UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

www.um.es/ijes

Where do Phonemes Come from? A View from the Bottom[1]

JOHN R. TAYLOR*
University of Otago, New Zealand

ABSTRACT

Infants have a remarkable ability to perceive all manner of phonetic contrasts. The phonological categories of a language, however, have to be learned from experience. Two learning paradigms are contrasted – supervised learning (where learners receive feedback on their categorization attempts) and unsupervised learning (where learners rely only on properties of the input). It is argued that unsupervised learning may be the appropriate paradigm, at least for the initial stages of acquisition. Thereafter, the emergence of phoneme categories draws on various kinds of knowledge available to the learner, including knowledge of articulation, and of literacy conventions. A concluding section emphasizes the taxonomic nature of the phoneme, and suggests that the special salience of a phonemic representation reflects the status of the phoneme as a basic level category.

KEYWORDS: phoneme; perception; structuralism; categorization; unsupervised learning; basic level

* ***Address for correspondence***: John R. Taylor. University of Otago, New Zealand. Department of English (Linguistics Programme). Division of Humanities, University of Otago, PO Box 56 Dunedin, New Zealand. Phone: 64 3 479 8952, Fax 64 3 479 8558. e-mail: john.taylor@stonebow.otago.ac.nz

I. INTRODUCTION

In a well-known passage, Saussure commented on what our mental experience would be like if we did not possess language:

> Psychologiquement, abstraction faite de son expression par les mots, notre pensée n'est qu'une masse amorphe et indistincte. … [S]ans le secours des signes, nous serions incapables de distinguer deux idées d'une façon claire et constante. Prise en elle-même, la pensée est comme une nébuleuse où rien n'est nécessairement délimité. Il n'y a pas d'idées préétablies, et rien n'est distinct avant l'apparition de la langue (Saussure, 1915: 155).[1]

Without language, Saussure claimed, thought would be inherently featureless and unstructured. For Saussure, it was language – more specifically, the conceptual categories symbolized by language – that gave structure to the amorphous substance that is prelinguistic thought. Saussure made an analogous claim about the sound substance of language. Without the mediation of a language and its phonological system, the speech signal would be equally indistinct and formless:

> La substance phonique n'est pas plus fixe ni plus rigide; ce n'est pas un moule dont la pensée doive nécessairement épouser les formes, mais une matière plastique qui se divise à son tour en parties distinctes pour fournir des signifiants dont la pensée a besoin (Saussure, 1915: 155).[2]

We can, of course, only speculate about the mental life of a person without language – a new born infant, for example, or a wild child. We are on firmer ground when it comes to our perception of the speech signal in ignorance of the linguistic categories which it encodes. When we are listening to a language which is totally unknown to us, Saussure's metaphor of the 'nébuleuse', where nothing seems clearly delineated, seems particularly apt. Learning the language consists inter alia of learning to 'make sense' of the acoustic signal, segmenting it into distinct units, classifying the units and their combinations, and, ultimately, recognizing in the signal the expression of meaningful words and phrases. This is a process which each child (with the exception, of course, of the profoundly deaf) must go through. In this paper, I comment on some aspects of this remarkable achievement, with special focus on the emergence of segmental categories.

II. PHONOLOGICAL UNITS

One might suppose that mastering the sound system of a language would consist in learning to make progressively finer perceptual distinctions amongst the sounds that one encounters in the acoustic signal. Much ingenious experimentation, however, has demonstrated that this is not how first language acquisition proceeds. It is now well established that newborn infants are exquisitely sensitive to speech sounds, being able to discriminate all manner of contrasts which are utilized in the various languages of the world (Aslin *et al.*, 1998; Aslin *et al.*, 1981; Eimas *et al.*, 1971; Jusczyk, 1997; Kuhl, 1987; Werker & Tees, 1984). While this remarkable ability surely facilitates entry into the sound system of whatever language a child is going to learn, the ability to discriminate sounds is not sufficient for phonological acquisition to take place. A person able to perceive all manner of acoustic-phonetic differences would be rather like Luria's (1968) mnemonist, or the fictitious Funes of Borges's (1964) story – individuals with a phenomenal ability to notice and remember every detail of their experiences but who, as a consequence, are unable to generalize and form abstractions. For speech perception to get under way, it is necessary for *categories of acoustic events* to be recognized in the kaleidoscope of auditory impressions.[3] Some chunks of the acoustic signal need to be regarded, in the phonological system that is being acquired, as being 'the same' as other chunks. The first question we need to ask, therefore, concerns the nature of these chunks that the learner needs to identify. There are at least three plausible candidates with regard to the linear segmentation of the speech signal (the list is not exhaustive, and the kinds are not mutually exclusive): Words, syllables, and parts of syllables.

