Blind sharp-sighted stones
snore inconsistently

On the rapid perceptual processing of speech: From signal information to phonetic and grammatical knowledge[1]

Jan Anward & Björn Lindblom
Department of Linguistics
Stockholm University, Sweden

Abstract

There is one aspect of normal speech comprehension that simultaneous interpreting highlights in particular: *its speed*. It appears fair to say that speech researchers are still far from being able to propose detailed algorithmic accounts of how listeners go about perceptually decoding speech. While several influential theories have been proposed and pursued experimentally, there is still no agreement on the mechanisms which make speech perception so singularly fast and robust. However, in on-going research, certain developments can be discerned that eventually may shed some light on both the mapping problem and its temporal course. Here we pay special attention to a type of mechanism whose prospects of meeting the criterion of processing speed appear particularly favorable: the so-called *exemplar-based* models of speech perception. In such models, phonetic and grammatical entities & rules arise developmentally as *emergent* consequences of the listener's cumulative perceptual experience. Our discussion leads us to conclude that, in comparison with many traditional approaches, exemplar models possess several attractive features. They provide a mechanism for automatically sorting out the acoustic context-dependence of phonetic segments. They permit grammatical analyses to extrapolate to previously unheard, but well-formed utterances. Phonetic and grammatical analyses are seen to occur without 'inference making' or 'hypothesis testing' operating with minimal demands on cognitive operations, i.e., without, strictly speaking, any 'top-down' processing. Last but not least, they have the necessary temporal characteristics. In our opinion, the proposed models appear to offer the kind of perceptual processing that both normal language use and simultaneous interpreting could use.

1. The problem and a phonetic perspective

1.1 *From signal to percept: fast processing of a complex input*

The aim of this paper is to present a brief overview of current work on speech perception and to do so in the context of simultaneous interpreting. That goal provides an opportunity to focus on a rather incompletely understood aspect, namely the *speed of perceptual processing*.

It is fair to say that modern phonetics does not yet offer a comprehensive, unified account of how listeners process speech signals. Several interesting theoretical frameworks have been proposed and examined experimentally (more on that topic anon),

[1] Presented at the *International Symposium on Language Processing and Interpreting*, Stockholm University, 21-22 February, 1997.

but nevertheless much remains to be discovered. We are still far from understanding the mechanisms that make normal speech perception so fast and robust. Rather a high-priority item on the agenda seems to be the formulation of simplified, heuristic models that can guide experimentation and suggest testable, 'in-principle' solutions to the problem of describing both the nature and origin of speech percepts. This state of affairs is linked to one of the classical issues of experimental phonetics: the infinite variety of natural speech patterns.

It is against the background of variability that the speed of processing seems so striking. Reviewing work on the *ear-voice span*, Gerver (1976) mentions a study of simultaneous interpreting and shadowing from prerecorded French into English. Twelve professional conference interpreters participated and performed their task under several different conditions of signal-to-noise ratio. It was found that the ear-voice span increased as a function of decreasing S/N, but remained fairly stable independent of listening conditions. The lag was about 5.7 words for interpreting and about 2 words for shadowing. Other estimates are also mentioned. A lag of at least 2 seconds is reported in one study (Oléron and Nanpon 1964), a figure that tends to increase with increasing difficulty of the input (Treisman 1965).

Simultaneous interpreting involves several stages, e.g., switching attention from input to output, comprehending the source message as well as the planning and execution of the utterance in the target language. Those activities take time. It appears natural to ask, How could there be time also for the process of converting the signal into a speech percept, particularly if the redundancy of the input serves as a trigger for the application of *top-down* knowledge. How could there be time for such top-down processes as 'inference making' and 'hypothesis formation'? Or are those questions asked in the wrong way? What is the nature and origin of top-down and other real-time mechanisms? Towards the end of this chapter we shall come back to those questions.

1.2 The infinite variety of signal patterns

The variability problem becomes evident when we contrast speech as a psychological phenomenon with speech as a physical signal. What language users (and linguists) call the "same" utterance normally appears in the guise of innumerable different physical

shapes: the phrase *'psychologically the same, but physically not the same*' is a nutshell statement of this paradox which is a variant of the classical problem of perceptual constancy in experimental psychology.

We need not illustrate the variability issue in great technical detail. A few simple everyday observations will convince us that it is real. First consider speaker identity. We are good at recognizing people by simply listening to them. We know that different speakers sound different, even though they speak the same dialect and utter the same phonetic segments, the same syllables, words and phrases. The use of the word "same" here poses a challenge for the phonetician, for "identical" speech samples from different speakers are far from identical as physical patterns. For one thing, speakers are built differently. Age and gender are correlated with anatomical and physiological variations, such as differences in the size of the vocal tract and the properties of the vocal folds. Communicating by telephone, we are usually able to tell correctly from the voice alone whether the person is young or old, or is male or female. Evidently we are able to do so from the cues in the acoustic signal. So what are the acoustic events that account for the fact that we hear the "same words" whether they are produced by speaker A or by speaker B? That is in fact a very important question in all kinds of contemporary research on speech.

When we limit our focus to the speech of a single person, we realize that the problem of variability still remains a major one. There is the factor of speaking style and situation. There is an extremely large number of ways in which the syllables and phonemes of the "same" phonetic forms could be spoken. Usually without being fully aware of it, we speak in a manner that depends on who we are talking to. We sound different talking to a baby or a dog, or to someone who is hard of hearing, has only a partial command of the language spoken, or speaks a different dialect. In response to noisy conditions, we spontaneously raise our vocal effort. We articulate more carefully and more slowly in addressing a large audience under formal conditions than when chatting with an old acquaintance. We sometimes tend to mumble and speak more to ourselves than to those present, often with drastic reduction of clarity and intelligibility as a result. The way we sound is affected by how we feel, our voices reflecting the state of our minds and bodies.

Clearly, the speech of a given individual mirrors the intricate interplay of an extremely large number of communicative, social, cognitive and physiological factors.

A third factor becomes evident as we narrow the scope still further. The term 'lab speech' is used with reference to samples of a single person's speech produced in a specific speaking style, at a particular vocal effort and fixed tempo and from a list of test items, not spontaneously generated by the speaker, but chosen by the experimenter. Even such restricted conditions do not prevent the articulatory and acoustic correlates of phonological units from exhibiting extensive variations. Here the variability arises because linguistic entities are not produced clearly delimited one by one in a neat sequence. Rather they form a seamless stream of movements that overlap in time and whose physical shapes depend crucially on how the language in question builds its syllables and how it uses prosodic dimensions such as timing, stress, intonation, and tonal phenomena. This interaction is present irrespective of the language spoken, and whether we observe speech as a motor or acoustic phenomenon. We are referring to the universal aspects of articulatory organization known as coarticulation.

2. Some traditional paradigms in speech perception research

How have major theories of speech perception attempted to deal with the above-mentioned variability problem? How do they address the 'speed of processing' issue?

2.1 "Speech is rather a set of movements made audible than a set of sounds produced by movements" (Stetson 1951).

This quotation applies to so-called gestural accounts: the Motor Theory of Speech Perception (Liberman & Mattingly 1985, 1989), Direct Realism (Fowler 1986, 1991, 1994), Dynamic Specification (Strange 1989a, 1989b), Articulatory Phonology (Browman & Goldstein 1992) and its application to phonological development (Studdert-Kennedy 1987, 1989). Their common denominator is the 'phonetic gesture':

> "… the invariant source of the phonetic percept is somewhere in the processes by which the sounds of speech are *produced*" (Liberman & Mattingly 1985:21; our italics).

> "An event theory of speech production must aim to characterize articulation of phonetic segments as overlapping sets of coordinated gestures, where each set of coordinated gestures conforms to a phonetic segment. By hypothesis, the organization of the vocal tract to produce a phonetic segment is invariant over variation in segmental and suprasegmental contexts." (Fowler 1986:11).

According to the proponents of the Motor Theory,

> ".. there is simply no way to define a phonetic category in purely acoustic terms…" (Liberman & Mattingly 1985:12),

the reason being that speech is a highly *encoded* (strongly coarticulated) phenomenon. Coarticulation is seen as an evolutionary adaptation that occurred to make speech production more rapid. However, this development had complex acoustic consequences and a specialized '*phonetic module*' capable of extracting the phonetic gestures from the speech wave "developed concomitantly" (Liberman & Mattingly 1985:7). Accordingly, the Motor Theory explains the variability of speech as follows: It assumes that there is a straightforward relationship between the linguistic units of an utterance on the one hand, and its speech sounds *qua* phonetic gestures on the other. Furthermore, it is the interaction among the gestures that is responsible for the lack of constancy in their acoustic correlates. Arguing that '*speech is special!'*, Liberman has presented a body of experimental evidence that he takes to indicate that we listen to speech differently from how we hear other sounds.

One example is *duplex perception* (Liberman & Mattingly 1989). This phenomenon refers to the demonstration that a given acoustic attribute (the transition of the third formant in /da/ vs /ga/ syllables) can either be perceived auditorily as a 'chirp', or phonetically as an integral part of the spectral patterns of /da/ or /ga/.

Another finding invoked by Liberman and colleagues is the *confusion of temporal order* reported by Warren & Warren (1970). This effect was observed when listeners, listening to repeated presentations of a high tone, a buzz, a low tone and a hiss, were asked to determine their order either verbally or by ordering four cards. As the stimuli were given shorter and shorter durations, subjects found the task increasingly difficult. For instance, at (roughly syllabic) stimulus durations of 200 ms, they were not able to avoid errors. However, when the arbitrary noises were replaced by spoken numbers also 200 ms long, correct responses were made. For Liberman, the non-speech results demonstrate the slower auditory mode of perception, whereas the responses to the speech stimuli are seen as evidence for the hypothesized faster phonetic module.

According to Direct Realism, speech perception does not require a specialized 'phonetic module'. Inspired by Gibson's ecological psychology (Gibson 1972, 1979)

direct realists assume that the brain has evolved general mechanisms that deal with the environment by extracting perceptual invariants about it. Such mechanisms are assumed to come into play also in speech processing.

