whose onsets can be localizable in time and which therefore define a sequence of time intervals. An example is the successive sounds of the horse's hoofs. An attending rhythm is internal to the perceiver. It is a 'self-sustaining oscillation' which can be thought of as generating an expectation.

When the individual attends to an environmental (external) rhythm, the phase and the period of the internal attending rhythm's oscillations become coupled, via entrainment, to those of the external rhythm, creating stable attractor states. The period of the oscillation reflects the rhythmic rate, or overall tempo, while the phase relationship between the coupled external and internal attending rhythms expresses the listener's expectation about when an (external) onset, or beat, should happen.

Changes in the beat are dealt with as follows. When the external rhythm is momentarily perturbed, then the phase relationship becomes desynchronised, but it can be adjusted to its original phase when the perturbation is not large. So when the only perturbations in an ongoing external rhythm are random and relatively minor, entrainment can continue through small phase adjustments of the attending rhythm. However, when the external rhythm changes to a new overall rate, the attending rhythm must change its period rather than continually adjust its phase. If the period does not change to reflect the new rate, then all succeeding external onsets will be expected at the wrong time and heard as either late or early, because the phase relationship will be wrong. Thus, while period coupling captures the overall rate, phase coupling prevents synchrony between the external and attending rhythms from being lost when the external rhythm is momentarily perturbed, and can capture the disparity between an expected onset and an actual onset.

Within each cycle of an attentional rhythm, the allocation of attentional energy is also modelled with a variable termed 'focus of attention'. There is a pulse of attentional energy, whose locus is determined by phase and whose extent is determined by focus of attention. It contributes an expectancy region where attentional energy is nonzero: a narrow pulse reflects an expectation that an event will take place within a narrow time range. A broader pulse reflects greater uncertainty about when an external event will happen. Focus allows the model to make predictions about the noticeability of time-changes. For instance, a large deviation from expectancy is more likely to be noticed than a small one, and any deviation is more likely to be noticed if attention is focused narrowly rather than broadly. The distinguishing principle is the same as the distinction between narrow-band and wide-band spectrograms. Wide-band spectrograms are made with a filter that has a shorter time window than that used to make a narrow-band spectrogram. Because the window is short, it allows more precise temporal definition—which is why wide-band spectrograms have sharper 'edges' between acoustic segments than narrow-band spectrograms.

Since most external rhythms are complex, involving ratio relationships among the phases and periods of their different temporal components, the model likewise employs multiple attending rhythms which are coupled to one another to preserve phase and period relationships. These relationships offer the potential for dynamically expressing relational information about the complex rhythm's underlying temporal form. Evidence from maturational changes in motor tapping and tempo discrimination is interpreted in the context of the model to suggest that the coupling of multiple oscillators develops with age and experience (Drake *et al.* 2000).

Among the appealing properties of Dynamic Attending Theory for our purposes are its explicit rejection of the view that temporal events like speech are time-warped sequences of underlying discrete elements whose serial order is of paramount importance. Instead, like ART, Dynamic Attending Theory incorporates a notion of internal psychological timescales which attune to those in the external environment. This broadly Gibsonian view is compatible with contemporary understanding of cyclical behaviour in biology, such as adaptation of circadian cycles to daylight hours (e.g. Sawyer *et al.* 1997; Kyriacou in press). As a result, both Dynamic Attending Theory and ART allow a prominent role for expectations and for informed (i.e. systematic, knowledge-driven) fluctuations in degree and focus of attention. This contrasts with psycholinguistic models which minimise the role of feedback (e.g. Norris *et al.* 2000), because, while expectations about some rhythms might not involve knowledge about the world, it seems to us that expectations about speech and many other rhythms (including music) must involve knowledge. The percept of rhythm in music is at least partly culturally determined (e.g. Stobart & Cross 2000), and the percept of rhythm in speech is at least partly language-specific.

Another appealing property of Dynamic Attending Theory is its simultaneous focus on different time windows, each suitably precise for the domain. Links between focus and expectations suggest that the better the listener knows the speaker, for example, the better he or she can predict upcoming events, and hence the faster and easier it will be to understand the message. The model thus has particularly interesting implications for making maximally efficient use of cues to segmental identity that are found in rhythmically predictable locations, such as landmark features and cues to place of articulation at the edges of syllables. When you expect these events, you adjust the focus of attention to a suitable time window to allow you to find them. When you do not expect or need them, you can transfer attention elsewhere.

Given the complexity of speech rhythms and their interactions with linguistic units of various kinds, we do not know to what extent Dynamic Attending Theory and mathematical formulations thereof could be adopted wholesale to form the basis of a speech understanding model. Its insights may be best viewed as a guiding principle showing what can be achieved when the idea of rhythmically-driven attention is explicitly built into a model. However, its similarities with principles used in other speech models such as Tuller *et al.*'s (1994) dynamical model of speech sound categorisation, described in Section 5.4. below, encourage us to suggest that principles used in Dynamic Attending Theory have a promising future in a polysystemic model of speech understanding.

5.3. Modelling adaptation to new speakers and rates: episodic representation

Adaptation to new speakers and rates is naturally accommodated in episodic theories of perception (Pisoni 1997b, Goldinger 1997) in which details of the speech signal that have traditionally been viewed as incidental are considered integral to the linguistic representation. Episodic theories see mental linguistic representation as an agglomeration of highly diverse language-related experiences, which is compatible with our ideas on multi-modal memory.

Exemplar models (e.g. Hintzman 1986, 1988; Nosofsky 1988) show the strongest commitment to episodic theories, retaining a separate memory trace for each stimulus experienced. When a new stimulus is matched against these multiple traces, it may excite anything from a highly specific response (as would be the case e.g. for an unusual word in a familiar or recently-heard voice), to an entirely generic one (e.g. to a common word in an unfamiliar voice). In these models, abstraction thus occurs dynamically during perception, rather than at storage.

There are growing numbers of exemplar models for speech, some of which have as a main objective the modelling of particular phonetic phenomena (e.g. Johnson 1997; Lacerda 1995), while others are not intended to model the details of the speech signal and use purely schematic, arbitrary representations of spoken words (e.g.

Goldinger's adaptation of Hintzman's MINERVA 2: Goldinger 1997, 2000). Both approaches are valuable, though for the long term we stress the importance of an input that is phonetically realistic. In this paper, however, we concentrate on the work of Goldinger, because his simplifications have allowed a clear focus to develop on the global contours of adaptation to speakers over time.

Goldinger (1997) found support for MINERVA 2's prediction that some stimuli excite a generic response, others a specific response. Subsequent investigations (Goldinger 2000) have mapped some of the task-specific ways in which these responses can change with repeated exposure to isolated word tokens. For instance, the effect of word frequency diminishes over time in a recognition memory test, but increases when the measure is imitation of a particular speech token. The interpretation is that the word frequency effect decreases in recognition memory tests because more traces have accumulated, which reduces the discrepancy between high- and low-frequency words, presumably specifically in this task. Conversely, it increases in imitation tasks because all the memory traces are of the particular token of the low-frequency word, so that production comes to reflect that token more and more closely.

So far, there has been little suggestion of how an appropriate set of episodic traces might be activated; as we have suggested, this process is unlikely to be as simple as a direct mapping of raw acoustic information onto stored traces, time-slice by time-slice, and indeed Goldinger (1997) emphasises that context will play a key role. Moreover, it is difficult to see how a simple exemplar-based matching process could convey some types of linguistic information, such as that provided by long-domain resonance effects: this type of acoustic-phonetic information may require some degree of abstraction in order to integrate it over long time domains (see Section 4.4.1 above). In any case, Goldinger's rationale for working with MINERVA 2 is primarily that, since it is a strict exemplar model, predictions derived from it offer a very stringent test of the extent to which episodic memory is implicated in linguistic processing. Since the model does predict a number of experimental results, it does strongly suggest that episodic memory plays a role in speech understanding. Of course, models that learn more abstract representations than MINERVA 2 can capture episodic memory for speech, provided that they learn in a flexible enough way that when the system needs to be able to make extremely fine discriminations, the details of individual exemplars can be retained. For instance, ART formalisms are compatible with Goldinger's ideas, since abstract prototypes and 'exemplar prototypes' can co-exist in these systems (Grossberg 2000b). Accordingly, even if some properties of speech turn out to be best explained not by a strict exemplar model, but perhaps by something more like ART, simulations with MINERVA 2 can make a useful contribution by showing where such a model should be constrained.

5.4. Modelling the emergence of linguistic categories

In the light of the overwhelming evidence that linguistic categories are plastic, a model should be able to produce appropriate changes in a particular phonetic percept as factors such as phonetic context and method of stimulus presentation vary. While many neural networks can simulate the influence of phonetic context, their success is usually due to training in which the network is told the correct categorisation for stimuli occurring in a variety of contexts. However, this type of training seems unlikely to be able to account for why repeated presentation of a stimulus may give different results depending on the order of presentation or the composition of the stimulus set.[2] One model which seems promising in this regard is the dynamical model of Tuller *et al.* (1994).