That competent hearers of a language perceive words in the stream of speech is self-evident. Listening to speech is essentially a matter of listening for words, and word-like units, and learning a language involves, amongst other things, learning the sound shapes of words.[4] Indeed, Jusczyk (1997: 108) suggests that the identification of words in the stream of speech is what "speech perception capacities are ultimately intended for", while others have proposed that the learner's identification of word-sized units may well bootstrap the whole language acquisition process (Beckman & Edwards, 2000; Beckman & Pierrehumbert, 2000).

During the earliest stages of language acquisition, it may well be the case that words are learned, stored, and retrieved as phonological wholes, without internal analysis (Jusczyk, 1997; Vihman, 1996). While a reliance on gestalt storage might be viable at a time when the child's linguistic repertoire consists of at most a couple of dozen items, the increasing size of the child's lexicon necessitates other, or additional storage modalities. This is because the number of holistic sound shapes that a person could reliably differentiate and commit to memory is severely limited. As the size of the lexicon increases, some kind of internal analysis of the word-sized units becomes necessary. Thus, pieces of one word might be identified with pieces of other words, the pieces themselves might in turn be broken down into even smaller units. In this way, a relatively small inventory of phonological units, and patterns for their combination, will be able to support the learning of a large and ever expanding lexicon.[5]

Candidates for the internal analysis of words are syllables, parts of syllables (such as onsets and rhymes), and, ultimately, consonant and vowel segments. Syllables, as units of analysis, would seem to be especially appropriate for languages such as Japanese and Māori, where the number of possible syllables in the language is quite limited. This is reflected in the katakana and hiragana writings systems of Japanese, in which each syllable is represented by a distinct symbol (exactly 46 are needed.). When the number of different syllables in a language increases, internal analysis once again becomes necessary. Thus, traditional accounts of Mandarin phonology analyze the 400 or so occurring syllables (this number disregards tonal differentiation) in terms of the combination of initials and finals, i.e. onsets and rhymes. For English, and other languages with complex syllable structures, in which the number of different syllables runs into the thousands, further analysis is necessary, namely into the individual phonemes (or, perhaps better, the positional allophones) which make up the syllables.

Words, syllables, and phonemes/allophones, as units of perception and representation, all raise the same problem, namely, that of acoustic variability. A word, syllable, or phoneme can be pronounced in a virtually unlimited number of ways according to the linguistic context of the unit (its immediate phonetic environment, its place within an intonation contour, the overall rate of speech, etc.) as well as speaker-dependent properties (dialect, gender, age, speaker-specific properties of the vocal tract, and even such factors as the state of the speaker's dentures).

Bloomfield (1933) had supposed that various manifestations of a phoneme would share some common acoustic features. The invention of spectrographic analysis in the 1940's, however, and early attempts to synthesize speech by concatenating invariant segments, brought home to phoneticians in a particularly dramatic way the lack of acoustic invariance associated with the units that we hear in the speech signal (Potter *et al*., 1947). Liberman and his colleagues (Liberman *et al*., 1967; Liberman & Mattingly, 1985) developed their 'motor theory' of speech perception largely in response to this state of affairs. Specifically, they sought to locate invariance, not in the signal itself, but rather in the motor commands which gave rise to the acoustic signal. Later versions of the theory located invariance, not in the motor commands themselves, but in a speaker's "intended phonetic gestures" (Liberman & Mattingly, 1985: 2), thereby pushing the invariants into a domain which in principle is out of reach of empirical observation.