> ".. Both the phonetically structured vocal-tract activity and the linguistic information … are directly perceived … by the extraction of invariant information from the acoustic signal … (Fowler 1986:24).

The details of how such mechanisms operate and how they attain their processing speed have not yet been spelled out, but one of the key assumptions of Direct Realism throws some light on this issue:

> "… perception must be *direct* and, in particular, *unmediated by cognitive processes of inference or hypothesis testing*, which introduce the possibility of error.… (Fowler 1986:4; our italics).

2.2 "… we speak to be heard in order to be understood" (Jakobson, Fant & Halle 1951/1963:13).

Although gestural accounts enjoy a great deal popularity today, their status nonetheless remains controversial. Many find the idea of speech being structured by auditory/acoustic goals more compelling.

There is a substantial body of experimental evidence also for this viewpoint. A phenomenon that indicates the primacy of auditory rather than gestural goals is *compensatory articulation*. In one experiment, speakers were asked to produce normal sounding vowels with an atypically large jaw opening maintained by a *'bite-block'* (Lindblom, Lubker & Gay 1979). Although the subjects had no previous experience of the task and were not allowed to practice beforehand, their output nonetheless closely matched normal vowels with respect to both auditory and acoustic properties. Evidently, irrespective of the experimental task, the same acoustic output was maintained. Acoustic constancy is achieved thanks to speech production being under so-called *output*- or *listener-oriented control*.

We can contrast the views of the gesturalists and the auditory school with a few remarks on coarticulation. Consider the pronunciation of the /k/ phoneme in "key" and "coo". The term coarticulation refers to the fact that the /k/ of "key" comes out as fronted in the context of /i/, a front vowel, whereas that of "coo" is more posterior in the environment of the back vowel /u/. The movements associated with the consonant are not completed before the articulatory activities for the vowels are initiated. As a result, there

is overlap among articulatory gestures. The movements of the consonant are *co-articulated* with those for the vowel.

To the gesturalists, such observations suggest that there is a deeper production level at which linguistic units have not yet begun to interact and to merge into complex articulatory patterns. To them this is the level at which the correspondence between linguistic units and phonetic correlates should be sought.

> "… the gestures do have characteristic invariant properties, …, though these must be seen, not as peripheral movements, but as the more remote structures that control the movements. *These structures correspond to the speaker's intentions.*" (Liberman & Mattingly 1985:23; our italics).

Öhman's numerical articulatory model (1967) can be said to give a formalized description of the gesturalist view of coarticulation.

However, several investigators have resisted the conclusion that "there is simply no way to define a phonetic category in purely acoustic terms…" They have continued their search for phonetic invariance at the acoustic level. One approach that has had some success in this respect is the so-called *locus equation* paradigm (Sussman, McCaffrey & Matthews 1991).

The term '*locus'* was introduced to describe place of articulation in consonants (Liberman, Delattre, Cooper and Gerstman 1954). According to the original usage it refers to an ideal onset or offset frequency of a formant transition. The initial expectation was that each place would exhibit its own characteristic locus pattern with fixed frequencies for all formants. However, it soon became clear that the formant patterns of different places of articulation were not constant but tended to show drastic variations depending on context. It can be shown that, for a given place, onset and offset formant transitions are strongly dependent both on the preceding and the following vowel. The ranges for F2 and F3 (the most significant formants for place) show a great deal of overlap for /b/, /d/ and /g/ in different vowel contexts. This extensive vowel-dependence, or coarticulation, makes it impossible to describe each place, as suggested by the original locus theory, in terms of a single formant pattern. However, the context-dependent variation is highly systematic. When, for a given place and set of vowels, the transition onset of F2 is plotted against the value of F2 in the adjacent vowel, data points closely approximate a pattern which is accurately described by a straight line, the 'locus equation'. With this type of representation, each place is distinctively specified by the

slope and intercept of its locus equation. This result represents a form of 'relational' invariance that might, as some investigators (Sussman, Fruchter, Hilbert & Sirosh 1998) have suggested, be exploited by the listener during perceptual processing.

3. A new approach to an old problem.

How do listeners cope with a variable and underspecified speech signal? How do we picture the step from a physical signal that is variable and context-dependent to linguistic units that are the opposite, namely invariant and context-free?

The traditional view (Table 1) has been to regard that step - the mapping from signal to percept - as complex. For instance, in letting a 'phonetic module' do the job, motor theorists tacitly subscribe to a division of labor implying *'complex-processing/simple units'*.

Table 1

	TRADITIONAL VIEW	**RECENT TREND**
MAPPING ***(processing)***	**Complex**	**Simple**
REPRESENTATION ***(units)***	**Simple**	**Complex**

An alternative philosophy **(Johnson&Mullenix in press)** is gradually emerging. It turns the answer around by advocating *'simple-processing/complex units*' (Table 1). In traditional accounts, a canonical linguistic representation is derived from the signal and irrelevant, 'non-linguistic' variations are thrown away. As we saw in reviewing the invariance problem, describing how this mapping occurs is an extremely difficult task for both the gestural and the auditory viewpoints. What proponents of 'simple-mapping/complex units' propose is to treat variability not as an unwanted interference, but as information which is systematic and which should therefore be retained and used. For an early example of such an approach to speech perception see Klatt (1979, 1989).

Let us provide some instructive experimental data.

3.1 An experiment on speech perception by animals

Kluender, Diehl and Killeen (1987) trained Japanese quail to peck in response to /dis/, /dus/, /dæs/ and /das/, but to avoid pecking when hearing the corresponding /bis/, /bus/, /bæs/, /bas/, or /gis/, /gus/, /gæs/, /gas/ sequences. After training, the birds successfully mastered this task. The quail were then presented with new stimuli containing the same consonants but different vowels: /ɪ ɛ ʊ ʌ eɪ oʊ ɔɪ ɚ/. The results clearly showed that the birds were able to generalize their responses to the new stimuli. They made significantly more pecks per minute to the syllables containing /d/ than to the other stimuli.

From these results we gather that the stimulus syllables were sufficiently rich acoustically to support correct categorization of place of articulation in stop consonants. This finding is clearly at variance with the conclusion of motor theorists (cf quotation above from Liberman & Mattingly 1985:12). Phonetic categories do indeed appear to be definable acoustically, at least for the isolated, clearly spoken syllables of this experiment.

3.2 Exemplar models

The next question becomes, What form does this acoustic place specification take? One recent promising line of research is the application of so-called *'exemplar models'* (Estes 1993, Hintzman 1986) to speech perception. We cannot do full justice to this paradigm. Suffice it to say that it has attracted the attention of many speech researchers and seems to be rapidly gaining ground.

Exemplar models assume that memories, including phonetic ones, are built by experience. They hypothesize that it is the set of all perceived instances of the category defines that category. Phonetic categories are thus formed by storing the auditory patterns set up by individual utterances. Every time a certain word is spoken, aspects of its acoustic, or more precisely, (a storage-efficient version of) its auditory attributes are read into phonetic memory. These auditory patterns get labeled as the meaning of the word becomes known (also as a result of experience). The end product is a network of sound-based memory structures each one linking sound (a set of auditory patterns) to meaning (a semantically and grammatically defined category).

3.3 Vowel recognition and speaker identification: an exemplar-based approach

How do exemplar models handle variability and contextual variation? By way of introduction, let us first turn to a study by **Johnson (in press)** who built an exemplar model for speech perception and tested its performance in a vowel identification and speaker recognition experiment. He asked 14 men and 25 women, all native speakers of American English, to read the words *aid, awed, had, head, heed, hid, hood, hud, odd, owed, and who'd*. Five measurements were made at the midpoint of the vowels including and the lowest three formant frequencies, fundamental frequency and vowel duration. An exemplar was defined by (i) these five acoustic properties plus (ii) information on the identity of the word, (iii) the sex of the speaker and (iv) the identity of the speaker. To simulate recognition, each token was treated as a new stimulus and was compared with each of the rest of the exemplars. For every exemplar, its similarity (Nosofsky 1988) to the token to be identified was calculated using the five-parameter descriptions. A high or low score was an indication that the token had produced a correspondingly high or low 'activation level' of the exemplar. Category by category, the activations were summed over all the exemplars. The token's identity was determined by assigning it to the category that showed the largest activation sum. Johnson found that this procedure resulted in 80% correct vowel recognition which is comparable to that of human listeners' in identifying steady-state synthetic vowels (Ryalls & Lieberman 1982). The sex of the speaker was correctly identfied for 96% of the stimuli.

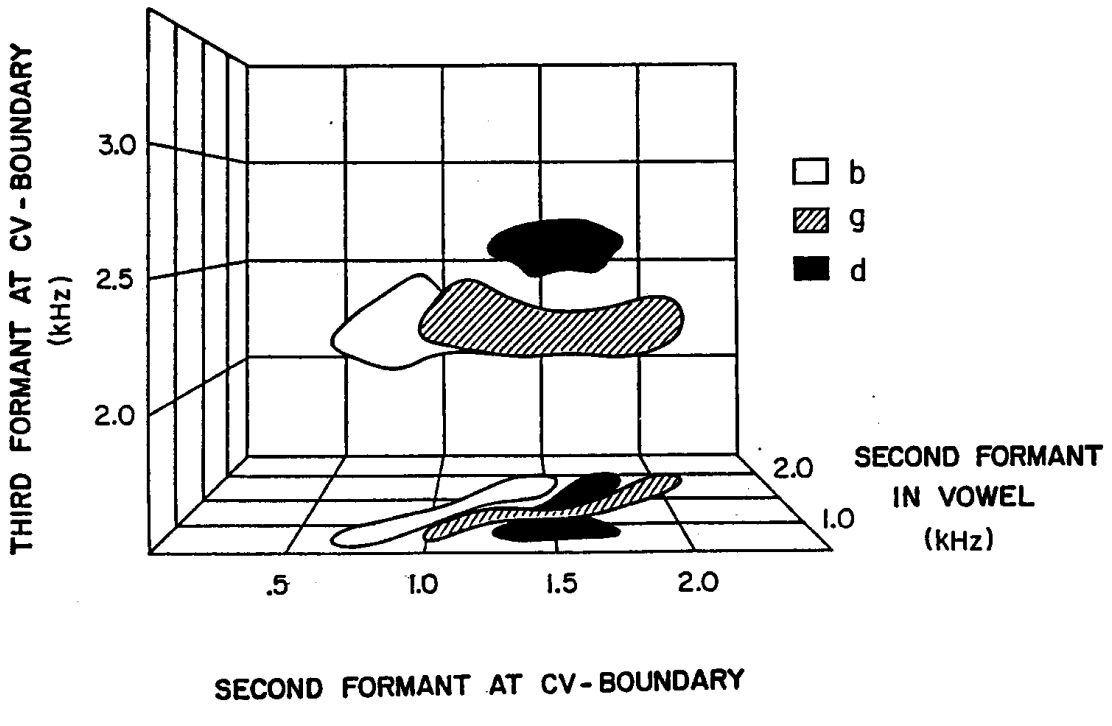
3.4 Representing place correlates of stop consonants as multi-dimensional 'clouds'.