Tuller *et al.* (1994) note that speech sound categorization is nonlinear in that when a control parameter is varied over a wide range, often no observable change in behaviour results, but when a critical value is reached, behaviour may change qualitatively or discontinuously. (Note the congruence with quantal theory.) To investigate the dynamics of this process, they varied the duration of a silent gap between a natural utterance of /s/ and a synthetic vowel /eɪ/ in a *say-stay* continuum like those typically used in categorical perception experiments. When the stimuli were presented in sequential order (the duration of the silent gap either increasing then decreasing, or first decreasing then increasing), it was relatively rare for listeners to hear *say* switch to *stay* (with increasing gap duration) at the same gap duration as *stay* switched to *say* (with decreasing gap duration). Instead, and as expected, listeners usually switched either relatively late in the series (response perseveration, or hysteresis) or early (enhanced contrast).

The dynamical model developed to account for these well-known psychophysical context effects uses equations of motion to describe the temporal evolution of the perceptual process, especially factors affecting stability *versus* plasticity of perceptual classes. In broad outline, each perceptual category is represented as an attractor. The attractor's range of influence is represented as a 'basin of attraction' whose breadth and depth can vary with stimulus conditions. Deeper and narrower basins are associated with more stable percepts. An acoustic parameter (in this case the duration of the silent gap) influences the perceptual dynamics. When the value of this parameter is low, only a single attractor exists, corresponding to the percept of *say*. This situation is maintained for increasing values of the parameter corresponding to gap duration (although the basin within which attraction occurs becomes progressively shallower and narrower), until a critical value is reached, at which point an additional attractor corresponding to the percept of *stay* is created. Both attractors coexist until a second critical value is reached at which the attractor corresponding to *say* disappears. Then, once again, only a single attractor is present, with the basin of attraction becoming deeper and wider with increasing gap duration. When both attractors coexist, random disturbances (corresponding to fatigue, attention, boredom, and so on) can produce spontaneous switches between the two percepts.

The effects of hysteresis and enhanced contrast are accounted for by including in the model a term corresponding to the number of perceived repetitions of a stimulus, so that when the listener hears many repetitions of a stimulus (or of different stimuli that are interpreted as belonging to the same category) the location of the boundary for that category shifts. The effects of this term are, in turn, sensitive to cognitive factors (represented in the model as the combined effects of learning, attention, and experience). Among the resulting predictions are that factors such as learning and experience with the stimuli will make listeners less likely to cling to the initial perceptual state (hysteresis), and more likely to undergo an early switch in percept (enhanced contrast).

The appeal of this model is that it accounts for the emergence and plasticity of linguistic categories by making their inherent relational nature fundamental to perception, and allowing for multiple influences on the final, context-sensitive percept. This contrasts with the more traditional approach of listing the attributes of a particular category. We admire the focus on the temporal evolution of the percept, and on the influence of recent sensory experience on perceptual decisions (cf. also Ding *et al.* 1995). The model focuses mainly on properties of a restricted stimulus set—the way the stimuli are presented—since this offers an interesting window on the stability of percepts. Although we are more interested in the effects of a more varied phonetic context, as found in normal listening situations, we expect that modelling these effects must similarly take into account rich perceptual dynamics, as some of Tuller's later work suggests is the case (e.g. Rączaszek *et al.* 1999).

There are parallels in Tuller's model with, on the one hand, Goldinger's investigations of the effects of repeated exposure to identical or very similar tokens, and on the other hand Grossberg's modelling of the dynamic emergence of categorisation judgements. Where Tuller *et al.*'s model differs from Grossberg's is in its explicit focus on the plasticity of categories defined in terms of immediate relations in the sensory signal (as well as in terms of other factors such as fluctuations in attention). In contrast, the focus of an ART model such as PHONET is on achieving an invariant percept for any given ratio of activity between the model's two processing streams, which implies that the model consistently produces the same response given successive identical inputs. As we have made clear, ART systems do have complex learning dynamics. For example, each time a resonant state occurs in the system, learning takes place, which presumably means that multiple repetition of a stimulus modifies the structure of the category onto which that stimulus is mapped. It is certainly possible, then, that the important aspect of plasticity captured by Tuller *et al.*'s model could be simulated in the ART framework, but this remains to be explored.

5.5. Modelling acquisition

Language acquisition plays an important role in the kind of model of speech understanding represented by Polysp. In part this is because listeners' capacity to learn about speech does not stop once a language has been acquired, so that, for us, acquisition and adult behaviour need not be radically discontinuous. Polysp assumes that the basic processes are similar, although the adult's greater experience means that actual patterns of neural activation will often differ, so that the details of how a particular piece of knowledge is learned and structured may also differ between adults and children. Understanding of how babies and young children form speech sound categories should thus provide insights into the nature of their plasticity, and may provide insights into how adults use them.

Another reason why acquisition must feature in a phonetically-sensitive polysystemic model of speech understanding is that the processes presumed to underlie the development of speech sound categories and phonology are the same as those underlying the developing organisation of grammar and meaning. This view is supported by increasing evidence that sensitivity to the distribution of sound patterns in the speech signal provides a basis from which infants learn about properties of language, such as the nature of different word classes (e.g. Smith 1999). Thus the infant and young child constructs all his or her knowledge of the language in parallel from the primary speech input, which is usually connected speech. Just as speech perception or understanding does not involve an orderly sequence in which the signal is mapped onto phonological units which are then mapped onto grammatical units and then onto meaning, so there is no learning sequence of the sort in which phonology provides the input to grammar which in turn provides the input to meaning. Indeed, logically, it seems more likely that meaning and the phonetic signal are closely related, and that grammatical and phonological systems arise later, as frequent patterns of activation become associated into common structures, or functional groupings. Looked at this way, phonology and grammar are, as it were, by-products of the association of meaning with particular sound patterns.

One recent development which reflects well some of our views on acquisition is Plaut & Kello's (1999) distributed connectionist model of the emergence of phonology in infants from the interplay of speech production and comprehension. In the model acoustic input is mapped directly onto a semantic level, via a level of phonological representations which are not predefined, but are learned by the system under the pressure of understanding and producing speech. Accessing meaning is of central importance, therefore, and phonological representations emerge to support this goal by recoding time-varying acoustic input into a more stable format. The idea that phonological representations are emergent tallies well with the neuropsychological evidence discussed by Coleman (1998, ms) and Pulvermüller (1999) in Section 4.4.1 above. Unfortunately, in order to make the model computationally tractable, Plaut & Kello were forced to simplify the phonetic parameters so that much of the important detail is neglected.

Jusczyk's WRAPSA model (Word Recognition and Phonetic Structure Acquisition, Jusczyk 1993, 1997) includes much more fine phonetic detail than Plaut & Kello's model can, because it has not been computationally implemented. WRAPSA is potentially richly polysystemic, synthesizing work from a number of areas, including phonetic temporal structure, linguistic categories, attention and memory, and assuming that input patterns derived from weighted auditory properties are matched against stored traces in the manner proposed by Hintzman (1986, 1988). In these respects it is immensely appealing, although its focus on word identification as a final goal seems a limitation compared with Plaut & Kello's attempt to map the input directly onto meaning.

WRAPSA's treatment of phonetic temporal structure is particularly promising in that it combines attention to auditory fine detail with integration of information over a window larger than the segment. The auditory analyzers are assumed to carry out some temporal tagging of features that co-occur within the same syllable-sized unit. (A syllable-sized unit has a local maximum between two local minima in the amplitude contour.) We have emphasised, on the other hand, that when adults understand speech, they must integrate information over multiple time domains, some of which are longer than the syllable. It seems probable that the domain or range of domains used by listeners should change during maturation (cf. Drake *et al.* 2000). It is interesting in this regard that infants may use a smaller window for processing information than adults, due to limited attentional and memory capacities, and there is evidence that 'starting small' could help infants identify perceptually relevant units in the input (cf. Elman 1999).

WRAPSA also shares with Polysp the properties that potential lexical units are not initially represented in fully phonetically specified form, and are not decomposed into phoneme-like segments, although temporal relationships between the acoustic properties within each unit are represented. For instance, infants' early perceptual representations might encode enough information for some of the relational contrasts of the prosodic hierarchy to be abstracted, although the entire set of contrasts would not yet be in place. Jusczyk proposes that early representations are syllabic, that is, focused around the middle of a fully-specified prosodic hierarchy of an utterance. We would add to this proposal the speculation that highly reliable acoustic information is more likely to be represented than less clear information, and that this distinction can cut across that of syllabic identity. For example, a strong syllable can be identified with high certainty even when its component phones are only partially identified; on the other hand, it may not always be clear, at least to novice listeners, how many unstressed syllables are present, especially in a sequence of unaccented syllables. For example, the phrase *(of) course it is* often has no voiced /ɪ/ for *it,* and instead that syllable is marked by durational changes and possibly some palatalisation of the other sounds in that vicinity e.g. [k^{h}ɔs:tɨz̬̥]; these preserve some of the rhythmic and critical segmental features of the carefully-spoken phrase. Likewise, a strident fricative can be identified with high certainty when there is only low certainty about which syllable it is a member of, or, as in this example, about how many lexical syllables it represents. Here, the integration between rhythmic and segmental structures will help make the phrase readily understandable.

In the light of the evidence (DeCasper & Fifer 1980; Jusczyk *et al.* 1992; Jusczyk *et al.* 1993) that infants in the first year of life retain detailed memory for the speaker's voice, which is normally classed as nonlinguistic information, Jusczyk (1997:227) suggests that it is paradoxical that phonetic representations might be less than fully detailed. He proposes instead that infants' capacities for representing the internal structure of words and syllables may not yet be fully developed; this might be the case if full phonetic representation required links between details for production as well as perception. As should be clear by now, our explanation is rather different. Nonlinguistic and linguistic representations of utterances are not fed by separate strands of information; instead, the same phonetic information is implicated in both. On this view, even incompletely specified phonetic representations are sure to contain information about the speaker, because voice information is bound up as part of the signal and is reflected in many, if not all, phonetic parameters.