The invariance problem is a familiar one to categorization researchers. In fact, the (largely unsuccessful) search for acoustic constants in the speech signal following the invention of spectrographic analysis is merely a variation on the theme of the non-viability of 'classical' categories in general. Classical categories, it will be recalled, are defined in terms of a set of necessary and sufficient features. Especially from the 1970's onwards, it became apparent that most categories that people operate with – for example, the categories that are conventionally named by the lexemes of their language – are not in fact susceptible to classical definitions; moreover, the features which supposedly define the categories are subject to the very same problem (Taylor, 2003b, to apppear). In light of these findings, various alternative models of categorization were developed. These included prototype models (in which categories are centred around 'good examples'), probabilistic models (in which categories are defined in terms of weighted probabilities of features), and exemplar models (where categories are constituted in terms of the similarity of already encountered instances). In view of this extensive research (reviewed in Murphy, 2002; see also Mompeán, 2002: Ch. 1) it should come as no surprise that phonemes, syllables, and words should also resist definition in terms of sets of invariant acoustic features.

III. SUPERVISED OR UNSUPERVISED LEARNING?

The 'discovery' of the phoneme has been described as "one of the most magnificent achievements of linguistic science" (Krámský, 1974: 7). The hyperbole of this statement conceals the fact that the phoneme concept is by no means a modern invention. It is the basis of all alphabetic writing systems (though, to be sure, few writing systems are consistently phonemic), and even speakers of unwritten languages are reported to have intuitive access to the phonemic structure of words.[6] Symptomatic of the popular acceptance of the notion is the fact that most monolingual and bilingual dictionaries nowadays give word pronunciations in some form of phonemic transcription. Yet, like many of the most basic concepts of linguistics – such as 'word', for example – a concise definition remains elusive, and indeed the phoneme concept has been, and remains, the subject of intense and ongoing theoretical controversies. Later in the paper I will touch on generative phonologists' rejection of the need for a distinct phonemic level of representation. In the meantime, I focus on some of the controversies which engaged the linguistic community in pre-generative days. Indeed, a glance at the journals of the time – as well as at the contents page of Joos's (1957) influential *Readings in linguistics* – gives the impression that the history of North American linguistics during the mid decades of the last century was in large part a confrontation with the problematics of the phoneme concept.

A major issue in pre-generative times concerned the criteria by which the phonemes of a language are to be established. One of the orthodoxies of the time was the prohibition on the 'mixing of levels' (Bloch, 1948; Hockett, 1942). The idea was that the investigation of a language should proceed in a strictly 'bottom-up' fashion. The investigating linguist first made detailed phonetic transcriptions of a corpus of native speaker utterances. Observation of the distribution of phonetic segments ('phones') would then permit the allocation of these segments to a fixed set of phonemes, accompanied by statements for the possible realizations of each phoneme in various contexts. Importantly, phonemic analysis was to be conducted without any reference to 'higher' levels, such as the words and morphemes of the language, nor, or course, to their meanings.[7] Subsequently, linguistic analysis would proceed to the identification of

allomorphs and their allocation to morphemes (again, without reference to their meaning), followed by the identification of word classes, syntactic patterns, and so on (Harris, 1951).

These discussions (for a review, see Heitner, 2005) may strike the modern reader as very arcane. Pike (1947), for one, ventured to state that no field linguist would ever proceed in the way demanded by the orthodoxy of the time, by ignoring meaning and strictly excluding any 'top-down' analysis. Nevertheless, I would suggest that the issues that were discussed in the 1940's and 1950's do relate to a matter which is very much of modern concern. Updating the discussions of more than half a century ago into more modern terminology – and fudging the distinction between the linguist's analytic procedures and the processes of language learning by children (and by machines)[8] – the question would be whether phoneme categories can emerge in unsupervised as opposed to supervised learning conditions. In supervised learning, the learner (whether human or machine) is presented with a set of stimuli which are labeled as members or non-members of the target category or categories (the labeling may take the form of feedback on the correctness or otherwise of the learner's attempts at categorization). Subsequently, the learner may be tested on new stimuli, which are presented without labeling or feedback, with the aim of determining how well the categories have been learned, and how 'ambiguous', or otherwise problematic stimuli will be handled. In unsupervised learning, on the other hand, the learner is simply presented with a set of stimuli and is required to group them into categories. The stimuli are not labeled, no feedback is provided, nor is the learner given any hints as to how many or what kinds of categories are to be formed.