We can now apply the philosophy of the 'exemplar model' to the perceptual performance of the Japanese quail. We shall use some speech samples described in an investigation by Öhman (1966). They were not unlike the stimuli presented to the quail. The test words were symmetrical and asymmetrical V_1CV_2 sequences. They contained all possible combinations of /b d g/ and /y ɸ ɑ o u/ spoken by a Swedish speaker. A total of 75 test items, each spoken at least five times, was analyzed.

Table 2

V_1	C	V_2
y		Y
ø	b	ø
ɑ	d	ɑ
o	g	o
u		u

As expected from what is known about co-articulation, the observed formant transitions to and from the consonants were highly vowel-dependent. For each place, the formant patterns at the boundaries of the consonant were strongly dependent on the identities of the preceding and following vowels. At the release of the stop, the formant pattern depended not only on the identity of the following, but also on the preceding vowel. Conversely, at the beginning of the closure, there was influence from both vowels. In Fig 5 of Öhman (1966), the ranges of the formant transition onsets and offsets for /b/, /d/ and /g/ show extensive overlap. Accordingly, if single F2 and F3 'locus' values were used to describe place of articulation, no unambiguous separation of the three categories would be obtained. In this sense, Liberman& Mattingly are certainly correct in stating that there is no way of defining a phonetic category in purely acoustic terms. However, as shown by the 'exemplar model' of Fig 1, that conclusion turns out not to be valid.



The diagram summarizes some of Öhman's findings. At the top, a stylized spectrogram indicating the selected measurements. They include (i) the onset of the F2 transition (plotted along the x-axis in the three-dimensional diagram), (ii) the onset of the F3 transition (y-axis) and (iii) the F2 value at the V_2-steady-state (z-axis).

The three "clouds" of Fig 1 enclose the measurements from all the test words. (For each cloud there were 125 data points (25 contexts times 5 repetitions), omitted after being enclosed by smooth contours). The drawn-out shapes illustrate the effects of coarticulation. We note that, if we look at the clouds in two dimensions only, for instance at the configurations projected on the ‘ground floor’, there is definitely overlap, but as we use all three dimensions we see that the three clouds do not use any common ‘air space’. They are separated from each other. We conclude that, while it is true that coarticulation eliminates absolute acoustic invariants, it does not jeopardize the separability of the categories, for, in three-dimensional space, the phonetic correlates of the three stops are nevertheless sufficiently distinct.

3.5 An exemplar model of the perception of stop consonants

Provided that in perceiving the stimuli the quail had access to at least the three parameters of Figure 1, the information available in the acoustic signal ought to have been sufficient for the birds to disambiguate the place of the consonants. Admittedly, the three dimensions selected do not in any way make a complete list of the signal attributes that are known to convey place information, one obvious omission being the spectral dynamics of the stop releases. Adding such attributes to the consonant space would be an effective means of further increasing the separation of the three "clouds" and thus enhancing their distinctiveness.

With this analysis of the stop consonant data, we can now return to the perceptual responses of the quail. In the first experiment, the birds learned to distinguish between syllables with or without /d/. This process is in many ways similar to preparing Figure 1. In other words, the learning occurs because the auditory parameters of the individual stimulus exemplars get stored in memory in analog form, and because something equivalent to multi-dimensional "clouds" is eventually established in the neural networks of the quail brain. The following assumption does not seem unjustified: If the /b/-, /d/-

and /g/-tokens are separable acoustically, they should remain so at the neural level. If this assumption is correct, we begin to see how the quail were able to associate the appropriate pecking and non-pecking responses with each discrete "cloud". When the new stimuli of the second experiment were presented, the birds could relate the auditory parameters of that set to the information stored on the exemplars of the first experiment. For instance, a new stimulus falling inside the "/d/-cloud" leads to a pecking response, one that ends up on the outside inhibits pecking.

Our account of the quail findings assumes that the first experiment was like "training a neural net". It set up distinct representations of the three stimulus categories in a multi-dimensional memory space not unlike that of the three "clouds" of Figure 1. This explanation suggests that the reason why the quail were able to generalize their behavior in the second experiment was that the "clouds" of the new stimuli were more or less identical with the old ones. For example, the new /bVs/-tokens remained more similar to the old /bVs/-set than to either of the other two previously established "clouds", and analogously for the new /dVs/-tokens and /gVs/-tokens.

4. Preliminary conclusions

We began by asking how normal speech perception can be so fast and reliable despite the complex patterns that natural speech signals present to the listener.

In a sense, traditional approaches such as the Motor Theory and Direct Realism do address the issue of processing speed. They do so by postulating a ‘phonetic module’ and ‘smart mechanisms’ respectively. But, strictly speaking, by that move they do not actually solve the problems of variability and speed. They just postpone the solution, since the ‘phonetic module’ and the ‘smart mechanisms’ are nothing but ‘black boxes’ until their underlying mechanisms have been specified.

The search for phonetic invariance at the acoustic and/or auditory level is in a similar situation. Despite many decades of acoustic phonetic research, there have been very few detailed accounts of how the listener goes about processing the speech signal so as to derive the invariants assumed to be needed for recognition.

4.1 'Solving' the invariance problem

In the exemplar-based approach, 'distinctiveness' replaces 'invariance' as the criterion for recognition. If our account of the performance of the quail is on the right track, the quail did not 'learn to extract invariants' from the signal. They learned to discriminate /d/-syllables from other stimuli. That discrimination was not based on constancies in the speech wave but on judging the similarity of a stimulus in relation to other instances that had been experienced previously. Note that our account is perfectly analogous to the operation of Johnson's exemplar model in that a stimulus syllable's identity is determined by picking the cloud that the stimulus was most compatible with. To restate, the stimulus was identified with the category for which maximal exemplar activation was produced.

How does the exemplar model handle contextual variation? Significantly it does so by doing away with traditional notions such as 'target' or 'prototype' (Grieser & Kuhl 1989, Kuhl 1992). In the case of the quail stimuli we assumed that it was done by storing 'primary' consonant information (F2 and F3 transition onsets) along with 'secondary' contextual information (F2 vowel). Although the primary information is heavily coarticulated and ambiguous, this variability shows a lawful co-variation with the secondary information whose function becomes that of disambiguating and 'making sense of' the primary information. Evidently, the exemplar-based account exploits the same regularities as locus equations, but it gives a different account in two important respects: it is committed to neither linearity nor signal invariance.

4.2 Speed of perceptual processing.

Note that the proposed process is 'direct' in the sense of Direct Realism. There are no mediating cognitive processes of 'inference'. There is no 'hypothesis testing' (cf Fowler 1986:4). As a consequence, speed of processing is determined by the time taken for the signal to interact with the memory contents.

We can picture that comparison as an automatic process resembling the phenomenon of 'resonance' (Licklider 1952, Shepard 1984). Incoming stimuli simultaneously activate the entire contents of the phonetic memory. The activity set up by the input is sent - in parallel (Rumelhart & McClelland 1986) - to all the patterns stored in memory. The response of these patterns (=their resulting activation level) is a function of their

similarity to the stimulus. The same point could be expressed by saying that their response reflects the extent to which they 'resonate' with the input.

Our account of the quail experiment invoked 'phonetic knowledge' in the form of the 'clouds' of Figure 1. It assumes that even the seemingly trivial task of perceiving a few nonsense syllables cannot be done without 'signal-independent' information which is put into play even in speech perception by animals! The quail did not need a 'phonetic module' à la the Motor Theory, nor 'smart mechanisms' à la Direct Realism. They did not need 'signal invariance'. They simply needed cumulative phonetic experience of a signal patterned in systematic and principled, and thus for the listener interpretable, ways.

5. Grammatical structure in an exemplar-based model

We have argued that the rapidity with which speech perception - and speech production - are accomplished, particularly in simultaneous interpretation, mandates a model of linguistic competence which combines complex representation and simple mapping. Incoming utterances are stored in full phonetic detail, matched against patterns based on the full phonetic detail of previously stored utterances, and 'interpreted' through the way they resonate with these patterns.

Doesn't this commit us to a hopeless and obsolete view of grammatical knowledge? Can we seriously maintain that something like rote learning of utterances is a basic property of language acquisition and that syntactic, morphological and lexical knowledge is only indirectly represented as contrasts and similarities between stored exemplar utterances?

For a number of formulaic utterances, such as *Good morning*!, it obviously makes sense to say that they are directly produced and perceived as instances of exemplar utterances, based on previous experiences of similar utterance tokens. But if a poet springs to the stage and starts reciting[2]

> *Blind, sharp-sighted stones*
> *snore inconsistently*

[2] This example is obviously based on Chomsky's *Colorless green ideas sleeo furiously* (Chomsky 1957:15). The reason we provide a new version is that Chomsky's original utterance hardly qualifies as a novel utterance anymore, having been repeated any number of times and even provided with a kind of official interpretation (Jakobson 1959).

this mode of description is not available. How can we deal with the novelty of these lines? How can utterances be constructed from parts of old utterances, and perceived as being so constructed, in an exemplar-based model of speech production and speech perception?