In summary, we suggest that the processes represented by Polysp for an adult understanding speech are not fundamentally different from those of an infant who only partially understands the same speech, or even fails to understand it. The main difference between adults and babies in this respect is that the adult has a lot more experience—and hence available structures—within which to place the newly heard utterance. Polysp is compatible with aspects of Plaut & Kello's phonetic-semantic mapping in a self-organising connectionist framework and with Jusczyk's more clearly polysystemic, phonetically-detailed approach that builds on episodic memories. Polysp adds to these models a rather strong stand on the status of linguistic and non-linguistic sensory information: that it is only classifiable as linguistic or non-linguistic by virtue of its current functional role, which is determined not so much by the quality of the sensory information as by the listener's attention. This is one aspect of the plasticity fundamental to Polysp. One implication of this approach is that individuals are assumed to differ in how they structure the same information. In consequence, an individual's linguistic system need be neither fully specified nor correct in the linguist's sense: as long as it allows the individual to understand meaning correctly, the details of how that meaning is arrived at are immaterial, from the point of view of the success of the communication. That is why it is perfectly possible to believe throughout adulthood that the last syllables of *could've* and *would've* contain *of* rather than *have*. They sound as if they end in *of*, and there is no reason why the underlying message should be misunderstood because the construction is thought to contain *of*. It is also why most educated speakers of English do believe that these phrases end in *have*, because what they are taught to write in school (*have*) changes the way they structure this particular spoken unit.

5.6. Summary: attributes of Polysp in existing models of perception

Each of the models we have discussed addresses a different aspect of the ideas fundamental to Polysp: the complex temporal distribution of acoustic information relevant to linguistic distinctions, the role of rhythmically-governed expectations and knowledge in focusing listeners' attention on particular points in the sensory signal, the inherently dynamic nature of speech understanding, including the plasticity of linguistic-phonetic categories and their sensitivity to context, experience, and linguistic development. Most of them go much further than Polysp in that they are computationally implemented, and in general, we have commented more on their principles than their details. Importantly, though, our focus on phonetic detail does lead us to rule out certain approaches and favour others, especially those that have a realistic way of processing temporally distributed information.

Not only do these models offer a rather coherent set of principles compatible with Polysp's basic principles, but the principles themselves are usually thought to have some reality in how the brain works. For example, a recent handbook (Arbib 1995) includes articles that between them propose neural mechanisms for the general and detailed principles we have discussed, such as modified Hebbian cell assemblies, phase locking between external and internal rhythms, different responses to faster and slower events, coupled oscillators, attractors, feedback, and so on. Equally, it is clear from this same book that we cannot be certain that any of the proposed mechanisms functions as described. That we are beginning to have access to them is encouraging, however.

Of particular interest to us is the fact that the models we have discussed in this section rely less than many traditional models on using linguistic units such as phonemes to define their processing stages, and make more use of constructs with independent motivation from other disciplines such as neurobiology, dynamical systems theory or more general aspects of behaviour such as rhythm perception. It is because of their comparative agnosticism about the units of speech perception, combined with their rich processing dynamics, that models like these are relevant to Polysp and *vice versa.*

Why then is it worth developing Polysp? As a linguistic model, Polysp complements dynamic, self-organising and exemplar models such as these because it makes strong claims about the wealth of linguistic information that is systematically available from the sensory signal and the way this linguistic information is organised; it stresses more than most models that linguistic processing must draw upon fine-grained experience and expectations. Yet, on phonetic as well as cognitive grounds, Polysp also proposes that there is no one way to understand a speech signal: polysystemic linguistic structure can be identified by many routes, in many different orders or in parallel.

6. Concluding remarks

Polysp offers a framework that can potentially hold good for all the varieties of speech that speakers standardly produce and understand, and that is able to include all the linguistic and nonlinguistic knowledge that listeners use when they understand what another person is saying. Let us take an example to see how this might work. Most native speakers of English have an impressively wide variety of ways of conveying the meaning of *I do not know.* The most common forms probably range between *I don't know* and *dunno*, both of which can be pronounced in a number of different ways. However, there are many other variants, the most extreme forms of which can only be used in particular circumstances. For example, it is hard to say the fully expanded form *I do not know* without conveying some degree of exasperation. An even more extreme form has pauses between the words (*I….do…not…know*) and (in most cases) is so rude that it can only be used when the listener does not seem willing to accept that the speaker really does not know. At the other extreme, it is possible—again, only in the right circumstances—to convey one's meaning perfectly adequately by means of a rather stylized intonation and rhythm, with very weak segmental articulation, ranging between something like [ã̱ᵊ̃nːəᶷ] and [ə̞̃ə̃ə̞̃] (intonation not marked). This type of utterance could allow successful communication between relaxed family members, for example when A asks B where the newspaper is, and B does not know, but does not feel that she needs to stop reading her book in order to help find it. Notice that the intonation pattern alone is not enough: at least the vowels must be there ([m]s will not do), and the vowels must start more open (and probably more fronted) than they finish, just as in the more clearly-spoken utterance, so that, at least in this situational context, [ə̝̃ə̃ə̝̃] is nonsense whereas [ə̝̃ə̃ə̝̃] is not.

These minimalist utterances are understood with no difficulty in the type of situation we have described, but they are unlikely to be understood outside them. They can be explained within the Polysp framework by postulating a functional neuronal grouping (a cell assembly for instance) linking these speech variants to meaning and to the affective, socio-cognitive and other properties of the experienced situational sensations. That meaning, in turn, is part of another functional cell assembly that includes representations of other ways of expressing the same meaning in other circumstances. Directly, or indirectly via other links, these other assemblies will include representation of the actual lexical items *I do not know* or *I don't know* and their individual and sentence meaning. In essence, these cell assemblies describe relationships between phonetic forms and connotative meanings. One could speculate that denotative meaning might lie at the intersection of all these variants on connotative meaning. Another way to express this is that the phonetic form reflects the perceived pragmatic meaning of the utterance, which is represented at the neuronal level by nuances in the composition or functioning of particular cell assemblies.

In this type of system, no one type of information need be more important than any other type, and the same result can be arrived at from different starting points. When B says [ə̝̃ə̃ə̝̃], A learns a great deal more than simply that B does not know where the newspaper is. Equally, because A sees that B is deeply involved with her book, he will be less likely to interpret her [ə̝̃ə̃ə̝̃] as a sort of dysarthric grunt preparatory to a more helpful utterance. In other words, the meaning can be arrived at by linking the perceived multi-faceted (i.e. detailed) situation with perceived sound in a mutually reinforcing way, no matter which is attended to most at the outset, or which type of neuron fires first. What is crucial is that the experience is coherent, which is to say that its components must bear appropriate relationships to one another.

These arguments seem uncontroversial in that they are basically common sense. We suggest that the same type of argument can—and indeed should—be made for how we understand more standard forms of speech. We have already made the case that components of a speech signal must be coherent, bearing the right relationships with one another. We have shown that the acoustic fine detail of the sound signal is a rich source of information about all aspects of the message, and argued that understanding will be most efficient when these details are attended to rather than when they are abstracted from early in the perceptual process. These views suggest that speech understanding is ultimately rooted in multi-modal episodic memory and subject to continuous adjustment (plasticity) throughout life.

We have also suggested (Section 5.5) that the perceptual system may access meaning from speech by using the most salient sensory information from any combination of levels of formal linguistic analysis, and from this create a representation that may be incomplete in some respects. By implication, (a) although some units might normally be more fundamental than others (stressed syllables are good candidates, for example—cf. Cutler & Norris 1988; Jusczyk 1997; Greenberg 1999) there is no one basic unit of speech analysis; and (b) there must be perceptual processes which are distinguished by the duration of the phonetic information they process: some process short-time events, while others process more long-lasting information (Section 3.6). This requirement in a model of speech understanding brings it closer to models of syntactic processing and contrasts sharply with the many computational models of speech perception which use a standard time-window and rather abstract input, thereby risking distortion of the relative importance of different types of processing. Nearey engagingly endorses this criticism of most models, including his own: "For example, in my models, instead of temporally-evolving acoustic waveforms, I start with neatly packaged, preparsed 'cues' which feed just the right nodes in my models in just the right way. Furthermore, my models are simple static-pattern recognizers where all evidence is presented instantaneously, rather than emerging through time." (2000:343).

Another implication is that there is no obligatory order a listener must follow in order to understand speech, and the process of speech understanding can be circular. The physical signal will usually dominate and drive perceptual decisions, and processing itself is modulated or driven by the temporal nature of the speech signal, including its rhythm; but listeners must look for grammar as well as lexical identity in speech, and (normally) work out grammar and meaning simultaneously from the spoken signal. Phoneme or allophone strings will not allow this unless they include the type of polysystemic context-sensitivity intrinsic to Polysp, rather than just sensitivity to other phonemes in the immediate vicinity. And if the nature of their context-sensitivity is polysystemic as in Polysp, then phonemes become irrelevant to the process of understanding meaning from spoken utterances. Similarly, meaning might sometimes be understood before individual words and their grammatical relations are identified, and certainly before a complete phonological representation is arrived at (cf. Warren 1999; Grossberg 2000a). For this reason, one remembers the meaning (or perceived meaning) but not always the actual sounds or lexical items that were spoken, just as has been demonstrated for grammar (Sachs 1967). Moreover, individual differences stemming from 'incomplete' or idiosyncratic phonological and grammatical structures are likely to be the norm rather than unusual.