Let us briefly review how grammatical structure is manifested in verbal utterances, using the utterances in (1)[3], two successive utterances from an informal, multi-party conversation in Swedish[4] as examples.

(1) a. jakommerihågbara**ett**fall↓(p)

b. devaenmanme**rev**bensfrakturer↓(p)

Utterances in a language are typically interpreted by speakers of that language as STRINGS of lexical items that 'belong' to the language. For example, the utterances in (1) are effortlessly segmented by Swedish speakers as in (2).

(2) a. ja kommer ihåg bara **ett** fall↓(p)
b. de va en man me **rev**bensfrakturer↓(p)

(a. 'I come in.mind (=remember) just one case'
b. 'it was a man with fractured.ribs')

Not all strings of native lexical items constitute possible utterances in a language, though. If we produce the items in (2a) in alphabetical order, as in (3a), in reverse alphabetical order, as in (3b), or in an order of increasing size, as in (3c), the results are all ungrammatical:

(3) a. bara ett fall ihåg ja kommer
b. kommer ja ihåg fall ett bara
c. ja ett fall ihåg bara kommer

(a. 'just one case in.mind I come'
b. 'come I in.mind case one just
c. 'I one case in.mind just come)

Permissible strings in a language may be reduced, expanded, or permuted (Harris 1957)**.** In (2a), *bara ett fall* (only one case) has been reduced to *det* (it); in (2b), *bara ett*

[3] A syllable in boldface indicates sentence accent, ↓ indicates the of a falling intonation contour, and (p) indicates a pause.
[4] Talsyntax E1: 29, 1974. Thanks to Jan Einarsson for making this material accessible.

fall (only one case) has been expanded to *bara ett sånt fall* (only one such case); and in (2c), *bara ett fall* (only one case) has been fronted.

(2) a. ja kommer ihåg det
('I remember it')

b. ja kommer ihåg bara ett sånt fall
('I remember only one such case')

c. bara ett fall kommer ja ihåg
('Only one case do I remember)

Such operations of expansion, reduction and permutation are both structure-dependent, in the sense of Chomsky (1975), i.e. they apply to constituents, and category-dependent, i.e. they apply to particular classes of constituents. Thus, it is because *bara ett fall* (only one case) is a noun phrase that it may be reduced, expanded or fronted, not because it is, say, a substring comprising the third, fourth, and fifth item of another string. This means that speakers routinely treat utterances as strings of categorized items organized by relations of constituent structure.

CATEGORIAL properties of lexical items relevant to grammatical structure include part of speech and morphological properties. The lexical items of (2a) and (2b) have the following categorial properties:

(3) *ja* (I);
Pronoun (Pro): First person, singular, common gender (1, Sg, C)
de (it);
Pronoun (Pro): Third person, singular, neutre gender (3, Sg, N)
en (a);
Article (Art): Singular, indefinite, common gender (Sg, Indef, C)
ett (a);
Article (Art): Singular, indefinite, neutre gender (Sg, Indef, N)
man (man);
Noun (N): Singular, indefinite, common gender (Sg, Indef, C)
fall (case);
Noun (N): Singular, indefinite, neutre gender (Sg, Indef, C)
revbensfrakturer (fractured.ribs);
Noun (N): Plural, indefinite, common gender (Pl, Indef, C)
kommer (come);
Verb (V): Present tense (Pres)
va (was);
Verb (V): Past tense (Past)
ihåg (in.mind);
Particle (Prt)
me (with);
Preposition (P)
bara (just);

Quantifier (Q)

Thus, (2a) and (2b) will be interpreted as the following strings of categorized items:

(6) a.

ja	*kommer*	*ihåg*	*bara*	*ett*	*fall*
Pro	V	Prt	Q	Art	N
1, SG, C	PRES			SG, INDEF, N	SG, INDEF, N

b.

de	*va*	*en*	*man*	*me*	*revbens-frakturer*
Pro	V	Art	N	P	N
3, SG, N	PAST	SG, INDEF, C	SG, INDEF, C		PL, INDEF, C

CONSTITUENT STRUCTURE is by now fairly well understood (Chomsky 1968, Gazdar et al. 1985, Nichols 1986, Chomsky 1995). Any lexical item can be taken as head and construed with a number of dependents: a functional head (such as a determiner or a complementizer), at most three arguments (subjects and objects), a predicative, and any number of modifiers (attributives or adverbials). A head (with or without arguments and modifiers) may also be conjoined with another head of the same type.

In this vein, (2a) can be analyzed as successive construals of the verb *kommer* with a modifier and two arguments, and a determiner and a modifier of one of the arguments. The verb *kommer* is taken as head (h) and construed with the particle *ihåg* as modifier (m), with the pronoun *ja* as subject (s), and with the noun *fall* as direct object (o). The noun *fall*, in its turn, is construed with the article *ett* as determiner (d), and with the quantifier *bara* as modifier, yielding a sentence with a complex direct object, a noun phrase with *fall* as head.

In a co-representational syntax (Jespersen 1937, Kac 1978, Anward 1980, 1981), where structure is represented by relational templates which are mapped onto sequences of categorized words, these construals can be represented as (7).

(7)

ja	*kommer*	°	°	*ihåg*	*bara*	*ett*	*fall*
Pro	V			Prt	Q	Art	N
1, SG, C	PRES					SG, INDEF, N	SG, INDEF, N
f	h	s	h	m			o
					m	d	h

Similarly, (2b) can be analyzed as successive construals of the verb *va* with an argument and a predicative, a determiner and a modifier of the predicative, and an argument of the modifier of the predicative. The verb *va* is construed with the pronoun *de* as subject (s), and with the noun *man* as predicative (p). The noun *man*, in its turn, is construed with the article *en* as determiner (d), and with the preposition *me* as modifier. The preposition *me*, finally, is construed with the noun *revbensfrakturer* as direct object (o).

The co-representational structure of (2b) is shown in (8).

(8)

de	*va*	°	°	*en*	*man*	*me*	*revbens-frakturer*
Pro	V			Art	N	P	N
3, SG, N	PAST			SG, INDEF, C	SG, INDEF, C		PL, INDEF, C
f	h	s	h		p		
				d	h	m	
						h	o

The structures in (7) and (8) also lay out the shape of declarative main clauses in Swedish. In a Swedish declarative, or wh-interrogative main clause, the finite verb, the head (h) of the clause, comes in second position, preceded by a topicalized constituent in f position, which binds an empty element (°) in some other position. After the second position comes the position where subjects appear when something else is topicalized (s) and the position of verbs in subordinate clauses (h). In (7) and (8), it is the subject that is topicalized. Hence, there is an empty s position, bound by f. Moreover, the finite verb binds the empty second h position.

Armed with the preceding account of grammatical structure, we can now offer a more precise formulation of the initial problem of this section. How is it possible for a speaker to construct, in real time, the novel utterance

Blind, sharp-sighted stones
snore inconsistently

with something like the following structure

(9)

blind	*sharp-sighted*	*stones*	*snore*	*inconsistently*
Adj	Adj	N	V	Adv
		PL	PRES	
		s	h	m
m	m	h		

from scratch, using only parts of old utterances, and how can a hearer interpret such a novel utterance, in real time?

6. Solve et coagula: A generative account

Standard generative accounts of how it is possible to produce and understand novel utterances (Chomsky 1965) follow quite closely the ancient alchemical formula of SOLVE ET COAGULA: concrete utterances are dissolved in memory, leaving only a residue of general patterns, from which new utterances can be formed. Thus, the generative account of linguistic competence rests on the assumption in (10).

(10) Speakers extract lexical items and grammatical rules from utterances they hear and use these items and rules to produce and understand new utterances.

In early generative theory (Chomsky 1965), the utterances in (1) would have been taken to be assembled from a list of lexical items, those in (3), by means of a separately stored set of rules - for example the elementary relational templates in (11) - and interpreted by means of the same items and rules.

(11)[5] a. f h s° h° (p)

b. h o

c. h p

d. m* h m*

e. d h

[5] As before, f = foundation (topic / focus), h = head, s = subject, o = object, p= predicative, m = modifier, and d = determiner. A ° indicates an empty position, and a Kleene star (*) indicates that several instances of a relational term are permitted.

In later work, particularly in the new minimalist framework (Chomsky 1995), lexicon and grammar have been fused, following the lead of Categorial Grammar (Curry 1961, Montague 1974, Dowty 1982a, 1982b), and Tree-Adjoining Grammar (Joshi 1985, Frank & Kroch 1995). What is assumed to be extracted (acquired) are lexicalized constituent structures, constituent structures with lexically specified heads and lexical and morphological constraints on dependents.

The article *ett* (a), as it is used in (2a), the noun *fall* (case), as it is used in (2a), and the preposition *me* (with) as it is used in (2b), would be extracted as (12a), (12b), and (12c), respectively.

From these lexical templates, a 'novel' phrase, such as *me ett fall* ('with a case'), can be assembled, by means of the Merge operation of Chomsky (1995)[6]. All we have to do is substitute (12a), *ett*, for the Art constituent in (12b), retaining the d function of that constituent, and then substitute the result of merging (12a) and (12b) for the N constituent in (12c), retaining the o function of that constituent.

(12)

a.

ett
Art
NEUT
h

b.

	fall
Art	N
NEUT	NEUT
d	h

c.

me	
P	N
h	o

[6] Formally, in the framework used here, Merge allows any structure with X as head to substitute for an X slot in another structure, and take on the function of X in that structure.

The result is (13).

(13)

me	*ett*	*fall*
P	Art	N
	NEUT	NEUT
h		o
	d	h
	h	

In an analogous way, the lines *Blind, sharp-sighted stones / snore inconsistently*, structured as in (9), can be constructed and interpreted through successive retrieval and assemblage of lexicalized templates headed by the five lexical items making up those lines.