In sum, we have tried (a) to show that the search for a better model of speech understanding depends partly on acknowledging that the phonetic signal is a rich source of information about a large number of different types of linguistic and nonlinguistic information, (b) to outline the types of attributes a more comprehensive model should have and (c) to suggest processes or mechanisms by which these attributes can be built into a model. In common with a number of other recent theories, Polysp assumes that, rather than moving as fast as possible towards an abstract representation of speech, listeners retain the richness of acoustic information at least until the meaning has been identified. We are thus suggesting that the speech signal could be seen as an integral aspect of meaning rather than a relatively uninteresting carrier of meaning; and that phonetic categories behave just like other linguistic categories: they are emergent, dynamic, plastic throughout life, and not simply context-sensitive, but can only be discussed in relational terms.

Address of the authors:

Sarah Hawkins: sh110@cam.ac.uk
Rachel Smith: rhs20@cam.ac.uk
Department of Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, United Kingdom.

Acknowledgements

We acknowledge with gratitude the intellectual debt we owe to John Local and Richard Ogden for sharing their knowledge and ideas about phonetics, and for many hours of patient exposition of linguistic-phonetic structure; and to Noël Nguyen for collaborating in experiments and the development of theory about the role of systematic acoustic-phonetic fine detail in spoken word understanding. We also thank Friedemann Pulvermüller for helpful and stimulating discussions. This work was supported in part by EPSRC grants GR/N19595 and GR/L53069 to the first author, by Swiss National Science Foundation grant 83BC-056471 to Noël Nguyen and the first author, and by an AHRB research studentship to the second author.

Notes

[1] Norris *et al.* (2000:318) note that Merge would get the same results if it used features instead of phonemes prelexically. Since they do not say whether they would structure the features in terms of higher- and low-level nodes with different domains of application, one assumes that they have in mind discrete feature bundles that represent essentially the same thing as the phonemes they actually use. We would therefore expect the same results.

[2] Many of the experiments that demonstrate these properties of speech perception were conducted within the context of the early search for auditory neural feature detectors in the brain comparable to those found for vision. That approach was abandoned when it became clear that central, integrative processes seem to be at least as important to speech perception as relatively simple, automatic responses to the presence or absence of auditory properties (see Darwin (1976) and Eimas & Miller (1978) for reviews). At that time, demonstrations of phonetic trading relations were considered to provide one of the most damaging blows to the idea that auditory feature detectors would prove to underlie phonological features, especially if the trading relations were between complicated combinations of acoustic properties that tend to co-occur in natural speech because of the way the vocal tract works, rather than because of any acoustic similarities. However, the basic idea seems broadly compatible with more recent thinking about neuronal behaviour as characterised by complex interactions between different functional groupings of cell units, although the process is undeniably more complicated than had originally been hoped.

Bibliographical References

Abeles, Moshe (1982), *Local Cortical Circuits: an electrophysiological study*, Berlin, Springer-Verlag.

Abeles, Moshe (1991), *Corticonics: neural circuits of the cerebral cortex*, Cambridge, Cambridge University Press.

Abu-Bakar, Mukhlis & Nick Chater (1995), “Time-warping tasks and recurrent neural networks”, in Levy, Bairaktaris, Bullinaria & Cairns (1995:269-288).

Alfonso, Peter J. & Thomas Baer (1982), “Dynamics of vowel articulation”, *Language and Speech* 25:151-173.

Allen, Joseph & Mark S. Seidenberg (1999), “The emergence of grammaticality in connectionist networks”, in MacWhinney (1999:115-151).

Arbib, Michael A. (1995), *The Handbook of Brain Theory and Neural Networks*, Cambridge, MA: MIT Press.

Arbib, Michael A. (2000), "The mirror system, imitation, and the evolution of language", to appear in Nehaniv & Dautenhahn (2000) http://www-hbp.usc.edu/_Documentation/papers/arbib/LanguageEvolution1-00.doc.

Assmann, Peter F., Terrance M. Nearey & John T. Hogan (1982), "Vowel identification: orthographic, perceptual, and acoustic aspects", *Journal of the Acoustical Society of America* 71:975-989.

Barrett, Sarah E. (1997), *Prototypes in Speech Perception*, unpublished Ph.D. dissertation, University of Cambridge.

Barrett Jones, Sarah E. & Sarah Hawkins (in preparation), "The sensitivity of perceptual magnet effects to phonetic context".

Beckman, Mary E. (1986), *Stress and non-stress accent*, Netherlands Phonetic Archives 7, Dordrecht, Foris.

Beddor, Patrice Speeter & Rena Arens Krakow (1999), "Perception of coarticulatory nasalization by speakers of English and Thai: evidence for partial compensation", *Journal of the Acoustical Society of America* 106:2868-2887.

Beddor, Patrice Speeter & Handan Kopkalli Yavuz (1995), "The relation between vowel-to-vowel coarticulation and vowel harmony in Turkish", in Elenius & Branderud (1995,2:44-51).

Beddor, Patrice Speeter, Rena Arens Krakow & Louis M. Goldstein (1986), "Perceptual constraints and phonological change: a study of nasal vowel height", *Phonology Yearbook* 3:197-217.

Bell-Berti, Fredericka (1993), "Understanding velic motor control: studies of segmental context", in Huffman & Krakow (1993:63-85).

Benguerel, André-Pierre & Helen A. Cowan (1974), "Coarticulation of upper lip protrusion in French", *Phonetica* 30:41-55.

Best, Catherine T. (1994), "The emergence of native-language phonological influences in infants: a perceptual assimilation model", in Goodman & Nusbaum (1994:167-224).

Best, Catherine T. (1995), "A direct realist view of cross-language speech perception", in Strange (1995:171-204).

Boardman, Ian, Stephen Grossberg, Christopher Myers & Michael Cohen (1999), "Neural dynamics of perceptual order and context effects for variable-rate speech syllables", *Perception and Psychophysics* 61:1477-1500.

Bregman, Arthur S. (1990), *Auditory Scene Analysis, the Perceptual Organization of Sound,* Cambridge, MA, MIT Press.

Browman, Catherine P. & Louis Goldstein (1989), "Articulatory gestures as phonological units", *Phonology* 6:201-251.

Browman, Catherine P. & Louis Goldstein (1990), "Gestural specification using dynamically-defined articulatory structures", *Journal of Phonetics* 18:299-320.

Browman, Catherine P. & Louis Goldstein (1992), "Articulatory phonology: an overview", *Phonetica* 49:155-180.

Buxton, Hilary (1983), "Temporal predictability in the perception of English speech", in Cutler & Ladd (1983:111-121).

Carterette, Edward C. & Morton P. Friedman, eds. (1976), *Handbook of Perception*, vol. 7, New York, Academic Press.

CLUMECK, Harold (1976), "Patterns of soft palate movements in six languages", *Journal of Phonetics* 4:337-351.

COHEN, Antonie, J.F. SCHOUTEN & Johan T'HART (1962), "Contribution of the time parameter to the perception of speech", in *Proceedings of the IVth International Congress of Phonetic Sciences*, The Hague, Mouton: 555-560.

COLEMAN, John S. (1994), "Polysyllabic words in the YorkTalk synthesis system", in Keating (1994:293-324).

COLEMAN, John S. (1998), "Cognitive reality and the phonological lexicon: a review", *Journal of Neurolinguistics* 11:295-320.

COLEMAN, John S. (ms), "Phonetic representations in the mental lexicon", under review for publication in Durand & Laks (in preparation).

CONNELL, Bruce & Amalia ARVANITI, eds. (1995), *Phonology and Phonetic Evidence: Papers in laboratory phonology IV,* Cambridge, Cambridge University Press.

COUPER-KUHLEN, Elizabeth (1986), *An introduction to English prosody*, London, Edward Arnold.

CUTLER, Anne & D. Robert LADD, eds. (1983), *Prosody: models and measurements,* Berlin / Heidelberg, Springer-Verlag.

CUTLER, Anne, James M. MCQUEEN & Rian ZONDERVAN, eds. (2000), *Proceedings of the workshop on spoken word access processes (SWAP),* Nijmegen, Max-Planck Institute for Psycholinguistics.

CUTLER, Anne & Dennis NORRIS (1988), "The role of strong syllables in segmentation for lexical access", *Journal of Experimental Psychology: Human Perception and Performance* 14:113-121.

DARWIN, Christopher J. (1976), "The perception of speech", in Carterette & Friedman (1976:175-226).

DARWIN, Christopher J. & Roy B. GARDNER (1985), "Which harmonics contribute to the estimation of first formant frequency?", *Speech Communication* 4:231-235.

DECASPER, Anthony J. & William P. FIFER (1980), "Of human bonding: newborns prefer their mothers' voices", *Science* 208:1174-1176.

DING, Mingzhou, Betty TULLER & J. A. Scott KELSO (1995), "Characterizing the dynamics of auditory perception", *Chaos* 5:70-75.

DOCHERTY, Gerard J. & D. Robert LADD, eds. (1992), *Papers in Laboratory Phonology II: gesture, segment, prosody,* Cambridge, Cambridge University Press.