7. A holistic alternative

In the (slightly unorthodox) generative model that was outlined in the previous section, new, but rule-governed, combinations of pieces of old utterances constitute no problem at all. The model is designed to handle such cases.

However, this capacity is bought at the price of complex mapping of variable input onto invariable representations. An utterance must be broken down into minimal lexical units and reassembled from these units during perception and interpretation. As we have seen, this kind of model is ill-suited to handle real-life variability and speed of speech processing.

What we would like is a model that combines simple mapping with the kind of structures that are provided by the generative model. To do this, we assume that utterances are directly perceived, stored, and produced as structured wholes. In such a model, neither grammatical structure nor lexical items exist on their own, but are mere aspects of complex wholes. Instead of building utterances 'vertically', by segmenting the sound stream into words and fitting these words into a complex structure, hearers build utterances 'horizontally', by joining fully structured pieces together.

Assumption (10) of the generative account is thus replaced by assumption (14).

(14) Speakers store utterances they hear and produce as structured wholes and use them as models to produce and understand new utterances.

Leaving the question of where structure comes from until next section, let us explore how utterances stored as structured wholes can serve as models for new utterances.

The idea is that not only whole structures, but also highlighted parts of structures, can be recycled for further use. Let us define a simple focussing operation, Focus, which highlights parts of structures. Focus picks out part of a stored utterance as figure, treating parts that are directly related to the focussed part as immediate ground, and the remainder of the structure as unspecified background. The immediate ground can be less specified than in the original utterance. Minimally, just the relational term of the highlighted part is specified.

For example: Suppose (2b) has been stored as (8), repeated below in (15).

(15)

de	*va*	°	°	*en*	*man*	*me*	*revbens-frakturer*
Pro	V			Art	N	P	N
3, SG, N	PAST			SG, INDEF, C	SG, INDEF, C		PL, INDEF, C
f	h	s	h		p		
				d	h	m	
						h	o

If *me* (‘with’) is focussed in (15), its immediate ground will be more or less specified instances of *man* (‘man’) and *revbensfrakturer* (‘fractured ribs’), the parts of (15) that are directly related to *me*. One instance of (15) focussed on *me* is shown in (16).

(16)

me	
P	N
m	
h	o

This structure can then be combined with other focussed structures, by means a slightly different version of Merge[7]. For example, (13) can be assembled from (16) and (17), which is one version of (7) focussed on *fall* (‘case’), by substituting (17) for the o slot in (16).

[7] This version of Merge allows a structure with the external function f to substitute for a f slot in another structure, provided that the first structure does not contradict the categorial properties of the slot.

(17)

ett	*fall*
Art	N
SG, INDEF, N	SG, INDEF, N
	o
d	h

In this way, the lines *Blind, sharp-sighted stones / snore inconsistently*, structured as in (9), can be constructed from the focussed structures in (18) - which of course can 'come' from five different utterances. The procedure is as follows:

(18b) substitutes for s in (18a)

(18c) substitutes for s in (18a+b)

(18a+b+c) substitutes for s in (18d)

(18a+b+c+d) substitutes for ^ [independent utterance] in (18e)

(18) a.

blind	
Adj	N
	s
m	h
h	

b.

sharp-sighted	
Adj	N
	s
m	h
h	

c.

stones
N
PL
s
h

d.

	snore
	V
	PRES
	^
s	h

e.

	in- *consistently*
V	Adv
^	
h	m
	h

The simple operations of Focus and Merge allow speakers to recycle parts of old utterances that are closely similar to the minimal lexical items of a generative model. However, a holistic model also affords a straightforward description of more flexible and economic re-uses of linguistic experience.

For example, episodes in ordinary conversation are often built from a few utterance frames, into which the new information of each successive utterance is plugged. In the episode in which (1a) and (1b) occur, a short narrative monologue of 12 utterances, 4 of the utterances are built on the frame in (19), 4 of the utterances are built on the closely related frames in (20), 2 are built on the frame in (21), and 2 instantiate utterance formats that do not recur in the episode.

(19) (å) de va __
(‘and it was __’)

(20) åde __
(‘and it __ ‘)

å__ blev de __
(‘and __ became it __’)

(21) (å) ja kommer (inte) ihåg __
(‘and I don’t remember __)

In a holistic model of utterance production and perception, where maximally economical use is made of a stored pool of whole utterances, this kind of a structure

makes eminent sense. It is as easy to focus frames - (22), for example, which is (15) focussed on de va - as it is to focus classical lexical items. However, real-time processing is considerably speeded up by the use of larger pre-constructed chunks, instead of classical lexical items only.

(22)

de	*va*	°	°	
Pro	V			N
3, SG, N	PAST			
f	h	s	h	p

As pointed out by Chafe (1968), Pawley & Syder (1983), and Hopper (1991), among others, generative theories of linguistic competence are ill-equipped to handle such routine ways of speaking and therefore tend drastically to underestimate the extent to which ordinary speech draws on pre-constructed combinations - clichés, idioms, frozen metaphors, set phrases, reduced paradigmatic options, and frames. In contrast, a holistic theory can handle both routine and novelty.

Moreover, idiosyncratic properties of complete utterances turn out to be more important than is usually recognized in generative accounts of speech perception and production. In every speech community, there are received ways of saying things. Such received ways of saying things are a useful and powerful resource. They both speed up 'delivery', since they reduce the degrees of freedom involved in utterance planning, and serve as salient markers of social identity.

In the episode from which (1) was culled, the following two utterances appear, for example:

(23) och (p) eh de togs revben [påden]
('and it was.taken ribs on it')

sen fick patienten dågåhem [me en fast binda]
('then the patient got to go home with a fixed bandage')

These are clearly received ways of describing a certain kind of medical investigation and a certain kind of treatment. Other ways of saying the same things would most probably mark the speaker as non-professional or non-Swedish, or both.

8. Structure is post-hoc

If utterances are perceived as structured wholes, where does the structure come from? We have already argued that the perceived structure of an utterance is a consequence of its resonance with what is already stored. It is then natural to take grammatical structure to be one dimension of that resonance.

Following, among others, Hopper (1987), Elman (1995), Ochs, Schegloff & Thompson (1996), and Ford & Wagner (1996), we thus take grammar and lexicon to be emergent features of linguistic practice. In particular, we argue that grammatical structure and lexical items emerge as aspects of an utterance, when that utterance is matched with previously stored utterances.

To see what this means, consider (24), a short excerpt from a conversation between a speech therapist (L) and her patient Alma (A), an elderly woman.[8]

(24)

1.	L	men var är du **nu** då Alma
2.	A	va
3.	L	var är du **nu** då
4.	A	(p) här hit
5.	L	ja
6.	A	(p) jaa var e **ja**
7.	L	ja var e **de** (*laughter)*
8.	A:	var har ja hamna **nu** nånstans

(1.	L:	where are you **now** then, Alma;
2.	A:	what;
3.	L:	where are you **now** then;
4.	A:	(p) here to.here;
5.	L:	yes;
6.	A:	(p) ye-es where am **I**;
7.	L:	yes where is **that** (*laughter*);
8.	A:	where have I ended.up **now** somewhere;

Let us concentrate on the utterances in lines 3, 6, 7, and 8. The syntactic structures of these utterances are fairly straightforward, as can be seen from (25) below. We take the locative interrogative pronoun *var* (where) to be a preposed predicative complement of the copula *är / e*, as well as of the whole periphrastic perfect construction *har hamna* (have landed). And we introduce two new positions: one, x, for response words, other interjections, and subjunctions, and one, y, for extraposed and right-dislocated material.

[8] From a forthcoming study by Ing-Mari Tallberg. As before, boldface indicates sentence accent and (p), pause. Underlining indicates simultaneous talk and non-verbal sounds occur in italic within parentheses.

(25)

	x	f	h	s	h	p	m	y
3.		*var*	*är*	*du*		°	***nu***	*då*
6.	jaa	var	e	**ja**		°		
7.	*ja*	*var*	*e*	***de***		°		
8.		var	har	ja	hamna	°	**nu**	nånstans

These utterances demonstrate very concretely how new utterances are made from old ones. By focussing *varärdu* in 3 and substituting *ja* for *du*, 6 is made from 3, and by substituting *de* for *ja* in 6, 7 is made from 6.

Moreover, they show how structure can emerge from a sequence of utterances. If the pool of stored utterances is incrementally restructured by the simple principle (26), the fundamental 'discovery procedure' of structural linguistics (see e.g. Gleason 1961, chs. 5-7),

(26) **Segmentation**

Utterances which contain sufficiently similar parts
are segmented into that part, a preceding environment, and a following environment

then a stored unsegmented utterance XAY will both segment and be segmented by an incoming utterance ZAW:

$$XAY \Rightarrow \begin{matrix} XAY \\ ZAW \end{matrix} \Rightarrow \begin{matrix} X\text{-}A\text{-}Y \\ Z\text{-}A\text{-}W \end{matrix}$$

Likewise, a further incoming utterance ZBW will be segmented by Z-A-W:

$$\begin{matrix} X\text{-}A\text{-}Y \\ Z\text{-}A\text{-}W \\ ZBW \end{matrix} \Rightarrow \begin{matrix} X\text{-}A\text{-}Y \\ Z\text{-}A\text{-}W \\ Z\text{-}B\text{-}W \end{matrix}$$

In (24), after the utterance in line 3:

3. L var är du **nu** då
(where are you **now** then)

has been produced and received, the pool of stored utterances is simply (27).

(27)

*varärdu**nu**då*

When the utterance in line 6:

6. A (p) jaa var e **ja**
((p) ye-es where am **I**)

is produced and received, it segments and is segmented by the utterance in line 3, in accordance with principle (26). Since the vare piece of 6 is similar to the the varär piece of 3, the pool of utterances is restructured from (27) to (28).[9]

(28)

	varär	*du**nu**då*
jaa	vare	**ja**

The utterance in line 7:

7. L ja var e **de** (*laughter)*
(yes where is **that** (*laughter*))

changes (28) to (29).