DRAKE, Carolyn, Mari Riess JONES & Clarisse BARUCH (2000), "The development of rhythmic attending in auditory sequences: attunement, referent period, focal attending", *Cognition* 77:251-288.

DUFFY, Susan A. & David B. PISONI (1992), "Comprehension of synthetic speech produced by rule: a review and theoretical interpretation", *Language and Speech* 35:351-389.

DURAND, Jacques & Bernard LAKS (in preparation), *Phonology: from phonetics to cognition*. Oxford: Oxford University Press.

DURLACH, Nathaniel I. & Louis D. BRAIDA (1969), "Intensity perception. I. Preliminary theory of intensity resolution", *Journal of the Acoustical Society of America* 46:372-383.

EIMAS, Peter D. & Joanne L. MILLER (1978), "Effects of selective adaptation on the perception of speech and visual patterns: evidence for feature detectors", in Walk & Pick (1978:307-345).

ELENIUS, Kjell & Peter BRANDERUD, eds. (1995), *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, KTH and Stockholm University.

ELMAN, Jeffrey L. (1990), "Finding structure in time", *Cognitive Science* 14:179-211.

ELMAN, Jeffrey L. (1999), "The emergence of language: a conspiracy theory", in MacWhinney (1999:1-27).

ELMAN, Jeffrey L. & James L. MCCLELLAND (1986), "Exploiting lawful variability in the speech wave", in Perkell & Klatt (1986:360-380).

FAULKNER, Andrew & Stuart ROSEN (1999), "Contributions of temporal encodings of voicing, voicelessness, fundamental frequency, and amplitude variation to audio-visual and auditory speech perception", *Journal of the Acoustical Society of America* 106:2063-2073.

FELLOWES, Jennifer M., Robert E. REMEZ & Philip E. RUBIN (1997), "Perceiving the sex and identity of a talker without natural vocal timbre", *Perception and Psychophysics* 59:839-849.

FIXMER, Eric (2001), *Grouping of Auditory and Visual Information in Speech*, dissertation submitted to fulfil the requirements of the Ph.D. degree, University of Cambridge.

FOUGERON, Cécile & Patricia A. KEATING (1997), "Articulatory strengthening at edges of prosodic domains", *Journal of the Acoustical Society of America* 101:3728-3740.

FOWLER, Carol A. (1986), "An event approach to the study of speech perception from a direct-realist perspective", *Journal of Phonetics* 14:3-28.

FOWLER, Carol A. & Dawn J. DEKLE (1991), "Listening with eye and hand: cross-modal contributions to speech perception", *Journal of Experimental Psychology: Human Perception and Performance* 17:816-828.

FOWLER, Carol A. & Mary R. SMITH (1986), "Speech perception as 'vector analysis': an approach to the problems of invariance and segmentation", in Perkell & Klatt (1986:123-139).

FROMKIN, Victoria A. (1985), *Phonetic Linguistics: essays in honor of Peter Ladefoged*, Orlando, Academic Press.

FUJISAKI, Hiroya & Takako KAWASHIMA (1970), "A model of the mechanisms for speech perception: quantitative analysis of categorical effects in discrimination", University of Tokyo *Electrical Engineering Research Institute, Division of Electrical Engineering, Annual Report* 3:59-68.

GERNSBACHER, Morton A., ed. (1994), *Handbook of Psycholinguistics*, San Diego, Academic Press.

GIBSON, Eleanor J. (1991), *An Odyssey in learning and perception*, Cambridge, MA / London, MIT Press.

GIBSON, James J. (1966), *The senses considered as perceptual systems,* Boston, Houghton Mifflin.

GOBL, Christer (1988), "Voice source dynamics in connected speech", *Speech Transmission Laboratory-Quarterly Progress and Status Report (STL-QPSR)*, Stockholm, KTH, 1:123-159.

GOLDINGER, Stephen D. (1997), "Words and voices: perception and production in an episodic lexicon", in Johnson & Mullennix (1997:33-66).

GOLDINGER, Stephen D. (2000), "The role of perceptual episodes in lexical processing", in Cutler, McQueen & Zondervan (2000:155-158).

GOLDINGER, Stephen D., David B. PISONI & John S. LOGAN (1991), "On the nature of talker variability effects on serial recall of spoken word lists", *Journal of Experimental Psychology: Learning, Memory and Cognition* 17:152-162.

GOODMAN, Judith C. & Howard C. NUSBAUM, eds. (1994), *The Fevelopment of Speech Perception: the Transition from Speech Sounds to Spoken Words*, Cambridge, MA / London, MIT Press.

GOTTFRIED, Terry L. & Winifred STRANGE (1980), "Identification of coarticulated vowels", *Journal of the Acoustical Society of America* 68:1626-1635.

GREEN, Kerry P., Gail R. TOMIAK & Patricia K. KUHL (1997), "The encoding of rate and talker information during phonetic perception", *Perception and Psychophysics* 59:675-692.

GREENBERG, Steven (1996), "Understanding speech understanding: towards a unified theory of speech perception", in *Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception*, 1-8.

GREENBERG, Steven (1999), "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation", *Speech Communication* 29:159-176.

GRIESER, DiAnne & Patricia K. KUHL (1989), "Categorization of speech by infants: support for speech sound prototypes", *Developmental Psychology* 25:577-588.

GROSSBERG, Stephen (1986), "The adaptive self-organization of serial order in behavior: speech, language, and motor control", in Schwab & Nusbaum (1986:187-293).

GROSSBERG, Stephen (2000a), "Brain feedback and adaptive resonance in speech perception", *Behavioral and Brain Sciences* 23:332-333.

GROSSBERG, Stephen (2000b), "How hallucinations may arise from brain mechanisms of learning, attention, and volition", *Journal of the International Neuropsychological Society* 6:583-592.

GROSSBERG, Stephen & Christopher W. MYERS (2000), "The resonant dynamics of speech perception: interword integration and duration-dependent backward effects", *Psychological Review* 107:735-767.

GROSSBERG, Stephen, Ian BOARDMAN & Michael COHEN (1997), "Neural dynamics of variable-rate speech categorization", *Journal of Experimental Psychology: Human Perception and Performance* 23:481-503.

GUENTHER, Frank H. & Marin N. GJAJA (1996), "The perceptual magnet effect as an emergent property of neural map formation", *Journal of the Acoustical Society of America* 100:1111-1121.

GUENTHER, Frank H., Fatima T. HUSAIN, Michael A. COHEN & Barbara G. SHINN-CUNNINGHAM (1999), "Effects of categorization and discrimination training on auditory perceptual space", *Journal of the Acoustical Society of America* 106:2900-2912.

HALLE, Morris & K.P. MOHANAN (1985), "Segmental phonology of modern English", *Linguistic Inquiry* 16:57-116.

HARLEY, Trevor A. (1995), *The Psychology of Language: from data to theory*, Hove, Psychology Press.

HARNAD, Stevan, ed. (1987), *Categorical Perception: the groundwork of cognition,* Cambridge, Cambridge University Press.

HAWKINS, Sarah (1995), "Arguments for a nonsegmental view of speech perception", in Elenius & Branderud (1995,3:18-25).

HAWKINS, Sarah & Noël NGUYEN (2000), "Predicting syllable-coda voicing from the acoustic properties of syllable onsets", in Cutler, McQueen & Zondervan (2000:167-170).

HAWKINS, Sarah & Noël NGUYEN (2001), "Perception of coda voicing from properties of the onset and nucleus of *led* and *let*", in Paul DALSGAARD, Børge LIDBERG & Henrik BENNER (eds.), *Proceeding of the 7th International Conference on Speech Communication and Technology (Eurospeech 2001 Scandinavia)*, vol. 1:407-410.

HAWKINS, Sarah & Noël NGUYEN (in press), "Effects on word recognition of syllable-onset cues to syllable-coda voicing", in Local, Ogden & Temple (in press).

HAWKINS, Sarah & Andrew SLATER (1994), "Spread of CV and V-to-V coarticulation in British English: implications for the intelligibility of synthetic speech", in *Proceedings of the 1994 International Conference on Spoken Language Processing,* Yokohama, vol. 1:57-60.

HAWKINS, Sarah & Paul WARREN (1994), "Implications for lexical access of phonetic influences on the intelligibility of conversational speech", *Journal of Phonetics* 22:493-511.

HEBB, Donald O. (1949), *The Organization of Behavior: a neurophysiological theory*, New York / London, Wiley.

HEID, Sebastian & Sarah HAWKINS (1999), "Synthesizing systematic variation at boundaries between vowels and obstruents", in Ohala, Hasegawa, Ohala, Granville & Bailey (1999,1:511-514).

HEID, Sebastian & Sarah HAWKINS (2000), "An acoustical study of long-domain /r/ and /l/ coarticulation", in *Proceedings of the 5th seminar on speech production: models and data* (ISCA), Kloster Seeon, Bavaria, Germany: 77-80.

HILLENBRAND, James M., Michael J. CLARK & Robert A. HOUDE (2000), "Some effects of duration on vowel recognition", *Journal of the Acoustical Society of America* 108:3013-3022.

HINTZMAN, Douglas L. (1986), "'Schema abstraction' in a multiple-trace memory model", *Psychological Review* 93:411-428.

HINTZMAN, Douglas L. (1988), "Judgments of frequency and recognition memory in a multiple-trace memory model", *Psychological Review* 95:528-551.