(29)

	varär	*du**nu**då*
jaa	vare	**ja**
ja	*vare*	***de***

And the utterance in line 8:

8. A: var har ja hamna **nu** nånstans
(where have I ended.up **now** somewhere)

identifies both *var* (where) and *nu* (now) as recurring segments:

(30)

	var	*är*	*du*		***nu***	*då*
jaa	var	e	**ja**			
ja	*var*	*e*	***de***			
	var	har	ja	hamna	**nu**	någonstans

[9] We are cheating a little here, for brevity, by disregarding the contributions of lines 1, 2, 4, and 5.

In this way, successive utterances form networks of related expressions, where each member of a network is structured by the whole network into units which contract syntagmatic and paradigmatic relations to other units licensed by the network. The relations created by the network in (30) are displayed in more compact form in (31), where columns constitute filler paradigms for the seven slots of a single syntagm. An optional slot (or, a slot that need not be filled) is indicated by the presence of ”-” in its filler paradigm.

(31)

-	var	är	du		-	-
jaa			ja		nu	då
ja			de			
		har		hamna		nånstans

(31) easily transcends the expressive limits of rote learning, and supports, on the basis of an experience of just four utterances, 108 potential utterances (3×1×2×3×1×2×3 = 108).

9. Functions and categories are emergent

We have seen that utterances become organized as strings of phrases and lexical items, when speakers use a stored pool of utterances as an optimal resource for the construction and interpretation of further utterances.

We will now go on to suggest that syntactic functions (constituent structure, as this notion was interpreted in section 5) and grammatical categories emerge, when speakers furthermore consistently use utterances to say something about the world.

Saying something about the world involves three basic capacities (Pylyshyn 1977): (i) the ability to analyze a situation into components; (ii) the ability to analyze an utterance into components; and (iii) the ability to associate a recurring utterance component with a recurring situation component. The gist of our proposal is that the human life-world provides an environment where all of these abilities naturally unfold.

The analysis of situations into components unfolds in roughly the same way as the analysis of utterances into components, through self-segmentation of the flow of experience in a sufficiently varied, yet optimally organized life-world. And when the two processes of segmentation are synchronized in practice and experience, utterance structure acquires the kind of stability that linguists codify as syntactic functions, grammatical categories, and lexical meaning.

We have already seen how the ability to analyze utterances into components naturally unfolds in a sufficiently varied, yet optimally organized verbal practice. We can assume, following Piaget (196), that the ability to analyze situations into components unfolds, in much the same way, in a sufficiently varied, yet optimally organized life-world. Thus, in an experience where the situations in (32) successively occur,

(32) I hold ball

I taste ball

you hold ball

I hold mug

these situations become segmented, by a principle equivalent to (26), into the components of ‘hold’, ‘taste’, ‘I’, ‘you’, ‘ball’, and ‘mug’.

Consider, in contrast, the famous alarm calls of vervet monkeys, described by Marler (1980) as: bark, chirp, chatter, *nyow, rraup,* and *uh*. An alarm call is an indexical sign, which serves to record the dynamics of the ongoing situation, the situation in which the call occurs, and can not be extended to signify a remembered threat or a projected threat. In a Wierzbicka style of semantic explication (Wierzbicka 1996), the meaning of a general alarm call, for example the *uh* of vervet monkeys, might be explicated as in (33):

(33) *uh*
I say this to you about this place at this time
I sense danger
Watch out

What has made vervet monkeys famous is a set of more specialized alarm calls. Vervet monkeys seem to have distinct alarm calls for snakes, leopards, and eagles. The meaning of the alarm call for eagles, for example, might be explicated along the following lines:

(34) *rraup*
I say this to you about this place at this time
I sense danger:
I see an eagle
Flee downwards

Nevertheless, it would be wrong to say that *rraup* signifies 'eagle'. This call still signifies an ongoing global situation, although it does not signify it as 'danger' simply, but

as 'danger: eagle', and the ‘eagle’ part of (34) can not be recycled to make up part of the mening of some other call, for example a call meaning ‘Let’s play eagle!’.

We conjecture that the life-world of vervet monkeys is not varied enough for ‘eagle’ to be identified as an independent segment of experience. In our view, vervets would not be able to develop a sign for 'eagle', unless their life-world would involve them in another salient relation, besides being potential prey, to eagles. For example, if vervet monkeys also sometimes fed on eagles, then 'eagle' would stand a chance of becoming a recurring segment of vervet monkey experience, and qualify as a potential signified for this species.

Which are then the recurring segments of human praxis and experience that drive the semiotics of human languages? If we take our inspiration from a Heideggerean analysis of the human life-world (Heidegger 1927, Pöggeler 1989), where human existence means being 'thrown' into a world alongside other humans and things, in time and history, not as a passive epistemological spectator, but as a concerned participant, involved in significant interactions with those humans and things, and at the same time are careful to make room for Wierzbicka’s semantic primitives and lexical universals (Wierzbicka 1996) in it, we might come up with something like (35).

(35) **A Human Life-World**

> Being human means living in a world together with other humans and with animals, plants, and things, as a concerned participant, involved in events, i.e. significant interactions with people, animals, plants, and things. Humans, animals, plants, things, and events are situated in space, occupy places. Events have an inherent orientation in time, and being human also means being situated in time, having a history, based on memory of past events, and a future, based on projected events. Humans, animals, plants, things, and events can be judged as good or bad, big or small, one or many, present or absent, same or different, and related in time and space.

In this, admittedly crude, life-world, the flow of experience will typically be segmented into larger episodes (Korolija 1998), sequences of situations, where a human, the main protagonist of the episode, starts out being located in a particular place, at a particular time, with certain properties, and is then involved in a chain of events, alone or together with other humans and/or things, also initially located and with certain properties, which, like the location and properties of the main portagonist, may change during the course of events. In Wierzbickan terms (Wierzbicka 1996, 1998), an episode might look something like this:

This person is at this place
This is before now
This person does something to other persons
Because of this, these persons become small

Thus, an episode involves a relatively stable cast of 'actors', and an evolving plot, where these actors pass through a series of different situations, each one minimally defined by an event, a time, a place, a relation or a property. In formal semantic terms, such situation-defining phenomena are functions from actors to situations. Experienced episodes are segmented into situations, which in turn are segmented into actors and plot components.

The ability to associate a recurring utterance component with a recurring situation component naturally unfolds in an activity, when vocalizing is synchronized with the flow of experience in such a way that the most salient difference between two successive utterances is associated with the most salient difference between the two successive situations in which the utterances are embedded.

The way this is done has been known for millennia. It is the basic insight that just as plot components are taken as functions from actors to situations, segments linked to situational contrasts in plot, plot segments, are taken as functions from segments linked to situational contrasts in actors, actor segments, to utterances. Plot segments are taken as heads of utterances, and actor segments as their argument dependents (subject or object).

To see what this means, let us return to the conversation excerpt in (24). The most salient difference between the utterances in lines 3 and 6:

(36)	3.	L	var är du **nu** då (where are you **now** then)
	6.	A	(p) jaa var e **ja** ((p) ye-es where am **I**)

is the contrast between *dunudå* and *ja*. The most salient difference between the situations in which 3 and 6 are embedded is a change in speech act rôles. In 3, the speech therapist is speaker and Alma is adressee; in 6, the rôles are reversed.

However, this change in speech act role is only interpretable against the background of a shared 'understanding', in the sense of Wootton (1997), by the speech therapist and

Alma of what is currently at issue in the unfolding activity, namely the present location of Alma. A partial explication of this understanding might run as follows.

(37) This person is at this place
This is now
This person does not know where this place is

Given this background, the situational difference linked to the formal difference between 3 and 6 can be explicated as the difference between (38) and (39), as a change in the identity of an actor.

(38) This person is at this place
This is now
This person does not know where this place is
You are this person

(39) This person is at this place
This is now
This person does not know where this place is
I am this person

Consequently, *dunudå* and *ja* are actor segments and hence syntactic arguments (subjects), and *vare*, the recurring part of 3 and 6, is a plot segment and hence syntactic head. In other words, when the utterance in line 6 is produced, embedded in an episode where (38) is followed by (39), the pool of stored utterances is restructured from (27) to (40).

(40)

	h	**s**
	varär	*du**nu**då*
jaa	vare	**ja**

It would involve us in too many technicalities to exhaustively ground the changes in constituent structure that are introduced by the utterances in lines 7 and 8 in (25). Let us just concentrate on the most important of these changes, the function of *nu*, which the utterance in line 8 establishes as a recurrent segment.

Semantically, *nu* corresponds to a plot component, although one which involves another plot component, rather than a person or thing, as actor (compare the ”This is now”-part of (37) - (39)). Syntactically, however, *nu* is not head, but a predicate modifier,

a function from an utterance head to another utterance head. While (40) exemplifies the basic construction type, predication, where a plot segment is head and an actor segment is dependent, the construal of *vare* with *nu* exemplifies another, secondary, construction type, modification, where a plot segment or actor segment is construed with a dependent, which would have been the head in a corresponding predication.

Modfication is a way of packing one more predication into an utterance. For example, instead of using the three consecutive utterances

Stones snore
The stones are blind
The snoring is inconsistent

we can use the single utterance,

Blind stones snore inconsistently

by 'downgrading' *blind* to an argument modifier, and *inconsistent*, to a predicate modifier.

As a rule of thumb we can say that an additional plot segment in a predication is a modifier, either of the head of that predication or of its dependent argument.

By means of this rule, *nu* is identified as a modifier, and since it corresponds to a plot component which involves the signified of *var e* as actor, *nu* is further identified as a predicate modifier. Hence, when the utterances in lines 7 and 8 are produced, the pool of stored utterances is restructured from (40) to (41).