HUFFMAN, Marie K. & Rena Arens KRAKOW (1993), *Nasals, Nasalization, and the Velum*, New York, Academic Press.

HUGGINS, A. William F. (1972a), "Just noticeable differences for segment duration in natural speech", *Journal of the Acoustical Society of America* 51:1270-1278.

HUGGINS, A. William F. (1972b), "On the perception of temporal phenomena in speech", *Journal of the Acoustical Society of America* 51:1279-1290.

HUME, Elizabeth V. & Keith A. JOHNSON, eds. (2001), *The Role of Speech Perception in Phonology*, New York, Academic Press.

IVERSON, Paul & Patricia K. KUHL (1995), "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling", *Journal of the Acoustical Society of America* 97:553-562.

IVERSON, Paul & Patricia K. KUHL (1996), "Influences of phonetic identification and category goodness on American listeners' perception of /r/ and /l/", *Journal of the Acoustical Society of America* 99:1130-1140.

JENKINS, James J., Winifred STRANGE & Thomas R. EDMAN (1983), "Identification of vowels in 'vowelless' syllables", *Perception and Psychophysics* 34:441-450.

JOHNSON, Keith (1990), "Contrast and normalization in vowel perception", *Journal of Phonetics* 18:229-254.

JOHNSON, Keith (1997), "Speech perception without speaker normalization: an exemplar model", in Johnson & Mullennix (1997:145-165).

JOHNSON, Keith & John W. MULLENNIX, eds. (1997), *Talker Variability in Speech Processing*, San Diego / London, Academic Press.

JOHNSON, Keith, Elizabeth A. STRAND & Mariapaola D'IMPERIO (1999), "Auditory-visual integration of talker gender in vowel perception", *Journal of Phonetics* 27:359-384.

JONES, Mari Riess (1976), "Time, our lost dimension: toward a new theory of perception, attention and memory", *Psychological Review* 83:323-355.

JURAFSKY, Daniel (1996), "A probabilistic model of lexical and syntactic access and disambiguation", *Cognitive Science* 20:137-194.

JUSCZYK, Peter W. (1993), "From general to language specific capacities: the WRAPSA model of how speech perception develops", *Journal of Phonetics* 21:3-28.

JUSCZYK, Peter W. (1997), *The Discovery of Spoken Language*, Cambridge, MA, MIT Press.

JUSCZYK, Peter W., David B. PISONI & John MULLENNIX (1992), "Some consequences of stimulus variability on speech processing by 2-month-old infants", *Cognition* 43:253-291.

JUSCZYK, Peter W., Elizabeth A. HOHNE, Ann Marie JUSCZYK & Nancy J. REDANZ (1993), "Do infants remember voices?", *Journal of the Acoustical Society of America* 93:2373.

JUSCZYK, Peter W., Elizabeth A. HOHNE & Angela BAUMAN (1999a), "Infants' sensitivity to allophonic cues for word segmentation", *Perception and Psychophysics* 61:1465-1476.

JUSCZYK, Peter W., Derek M. HOUSTON & Mary NEWSOME (1999b), "The beginnings of word segmentation in English-learning infants", *Cognitive Psychology* 39:159-207.

KEATING, Patricia A. (1985), "Universal phonetics and the organization of grammars", in Fromkin (1985:115-132).

KEATING, Patricia A., ed. (1994), *Phonological Structure and Phonetic form: Papers in laboratory phonology III*, Cambridge, Cambridge University Press.

KEATING, Patricia A., Taehong CHO, Cécile FOUGERON & Chai-Shune HSU (in press), "Domain-specific articulatory strengthening in four languages", in Local, Ogden & Temple (in press).

KELLY, John & John LOCAL (1986), "Long-domain resonance patterns in English", in *International conference on speech input/output; techniques and applications*, London, Institution of Electrical Engineers, Conference publication no. 258:304-309.

KELLY, John & John K. LOCAL (1989), *Doing phonology,* Manchester, Manchester University Press.

KELSO, J.A. Scott, Dan L. SOUTHARD & David GOODMAN (1979), "On the nature of human interlimb coordination", *Science* 203:1029-1031.

KEWLEY-PORT, Diane & Yijian ZHENG (1999), "Vowel formant discrimination: towards more ordinary listening conditions", *Journal of the Acoustical Society of America* 106:2945-2958.

KINGSTON, John & Randy L. DIEHL (1994), "Phonetic knowledge", *Language* 70:419-454.

KLATT, Dennis H. (1979), "Speech perception: A model of acoustic-phonetic analysis and lexical access", *Journal of Phonetics* 7:279-312.

KOZHEVNIKOV, Valerij A. & Ludmilla A. CHISTOVICH (1965), *Speech: articulation and perception,* Moscow-Leningrad, English translation: J.P.R.S., Washington, D.C., No. JPRS 30543.

KRAKOW, Rena Arens (1993), "Nonsegmental influences on velum movement patterns: syllables, sentences, stress, and speaking rate", in Huffman & Krakow (1993:87-113).

KRAKOW, Rena Arens (1999), "Articulatory organization of syllables: a review", *Journal of Phonetics* 27:23-54.

KRAKOW, Rena Arens, Patrice S. BEDDOR, Louis M. GOLDSTEIN & Carol A. FOWLER (1988), "Coarticulatory influences on the perceived height of nasal vowels", *Journal of the Acoustical Society of America* 83:1146-1158.

KUHL, Patricia K. (1992), "Infants' perception and representation of speech: development of a new theory", in Ohala, Nearey, Derwing, Hodge & Wiebe (1992,1:449-456).

KUHL, Patricia K. & Paul IVERSON (1995), "Linguistic experience and the 'perceptual magnet effect'", in Strange (1995:121-154).

KUHL, Patricia K. & Andrew N. MELTZOFF (1982), "The bimodal perception of speech in infancy", *Science* 218:1138-1141.

KWONG, Katherine & Kenneth N. STEVENS (1999), "On the voiced-voiceless distinction for writer/rider", in *Speech Communication Group Working Papers–Research Laboratory of Electronics,* MIT, XI:1-20.

KYRIACOU, Charalambos P. (in press), "The genetics of time", in *Time,* The Darwin College Lectures 2000, Cambridge, Cambridge University Press.

LACERDA, Francisco (1995), "The perceptual-magnet effect: an emergent consequence of exemplar-based phonetic memory", in Elenius & Branderud (1995,2:140-147).

LADD, D. Robert (1996), *Intonational Phonology,* Cambridge, Cambridge University Press.

LADEFOGED, Peter & Donald E. BROADBENT (1957), "Information conveyed by vowels", *Journal of the Acoustical Society of America* 29:98-104.

LARGE, Edward W. & Mari Riess JONES (1999), "The dynamics of attending:

how people track time-varying events", *Psychological Review* 106:119-159.

LASS, Norman J., ed. (1984), *Speech and Language: advances in basic research and practice,* vol.10, San Diego, Academic Press.

LEHISTE, Ilse (1972), "Manner of articulation, parallel processing, and the perception of duration", *Computer and Information Science Research Center Working Papers in Linguistics,* Ohio State University, 12:33-52.

LEVY, Joseph P., Dimitris BAIRAKTARIS, John A. BULLINARIA & Paul CAIRNS, eds. (1995), *Connectionist Models of Memory and Language*, London, UCL Press.

LIBERMAN, Alvin M. & Ignatius G. MATTINGLY (1985), "The motor theory of speech perception revised", *Cognition* 21:1-36.

LINDBLOM, Björn & Karen RAPP (1973), "Some temporal regularities of spoken Swedish", *PILUS (Papers from the Institute of Linguistics)*, Stockholm, University of Stockholm, 21:1-58.

LOCAL, John K. (1992), "Modelling assimilation in a non-segmental rule-free phonology", in Docherty & Ladd (1992:190-223).

LOCAL, John K. & Richard OGDEN (1997), "A model of timing for non-segmental phonological structure", in van Santen, Sproat, Olive & Hirschberg (1997:109-122).

LOCAL, John K., Richard A. OGDEN & Rosalind A.M. TEMPLE, eds. (in press), *Papers in Laboratory Phonology VI*, Cambridge, Cambridge University Press.

LOTTO, Andrew J., Keith R. KLUENDER & Lori L. HOLT (1998), "Depolarizing the perceptual magnet effect", *Journal of the Acoustical Society of America* 103:3648-3655.

LUCE, Paul A., Stephen D. GOLDINGER & Michael S. VITEVICH (2000), "It's good… but is it ART?", *Behavioral and Brain Sciences* 23:336.

MACMILLAN, Neil A., Rina F. GOLDBERG & Louis D. BRAIDA (1988), "Resolution for speech sounds: basic sensitivity and context memory on vowel and consonant continua", *Journal of the Acoustical Society of America* 84:1262-1280.

MACNEILAGE, Peter F., ed. (1983), *The Production of Speech,* New York, Springer-Verlag.

MACWHINNEY, Brian (1999), *The Emergence of Language,* Mahwah, Lawrence Erlbaum Associates.

MCCLELLAND, James L. & Jeffrey L. ELMAN (1986), "The TRACE model of speech perception", *Cognitive Psychology* 18:1-86.

MCGURK, Harry & John MACDONALD (1976), "Hearing lips and seeing voices", *Nature* 264:746-748.

MACCHI, Marian J. (1980), "Identification of vowels spoken in isolation versus vowels spoken in consonantal context", *Journal of the Acoustical Society of America* 68:1636-1642.