(41)

	h		**s**		**m**	
	var	*är*	*du*		***nu***	*då*
jaa	var	e	**ja**			
ja	*var*	*e*	***de***			
	var	har	ja	hamna	**nu**	någonstans

In this way, the synchronization of vocalization and experience allows syntactic functions and constituent structure to emerge.

The same kind of grounding also works for categorical properties of lexical items. There is a story to be told about morphological properties of lexical items, but it is much

too long to be told here. We content ourselves with a brief (but hopefully sufficient) indication of where parts of speech come from. As detailed in Anward (1995, fc.), parts of speech can successfully be reduced to patterns of spread over combinations of syntactic function and semantic category. In particular, core instances of received parts of speech can be reduced to the following combinations of function and category:

(42)

Function	**Semantic Category**	**Part of Speech**
Utterance	Situation	Interjection
Predicate	Event	Verb
Predicate modifier	Place	Adverb
	Time	Adposition
	Relation	Conjunction
	Property	
Argument modifier	Place	Demonstrative
Argument modifier	Property	Adjective
Argument modifier	Quantity	Quantifier, Numeral
Argument	Person / Thing	Personal Pronoun, Noun

Items with privileged access to one of these combinations can be labelled with the part of speech tag appropriate for that particular combination. Consequently, *ja*, *du* and *de* are labelled as pronouns, *nu* as an adverb, and *var e* and *var har* as (complex) verbs.

In sum, then, the short conversation extract in (24), embedded in the context of (37) - (39), is sufficient to structure the utterance in line 3 as in (43).

(43)

var *är*	*du*		***nu***	*då*
V	Pro		Adv	
h	s		m	

There is of course still a difference betweeen this analysis of the utterance in line 3 and the more complete analysis of the same utterance in (25), but this is a matter of degree, not of kind. Given more time and more utterances, the emergent structure of 3 will converge on the structure of (25).

10. Fast processing requires naïvism, not nativism: Summary of basic argument

We began with a review of theoretical paradigms of speech perception research focusing in particular on the criterion of processing speed. A widely shared tacit assumption was identified, *viz.*, the idea that perceptual processing involves *simple representations* (e.g., abstract invariant units such as 'phonemes', or 'gestures') but *complex mapping* (Johnson & Mullenix 1997). This observation led us to ask, 'If mapping is indeed based on complex computations (performed by say, a 'phonetic module', 'smart mechanisms', 'reconstructive rules', or by 'analysis-by-synthesis' and 'top-down' mechanisms such as 'hypothesis testing' and 'inference making'), how do those operations achieve their speed?'

To broaden the discussion, we then introduced a type of mechanism that turns the traditional assumption around making mapping simple and representations complex. The so-called *exemplar-based* models of perception and learning are instances of this sort of mechanism. Experimental phonetic evidence was presented indicating that these models exhibit a number of theoretically interesting characteristics.

For instance, perceptual processing is metaphorically like the phenomenon of 'resonance' (Licklider 1952, Shepard 1984). Incoming signal patterns simultaneously activate the entire contents of the memory. The response of a given memory pattern (=its resulting activation level) is a function of its similarity to the stimulus as measured by the extent to which that pattern 'resonates' with the input.

Such systems have several virtues. One is speed. Perceptual processing occurs the moment the memory structures are excited. Stored knowledge (primary information + context) is activated instantly by the signal (input information + context). Therefore there is strictly speaking no top-down flow of information, (no gestural recoding of the

auditory signal, no time-consuming computations, no application of reconstructive rules etc). In this respect, the models operate in a *direct* mode of perception (Fowler 1986).

Another advantage is the way such systems bypass the classical invariance problem - the difficulty of identifying invariant physical correlates of linguistic units. There are strong reasons to believe that phonetic variations are adaptations to the listener's need for information and the current communicative conditions – rather than linguistically irrelevant and variable "noise" superimposed on phonetic invariants (Lindblom 1996). On this view, speech transforms are lawful, not random and unpredictable. Regularities appear when signal patterns are interpreted in terms of segmental and prosodic context. In keeping with that scenario, exemplar models cope with the massive but systematic signal variability by retaining (storage-efficient versions of) context along with primary information. Exemplar models thus trade complex representations for explicit mapping rules. They hypothesize that 'categories' and 'rules' are implicitly present in the information stored. A specific category retains its identity in the face of all the variability, because multi-dimensional context is there to 'account for' it. In other words, context resolves ambiguity and thus guarantees the distinctness of the complex representations. (In simplified form, this was the point made in Figure 1).

We then momentarily left the phonetic domain raising the question whether exemplar models can also be applied in modeling the grammatical analysis that takes place during speech comprehension. Clearly, if the criterion of 'rapid processing' is to be met in a serious way, any proposal must necessarily also include grammar and higher levels of processing.

A few well-known objections immediately come to mind. If, as conventional wisdom has it, 'experience underdetermines the adult user's full grammatical and phonetic competence', how can an experience-based mechanism be expected to be useful in processing utterances that have never been heard before? How can it be seriously claimed that 'rote learning' of utterances is fundamental to language acquisition and that syntactic, morphological and lexical knowledge is represented only as implicit contrasts and similarities among stored exemplar utterances?

With the aid of samples o f recorded natural conversations we proceeded to present a detailed analysis aimed at countering these objections. We showed how a pool of stored,

initially unanalyzed utterances were incrementally restructured and decomposed by a principle of *self-segmentation*. After elaborating on this principle, which works by extracting 'recurring parts that make similar contributions to meaning', we were able to conclude that grammatical structures are indeed derivable from the regularities in the input and the context. Obviously utterances do not come specified with syntactic labels such as 'head', 'subject', 'predicative', 'object', 'modifier', 'determiner' etc, but – so the argument goes – such properties nonetheless become available, albeit they remain contextually embedded and therefore are defined only implicitly. 'Units' and 'rules' come into existence gradually as more and more exemplars accumulate in memory, and as the distributional properties of the stored materials become more and more salient. They are novel attributes, *emergents*, of the listener's cumulative perceptual experience.

To illustrate how it might be possible to extrapolate from old and limited information to novel cases, we considered the processing of a deliberately contrived phrase: *Blind sharp-sighted stones snore inconsistently*. Although a native speaker of English might arguably never have heard this particular line before, he would nevertheless succeed in parsing it and classifying it as 'grammatically well-formed'. Should we follow Chomsky (1957) in interpreting this observation to show that grammaticalness has nothing to do with utterance statistics? No, according to the present account, the native speaker's judgement derives from the fact that he correctly spots the constituent words. The identification of a word automatically links it with other elements having identical syntactical properties. This arises from the fact that hearing a word activates that word's phonetic template, but also its 'word class', that is the 'cohort' of which the word - for distributional reasons – is a member. Although cohorts do not have explicit labels, it appears reasonable to suggest that a given sequence of words sets up a 'neural trace' connecting the cohorts implicated by those words. This is the basis of the system's ability to extrapolate to new inputs. For '*Blind sharp-sighted stones snore inconsistently*', a trace would arise corresponding to the syntactic string of **Adj Adj N V Prt**. The obvious questions are, Did the listener ever come across that particular syntactic pattern? The answer is of course affirmative for the present case. Hence the utterance gets categorized as grammatical. Does the utterance make sense? Here the response is no, so the example is judged to be semantically anomalous.

Needless to say, exemplar models have a long way to go before we know how successful they will be. Nevertheless, in comparison with many of the traditional approaches they possess several attractive features.

Perhaps most significantly, their application imposes a principle of *methodological parsimony* which forces the investigator to make the most of the information in the input before resorting to extravagant assumptions about the cognitive structures that *'must'* be pre-specified for the language learner. By adopting a sort of *naïvism* instead of the strong *nativism* of e.g., Universal Grammar, we are able to propose that phonetic and grammatical structure does not necessarily have to come pre-specified by the learner's genetic endowment. It might as well arise developmentally from an interaction between the listener's cumulative perceptual experience and very general biological preconditions not necessarily specialized for Language.

Again, we should note that not only do exemplar-based systems have the necessary temporal characteristics. They are also in principle capable of automatically sorting out the acoustic context-dependence of phonetic segments. And they permit grammatical extrapolations to previously unheard, but well-formed utterances. Perceptual analyses occur without 'inference making' or 'hypothesis testing' operating with minimal demands on cognitive operations, i.e., without, strictly speaking, any 'top-down' processing.

In our opinion, that seems to be the kind of perceptual processing that both normal language use and simultaneous interpreting could use!

11. References.

Anward J (1980): ” Från yttrandeplanering till social struktur”, in B. Brodda & G. Källgren (eds): *Lingvistiska perspektiv*, Stockholm, 15-56

Anward J (1981): *Functions of passive and impersonal constructions. A case study from Swedish*, Diss., Dept of Linguistics, Uppsala University.

Anward J (1995): ”The Dao of Lexical Categories. Towards a Typology of Part-of-Speech Systems”, paper presented at a conference on 'Functional Approaches to Grammar', Albuquerque, New Mexico, July 1995.

Anward J (in press): ”Parts of Speech”, to appear in *Language Typology and Language Universals. An International Handbook*, Berlin, Mouton De Gruyter

Bregman A S (1990): *Auditory Scene Analysis*. Cambridge, MA, MIT Press.

Browman C & Goldstein L (1992): "Articulatory phonology: An overview", *Phonetica* 49:155-180.

Chafe W (1968): ”Idiomaticity as anomaly in the Chomskyan paradigm”, *Foundations of Language* 4: 109-127.

Chomsky N A (1957): *Syntactic structures*, The Hague, Mouton.

Chomsky N A (1965): *Aspects of the Theory of Syntax*, Cambridge, Mass, MIT Press.

Chomsky N A (1970): ”Remarks on Nominalization”, in Jacobs R A & Rosenbaum P S (eds): *Readings in English Transformational Grammar*, Waltham, Ginn and company, 184-221.

Chomsky N A (1975): *Reflections on Language*, New York, Pantheon.

Chomsky N A (1995): *The Minimalist Program*, Cambridge, Mass., MIT Press.

Curry H B (1961): ”Some logical aspects of grammatical structure”, in Jakobson R (ed): *The Structure of Language and its Mathematical Aspects*, Providence, American Mathematical Society

Dowty D R (1982a): ”Grammatical relations and Montague Grammar”, in Jacobson P & Pullum G K (eds): *The Nature of Syntactic Representation*, Dordrecht, Reidel, 79-130

Dowty D R (1982b): ”More on the categorial analysis of grammatical relations”, in Zaenen A (ed): *Subjects and Other Subjects*, Bloomington, Indiana University Linguistics Club.

Elman J L (1995): ”Language as a dynamical system”, in Port R F & van Gelder T (eds): *Mind as Motion: Explorations in the Dynamics of Cognition*, Cambridge, Mass., MIT Press, 195-223

Estes W K (1993): “Concepts, categories, and psychological science”, *Psychological Science 4*, 143-153.

Ford C & Wagner J, eds. (1996): *Interaction-based studies of language*, special issue of *Pragmatics* (6:3).

Fowler C A (1986): "An event approach to the study of speech perception from a direct-realist perspective", *Journal of Phonetics 14(1)*, 3-28.

Fowler C A (1991): "Auditory perception is not special: We see the world, we feel the world, we hear the world", *J Acoust Soc Am 89(6)*, 2910-2915.

Fowler C A (1994): "Speech perception: Direct realist theory", in Asher R E (ed): *Encyclopedia of Language and Linguistics*, Pergamon:New York, 4199-4203.

Frank R & Kroch A (1995): ”Generalized transformations and the theory of grammar”, *Studia Linguistica* 49: 103-151.

Gazdar, Klein, Pullum & Sag (1985): *Generalized Phrase Structure Grammar*, Oxford, Blackwell.

Gerver D (1976): “Empirical studies of simultaneous interpretation: a review and a model”, 165-207 in Brislin R W (ed): *Translation. Application and Research*, Gardner Press:New York.

Gibson J J (1972): "Outline of a Theory of Direct Visual Perception", in Royce, J R & Rozeboom, W W (eds): *The Psychology of Knowing*, Gordon&Breach:New York.

Gibson J J (1979): *The Ecological Approach to Visual Perception*, Houghton Mifflin:Boston, MA.

Gleason H A (1961): *An Introduction to Descriptive Linguistics*, 2nd ed. New York: Holt, Rinehart and Winston.

Grieser D & Kuhl P A (1989): "Categorization of speech by infants: Support for speech-sound prototypes", *Developmental Psychology 24(4)*, 577-588.

Grosjean F (1980): “Spoken word recognition processes and the gating paradigm”, *Perception & Psychophysics 28(4)*, 267-283.

Harris Z S (1957): ”Co-occurrence and transformation in linguistic structure”, *Language* 33: 283-340.

Heidegger M (1927): *Sein und Zeit*, Tübingen.

Hintzman D L (1986): “’Schema abstraction in a multiple-trace model’”, *Psychological Review 93*, 411-428.

Hopper P J (1987): ”Emergent Grammar”, *BLS* 13: 139-157.

Hopper P J (1991): ”Dispersed Verbal Predicates in Vernacular Written Narratives”, *BLS* 17.

Jakobson R (1967): ”Boas’ View of Grammatical Meaning”, reprinted in *On Language*, ed. by Waugh L R & Monville-Burston M, Cambridge, Mass, Harvard University Press, 1990, 324-331.

Jakobson R, Fant G, & Halle M (1963) *Preliminaries to Speech Analysis*, Cambridge, MA: MIT Press, (Originally published in 1951.)

Jespersen O (1937): *Analytic syntax*, reprinted 1969, New York, Holt, Rinehart and Winston.

Johnson K (1997): “Speech perception without speaker normalization: an exemplar model”, chapter 8 in Johnson K & Mullenix J (eds): *Talker variability in speech processing*, Academic Press.

Johnson K & Mullenix J (1997): “Complex representations used in speech processing: Overview of the book”, 1-8 in Johnson K & Mullenix J (eds): *Talker variability in speech processing*, Academic Press.

Joshi A K (1985): ”How much context-sensitivity is required to provide reasonable structural descriptions: tree adjoining grammars”, in Dowty D, Karttunen L & Zwicky A (eds): *Natural language parsing*, Cambridge, Cambridge Univesity Press, 206-250.

Kac M B (1978): *Corepresentation of grammatical structure*, London, Croom Helm.

Klatt D H (1979): "Speech perception: A model of acoustic-phonetic analysis and lexical access", *J of Phonetics 7*, 279-312.

Klatt D H (1989): "Review of selected models of speech perception", 169-226 in Marslen-Wilson W (ed): *Lexical representation and process*, Cambridge, Mass:MIT Press.

Kluender K R, Diehl R & Killeen P (1987): "Japanese quail can learn phonetic categories", *Science 237*, 1195-1197.

Korolija N (1998): *Episodes in Talk*, Linköping, Linköping Studies in Arts and Science

Kuhl P K (1992): "Innate predispositions and the effects of experience in speech perception: The native language magnet theory", 259-274 in de Boysson-Bardies B, de Schonen S, Jusczyk P, MacNeilage P & Morton J (eds): *Developmental neurocognition: Speech and face processing in the first year of life*, Kluwer:Holland.

Lashley K S (1951): "The problem of serial order in behavior", 112-146 in Jeffress L A (ed): *Cerebral mechanisms in behavior*, New York:Wiley.

Liberman A M, Delattre P C, Cooper F S & Gerstman L J (1954): "The role of consonant-vowel transitions in the perception of the stop and nasal consonants", *Psychological Monographs 68*, 1-13.

Liberman A & Mattingly I (1985): "The motor theory of speech perception revised," *Cognition 21*, 1-36.

Liberman A & Mattingly I (1989): "A specialization for speech perception," *Science 243*, 489-494.

Licklider J (1952): "On the process of speech perception", *J Acoust Soc Am 24(6)*, 590-594.

Lindblom B (1996): "Role of articulation in speech perception: Clues from production", *J Acoust Soc Am 99(3)*, 1683-1692.

Lindblom B, Lubker J & Gay T (1979): "Formant frequencies of some fixed-mandible vowels and a model of speech programming by predictive simulation", *J Phonetics 7*, 147-162.

Marler P (1980): "Primate Vocalization: Affective or Symbolic?", in Sebeok Th A & Umiker-Sebeok J (eds): *Speaking of Apes*, New York and London, Plenum Press, 221 - 230

Montague R (1974): *Formal philosophy*, edited by R. Thomason, New Haven: Yale University Press.

Nichols J (1986): "Head marking and dependent marking grammar", *Language* 62: 56-119.

Nosofsky R M (1988): "Exemplar-based accounts of relations between classification, recognition and typicality", *J Exp Psych: L, M & C*, 700-708.

Ochs E, Schegloff E A & Thompson S A, eds. (1996): *Interaction and grammar*, Cambridge, Cambridge University Press.

Öhman S (1966): "Coarticulation in VCV utterances: Spectrographic measurements", *J Acoust Soc Am 39(1)*, 151-168.

Öhman S (1967): "Numerical model of coarticulation", *J Acoust Soc Am 41*, 310-320.

Oléron P & Nanpon H (1964): "Recherches sur la traduction simultanée", *Journal de Psychologie Normal et Pathologique 62*, 73-94.

Pawley A & Syder F H (1983): "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency", in Richards J & Schmidt R (eds): *Language and Communication*, London, Longman.

Piaget J (1968): *Le structuralisme*, Que sais-je?, Paris, Presses Universitaires de France

Pylyshyn Z W (1977): ”What does it take to bootstrap a language?”, in J. Macnamara (ed): *Language Learning and Thought*, New York etc.: Academic Press, 37-45.

Pöggeler O (1989): *Martin Heidegger's Path of Thinking*, Atlantic Highlands, Humanities Press International.

Rumelhart D & McClelland J (1986, eds): *Parallel Distributed Processing*, Vol 2, MIT, Cambridge MA.

Ryalls J & Lieberman P (1982): “Fundamental frequency and vowel perception”, *J Acoust Soc Am 72*, 1631-1634.

Shepard R N (1984): "Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking and dreaming", *Psychological Review 91(4)*, 417-447.

Stetson R H (1951): *Motor phonetics a study of speech movements in action*, North Holland:Amsterdam.

Strange W (1989a): "Evolving theories of vowel perception", *J Acoust Soc Am 85*, 2081-2087.

Strange W (1989b): "Dynamic specification of coarticulated vowels spoken in sentence context", *J Acoust Soc Am 85*, 2135-2153.

Studdert-Kennedy M (1987): "The phoneme as a perceptuomotor structure", in Allport A, MacKay D, Prinz W & Scheerer E (eds): *Language, Perception and Production*, New York:Academic Press.

Studdert-Kennedy M (1989): "The early development of phonology", in von Euler C, Forsberg H & Lagercrantz H (eds): *Neurobiology of Early Infant Behavior*, New York:Stockton.

Sussman H M, McCaffrey H A & Matthews S A (1991): "An investigation of locus equations as a source of relational invariance for stop place categorization", *J Acoust Soc Am 90*, 1309-1325.

Sussman H M, Fruchter D, Hilbert J & Sirosh J (1998): “Linear correlates in the speech signal: the orderly output constraint”, *Behavioral and Brain Sciences 21*, 241-299.

Treisman A (1965): “The effects of redundancy and familiarity on translating and repeating back a foreign and a native language”, *British Journal of Psychology 56*, 369-379.

Warren & Warren (1970): "Auditory illusions", *Sci Am 223(6)*, 30-36.

Wierzbicka A (1996): *Semantics. Primes and Universals*, Oxford, Oxford University Press.

Wierzbicka A (1998): ”Anchoring linguistic typology in universal semantic primes”, *Linguistic Typology* 2: 141-194

Wootton A (1997): *Interaction and the development of mind*, Cambridge, Cambridge University Press.

Zlatev J (1997): *Situated Embodiment*, Diss, Stockholm, Dept of Linguistics, Stockholm University.