MAGEN, Harriet S. (1997), "The extent of vowel-to-vowel coarticulation in English and Japanese", *Journal of Phonetics* 25:187-205.

MANUEL, Sharon Y. (1990), "The role of contrast in limiting vowel-to-vowel coarticulation in different languages", *Journal of the Acoustical Society of America* 88:1286-1298.

MANUEL, Sharon Y. (1995), "Speakers nasalize /ð/ after /n/, but listeners still hear /ð/", *Journal of Phonetics* 23:453-476.

MANUEL, Sharon Y., Stefanie SHATTUCK-HUFNAGEL, Marie HUFFMAN, Kenneth N. STEVENS, Rolf CARLSON & Sheri HUNNICUTT (1992), "Studies of vowel and consonant reduction", in Ohala, Nearey, Derwing, Hodge & Wiebe (1992,2:943-946).

MARSLEN-WILSON, William & Paul WARREN (1994), "Levels of perceptual representation and process in lexical access: words, phonemes, and features", *Psychological Review* 101:653-675.

MARTIN, Christopher S., John W. MULLENNIX, David B. PISONI & W. Van SUMMERS (1989), "Effects of talker variability on recall of spoken word lists", *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15:676-684.

MASSARO, Dominic W. (1998), *Perceiving Talking Faces: from speech perception to a behavioral principle*, Cambridge, MA / London, MIT Press.

MATTYS, Sven L. (1997), "The use of time during lexical processing and segmentation: a review", *Psychonomic Bulletin and Review* 4:310-329.

MILLER, Joanne L. & Thomas BAER (1983), "Some effects of speaking rate on the production of /b/ and /w/", *Journal of the Acoustical Society of America* 73:1751-1755.

MILLER, Joanne L. & Alvin M. LIBERMAN (1979), "Some effects of later-occurring information on the perception of stop consonant and semivowel", *Perception and Psychophysics* 46:505-512.

MOORE, Brian C.J. (1997), *An Introduction to the Psychology of Hearing*, 4th edition, San Diego / London, Academic Press.

MULLENNIX, John W., David B. PISONI & Christopher S. MARTIN (1989), "Some effects of talker variability on spoken word recognition", *Journal of the Acoustical Society of America* 85:365-378.

NEAREY, Terrance M. (1995), "A double-weak view of trading relations: comments on Kingston and Diehl", in Connell & Arvaniti (1995:28-39).

NEAREY, Terrance M. (1997), "Speech perception as pattern recognition", *Journal of the Acoustical Society of America* 101:3241-3254.

NEAREY, Terrance M. (2000), "Some concerns about the phoneme-like inputs to Merge", *Behavioral and Brain Sciences* 23:342-343.

NEHANIV, Chrystopher & Kerstin DAUTENHAHN, eds. (2000), *Imitation in animals and artifacts*, Cambridge, MA: MIT Press.

NGUYEN, Noël & Sarah HAWKINS (1999), "Implications for word recognition of phonetic dependencies between syllable onsets and codas", in Ohala, Hasegawa, Ohala, Granville & Bailey (1999,1:647-650).

NÍ CHASAIDE, Ailbhe & Christer GOBL (1993), "Contextual variation of the vowel voice source as a function of adjacent consonants", *Language and Speech* 36:303-330.

NORRIS, Dennis G. (1992), "Connectionism: a new breed of bottom-up model?", in Reilly & Sharkey (1992:351-371).

NORRIS, Dennis G. (1994), "Shortlist: a connectionist model of continuous speech recognition", *Cognition* 52:189-234.

NORRIS, Dennis G., James M. MCQUEEN & Anne CUTLER (2000), "Merging information in speech recognition: feedback is never necessary", *Behavioral and Brain Sciences* 23:299-370.

NOSOFSKY, Robert M. (1988), "Exemplar-based accounts of relations between classification, recognition, and typicality", *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14:700-708.

NYGAARD, Lynne C. & David B. PISONI (1998), "Talker-specific learning in speech perception", *Perception and Psychophysics* 60:355-376.

NYGAARD, Lynne C., Mitchell S. SOMMERS & David B. PISONI (1994), "Speech perception as a talker-contingent process", *Psychological Science* 5:42-46.

OGDEN, Richard (1999), "A declarative account of strong and weak auxiliaries in English", *Phonology* 16:55-92.

OGDEN, Richard & John K. LOCAL (1994), "Disentangling autosegments from prosodies: a note on the misrepresentation of a research tradition in phonology", *Journal of Linguistics* 30:477-498.

OGDEN, Richard, Sarah HAWKINS, Jill HOUSE, Mark HUCKVALE, John LOCAL, Paul CARTER, Jana DANKOVIČOVÁ & Sebastian HEID (2000), "ProSynth: an integrated prosodic approach to device-independent, natural-sounding speech synthesis", *Computer Speech and Language*, 14:177-210.

OHALA, John J., Yoko HASEGAWA, Manjari OHALA, Daniel GRANVILLE & Ashlee C. BAILEY, eds. (1999), *Proceedings of the XIVth International Congress of Phonetic Sciences*, University of California, Berkeley.

OHALA, John J., Terrance M. NEAREY, Bruce L. DERWING, Megan M. HODGE & Grace E. WIEBE, eds. (1992), *Proceedings of the 1992 International Conference on Spoken Language Processing*, Banff, University of Alberta.

ÖHMAN, S.E.G. (1966), "Coarticulation in VCV utterances: spectrographic measurements", *Journal of the Acoustical Society of America* 39:151-168.

PARKER, Ellen M. & Randy L. DIEHL (1984), "Identifying vowels in CVC syllables: effects of inserting silence and noise", *Perception and Psychophysics* 36:369-380.

PATTERSON, Roy, http://www.mrc-cbu.cam.ac.uk/personal/roy.patterson/aim/.

PERKELL, Joseph S. (1986), "On sources of invariance and variability in speech production", in Perkell & Klatt (1986:260-263).

PERKELL, Joseph S. & Dennis H. KLATT, eds. (1986), *Invariance and variability in Speech Processes*, Hillsdale, Lawrence Erlbaum Associates.

PICKETT, J.M. (1999), *The Acoustics of Speech Communication: fundamentals, speech perception theory, and technology*, Needham Heights, Allyn and Bacon.

PICKETT, J.M. & Irwin POLLACK (1963), "Intelligibility of excerpts from fluent speech: effects of rate of utterance and duration of excerpt", *Language and Speech* 6:151-164.

PIERREHUMBERT, Janet & David TALKIN (1992), "Lenition of /h/ and glottal stop", in Docherty & Ladd (1992:90-117).

PISONI, David B. (1973), "Auditory and phonetic memory codes in the discrimination of consonants and vowels", *Perception and Psychophysics* 13:253-260.

PISONI, David B. (1997a), "Perception of synthetic speech", in van Santen, Sproat, Olive & Hirschberg (1997:541-560).

PISONI, David B. (1997b), "Some thoughts on 'normalization' in speech perception", in Johnson & Mullennix (1997:9-32).

PISONI, David B., Scott E. LIVELY & John S. LOGAN (1994), "Perceptual learning of nonnative speech contrasts: implications for theories of speech perception", in Goodman & Nusbaum (1994:121-166).

PLAUT, David C. & Christopher T. KELLO (1999), "The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach", in MacWhinney (1999:381-415).

POLS, Louis C.W. (1986), "Variation and interaction in speech", in Perkell & Klatt (1986:140-154).

PROTOPAPAS, Athanasios (1999), "Connectionist modeling of speech perception", *Psychological Bulletin* 125:410-436.

PULVERMÜLLER, Friedemann (1999), "Words in the brain's language", *Behavioral and Brain Sciences* 22:253-336.

PULVERMÜLLER, Friedemann (forthcoming), *Neuroscience of Language: from brain principles to mechanisms organizing words and serial order*, Cambridge, Cambridge University Press.

RĄCZASZEK, Joanna, Betty TULLER, Lewis P. SHAPIRO, Pamela CASE & Scott KELSO (1999), "Categorization of ambiguous sentences as a function of a changing prosodic parameter: a dynamical approach", *Journal of Psycholinguistic Research* 28:367-393.

RECASENS, Daniel (1989), "Long-range coarticulation effects for tongue-dorsum contact in VCVCV sequences", *Speech Communication* 8:293-307.

REILLY, Ronan G. & Noel E. SHARKEY, eds. (1992), *Connectionist Approaches to Natural Language Processing*, Hove, Erlbaum.

REMEZ, Robert E. (2001), "The interplay of phonology and perception considered from the perspective of perceptual organization", in Hume & Johnson (2001:27-52).

REMEZ, Robert E. & Philip E. RUBIN (1992), "Acoustic shards, perceptual glue", *Haskins Laboratories Status Report on Speech Research* SR-111/112:1-10.

REMEZ, Robert E., Philip E. RUBIN, David B. PISONI & Thomas D. CARRELL (1981), "Speech perception without traditional speech cues", *Science* 212:947-950.

REMEZ, Robert E., Philip E. RUBIN, Stefanie M. BERNS, Jennifer S. PARDO & Jessica M. LANE (1994), "On the perceptual organization of speech", *Psychological Review* 101:129-156.

REMEZ, Robert E., Jennifer M. FELLOWES & Philip E. RUBIN (1997), "Talker identification based on phonetic information", *Journal of Experimental Psychology: Human Perception and Performance* 23:651-666.

REMEZ, Robert E., Jennifer M. FELLOWES, David B. PISONI, Winston D. GOH & Philip E. RUBIN (1998), "Multimodal perceptual organization of speech: evidence from tone analogs of spoken utterances", *Speech Communication* 26:65-73.

REPP, Bruno H. (1982), "Phonetic trading relations and context effects: new experimental evidence for a speech mode of perception", *Psychological Bulletin* 92:81-110.

REPP, Bruno H. (1984), "Categorical perception: issues, methods, findings", in Lass (1984:243-335).

REPP, Bruno H. & Alvin M. LIBERMAN (1987), "Phonetic category boundaries are flexible", in Harnad (1987:89-112).

REPP, Bruno H., Alvin M. LIBERMAN, Thomas ECCARDT & David PESETSKY (1978), "Perceptual integration of acoustic cues for stop, fricative, and affricate manner", *Journal of Experimental Psychology: Human Perception and Performance* 4:621-637.

RIZZOLATTI, Giacomo & Michael A. ARBIB (1998), "Language within our grasp", *Trends in Neurosciences* 21:188-194.

ROSEN, Stuart & Peter HOWELL (1987), "Auditory, articulatory, and learning explanations of categorical perception in speech", in Harnad (1987:113-160).

RUMELHART, David E. & James L. MCCLELLAND (1986), "On learning the past tense of English verbs", in Rumelhart, McClelland & the PDP Research Group (1986:216-271).

RUMELHART, David E., James L. MCCLELLAND & the PDP Research Group (1986), *Parallel Distributed Processing: Vol. 2. Psychological and biological models*, Cambridge, MA, MIT Press.

SACHS, Jacqueline S. (1967), "Recognition memory for syntactic and semantic aspects of connected discourse", *Perception and Psychophysics* 2(9):437-442.

SALTZMAN, Elliot & J.A. Scott KELSO (1987), "Skilled actions: a task-dynamic approach", *Psychological Review* 94:84-106.

SALTZMAN, Elliot L. & Kevin G. MUNHALL (1989), "A dynamical approach to gestural patterning in speech production", *Ecological Psychology* 1:333-382.

SAWYER, L., Michael HENNESSEY, Alexandre A. PEIXOTO, E. ROSATO, H. PARKINSON, Rodolfo COSTA & Charalambos P. KYRIACOU (1997), "Natural variation in a Drosophila clock gene and temperature compensation", *Science* 278:2117-2120.

SCHWAB, Eileen C. & Howard C. NUSBAUM, eds. (1986), *Pattern Recognition by Humans and Machines: Volume 1, Speech Perception,* Orlando, Academic Press.

SCHWAB, Eileen C., James R. SAWUSCH & Howard C. NUSBAUM (1981), "The role of second formant transitions in the stop-semivowel distinction", *Perception and Psychophysics* 29:121-128.

SHANNON, Robert V., Fan-Gang ZENG, Vivek KAMATH, John WYGONSKI & Michael EKELID (1995), "Speech recognition with primarily temporal cues", *Science* 270:303-304.

SILIPO, Rosaria, Steven GREENBERG & Takayuki ARAI (1999), "Temporal constraints on speech intelligibility as deduced from exceedingly sparse spectral representations", *Proceedings of Eurospeech 1999:* 2687-2690.

SIMPSON, Greg B. (1994), "Context and the processing of ambiguous words", in Gernsbacher (1994:359-374).

SMITH, Linda B. (1999), "Children's noun learning: how general learning processes make specialized learning mechanisms", in MacWhinney (1999:277-303).

SMITH, Rachel & Sarah HAWKINS (2000), "Allophonic influences on word-spotting experiments", in Cutler, McQueen & Zondervan (2000:139-142).

SOLÉ, Maria-Josep (1995), "Spatio-temporal patterns of velopharyngeal action in phonetic and phonological organization", *Language and Speech* 38:1-23.

STEVENS, Kenneth N. (1983), "Design features of speech sound systems", in MacNeilage (1983:247-261).

STEVENS, Kenneth N. (1989), "On the quantal nature of speech", *Journal of Phonetics* 17:3-45.

STEVENS, Kenneth N. (1995), "Applying phonetic knowledge to lexical access", *4th European Conference on Speech Communication and Technology* 1:3-11.

STEVENS, Kenneth N. (1998), *Acoustic Phonetics*, Cambridge, MA / London, MIT Press.

STEVENS, Kenneth, Sharon Y. MANUEL, Stefanie SHATTUCK-HUFNAGEL & Sharlene LIU (1992), "Implementation of a model for lexical access based on features", in Ohala, Nearey, Derwing, Hodge & Wiebe (1992,1:499-502).

STOBART, Henry & Ian CROSS (2000), "The Andean anacrusis? Rhythmic structure and perception in Easter songs of Northern Potosí, Bolivia", *British Journal of Ethnomusicology* 9:63-94.

STRANGE, Winifred (1989), "Dynamic specification of coarticulated vowels spoken in sentence context", *Journal of the Acoustical Society of America* 85:2135-2153.

STRANGE, Winifred, ed. (1995), *Speech Perception and Linguistic Experience: issues in cross-language research*, Baltimore, York Press.

STRANGE, Winifred, Robert R. VERBRUGGE, Donald P. SHANKWEILER & Thomas R. EDMAN (1976), "Consonant environment specifies vowel identity", *Journal of the Acoustical Society of America* 60:213-224.

STRANGE, Winifred, Thomas R. EDMAN & James J. JENKINS (1979), "Acoustic and phonological factors in vowel identification", *Journal of Experimental Psychology: Human Perception and Performance* 5:643-656.

STRANGE, Winifred, James J. JENKINS & Thomas L. JOHNSON (1983), "Dynamic specification of coarticulated vowels", *Journal of the Acoustical Society of America* 74:695-705.

SUOMI, Kari (1993), "An outline of a developmental model of adult phonological organization and behavior", *Journal of Phonetics* 21:29-60.

SWINNEY, David A. (1979), "Lexical access during sentence comprehension: (re)consideration of context effects", *Journal of Verbal Learning and Verbal Behavior* 18:545-569.

TAJIMA, Keiichi & Robert F. PORT (in press), "Speech rhythm in English and Japanese", in Local, Ogden & Temple (in press).

TULLER, Betty, Pamela CASE, Mingzhou DING & J.A. Scott KELSO (1994), "The nonlinear dynamics of speech categorization", *Journal of Experimental Psychology: Human Perception and Performance* 20:3-16.

TUNLEY, Alison (1999), *Coarticulatory Influences of Liquids on Vowels in English*, unpublished Ph.D. dissertation, University of Cambridge.

VAN SANTEN, Jan P.H., Richard W. SPROAT, Joseph P. OLIVE & Julia HIRSCHBERG, eds. (1997), *Progress in Speech Synthesis,* New York, Springer.

VAN TASELL, Dianne J., Sigfrid D. SOLI, Virginia M. KIRBY & Gregory P. WIDIN (1987), “Speech waveform envelope cues for consonant recognition”, *Journal of the Acoustical Society of America* 82:1152-1161.

WALK, Richard D. & Herbert L. PICK (1978), *Perception and Experience,* New York, Plenum Press.

WARREN, Paul & William MARSLEN-WILSON (1987), “Continuous uptake of acoustic cues in spoken word recognition”, *Perception and Psychophysics* 41:262-275.

WARREN, Richard M. (1970), “Perceptual restoration of missing speech sounds”, *Science* 167:392-393.

WARREN, Richard M. (1984), “Perceptual restoration of obliterated sounds”, *Psychological Bulletin* 96:371-383.

WARREN, Richard M. (1999), *Auditory Perception: a new analysis and synthesis*, Cambridge, Cambridge University Press.

WARRINGTON, Elizabeth K. & Rosaleen A. MCCARTHY (1987), “Categories of knowledge: further fractionations and an attempted integration”, *Brain* 110:1273-1296.

WARRINGTON, Elizabeth K. & Tim SHALLICE (1984), “Category specific semantic impairments”, *Brain* 107:829-854.

WERKER, Janet F. & John S. LOGAN (1985), “Cross-language evidence for three factors in speech perception”, *Perception and Psychophysics* 37:35-44.

WERKER, Janet F. & Richard C. TEES (1984), “Phonemic and phonetic factors in adult cross-language speech perception”, *Journal of the Acoustical Society of America* 75:1866-1878.

WEST, Paula (1999a), “The extent of coarticulation of English liquids: an acoustic and articulatory study”, in Ohala, Hasegawa, Ohala, Granville & Bailey (1999,3:1901-1904).

WEST, Paula (1999b)ɹ “Perception of distributed coarticulatory properties of English /l/ and / /”, *Journal of Phonetics* 27:405-426.

WHALEN, Douglas H. (2000), “Occam’s razor is a double-edged sword: reduced interaction is not necessarily reduced power”, *Behavioral and Brain Sciences* 23:351.

WHALEN, Douglas H. & Sonya SHEFFERT (1997), “Normalization of vowels by breath sounds”, in Johnson & Mullennix (1997:133-144).

ZELLNER KELLER, Brigitte & Eric KELLER (forthcoming), “Representing speech rhythm”, to appear in *Working Papers of COST 258,* University of Lausanne.

ZUE, Victor W. (1985), “The use of speech knowledge in automatic speech recognition”, in *Proceedings of the Institute of Electrical and Electronics Engineers* 73:1602-1615.