It will be apparent that a strict application of the dogma of the separation of levels is in essence a prescription for unsupervised learning. Indeed, linguists of the time were much concerned with developing a set of 'discovery procedures' – that is, a set of algorithms – which would correctly, and 'automatically', identify the phonemes of a language, given only a narrow phonetic transcription. The phonemic analysis would 'emerge' from the phonetic properties of a corpus, without the analyst needing to be aware that two phonetically similar stretches were merely variant pronunciations of the same word (i.e., that the pronunciations were in free variation), or whether they in fact constituted pronunciations of different words (i.e., constituted

minimal pairs). If access to this latter kind of information were to be available, we would be in the domain of supervised learning.

The period in question – the 1940's and 1950's – is commonly referred to as the heyday of Bloomfieldian linguistics, reflecting the towering influence of Bloomfield's monograph *Language* (1933). It may be interesting, therefore, to recall Bloomfield's position on phonemic analysis. We have already referred to Bloomfield's belief that a sufficiently sophisticated acoustic analysis would eventually reveal the invariant properties definitional of each phoneme of a language. For Bloomfield, however, the search for these invariant properties could not be basis of phonemic analysis. Rather, linguistic analysis was based on what for Bloomfield was the "fundamental assumption of linguistics", namely, that "in every speech-community some utterances are alike in form and meaning" (1933: 78). Thus, according to Bloomfield,

> …even a perfected knowledge of acoustics will not, by itself, give us the phonemic structure a language. We shall always have to know which of the gross acoustic features are, by virtue of meanings, 'the same', and which are different for the speakers (Bloomfield, 1933: 128).

The irony of Bloomfield's position has not escaped some commentators (Harris, 1973; Taylor, 2003b). Bloomfield, who was so intent on excluding 'mentalistic' notions, such as 'meanings', from linguistic analysis, had to postulate 'sameness of meaning' as a prerequisite for any linguistic analysis at all. Be that as it may, the relevance of Bloomfield's observation to the present topic will be evident. Bloomfield was proclaiming the impossibility, in principle, of unsupervised language learning.

And, indeed, common sense would seem to be on Bloomfield's side. Consider the acquisition of word meanings. The child encounters a range of creatures of different shapes, sizes, colours, and habitats, and exhibiting different temperaments and behaviours. Some of these creatures are referred to as 'dogs', others bear different labels, such as 'cat', 'rabbit', 'cow', 'mouse', as well as 'animal' and 'pet', The child's task, now, is to work out the criteria for this classification, on the assumption that the different uses of *dog* are 'the same in meaning', that is, that they designate one and the same category of entities. It was in such terms that Brown (1958)

presented 'The Original Word Game'. On this account, the learner's task would be exactly analogous to that confronted by subjects participating in a supervised learning experiment. Some such process would seem to be indicated, if only because different speech communities typically categorize the environment in different ways. For example, English, French, and German distinguish 'rats' from 'mice', whereas Italian does not, both kinds of creature bearing the label 'topo'. Language-specific categories presumably do not, and could not, emerge from simple observation of the world; they have to be transmitted from one generation of speakers to the next by engagement in the 'word game'.

Appealing as it is, the 'word game', and the parallels with supervised learning, may not be the whole story. In a supervised learning experiment, a subject is presented with an array of experimental stimuli and is explicitly informed about their category membership. The counterpart of this situation in language acquisition would be that a word is explicitly associated with its referent on each occasion of its use. Yet it is not always the case that words, even words which designate easily observable entities, are uttered in the presence of their referents, and even when they are, the child still has to figure out just which features of the environment are to be matched with a given word. Gleitman (1990), in addressing the common belief that words are learned by ostension, urges us to 'look and see' whether words are indeed spoken in situations in which their referents are perceptually salient to the learners. She concludes that, in many cases, they are not. Indeed, detailed observations by Gleitman suggest that learning by ostension may actually be the exception rather than the rule. And in the case of words designating 'abstract' entities and processes, such as *think, believe*, and *know*, the words' referents may not be candidates for ostension at all. The task faced by the language learner, then, is not simply one of working out the correct categorization of an array of labeled stimuli. The learner must first discover what the stimuli are that are to be categorized.

The matter becomes more complicated still when we bear in mind that word learning is not only a question of learning semantic categories, the word forms themselves have to be learned. The learner, namely, has to realize that the multifarious ways in which *dog* can be pronounced all count as pronunciations of 'the same word'. The learner could, in principle, explore the hypothesis that variations in the duration of the vowel, or whether the final consonant

is released or not, might correlate with meaning differences, e.g. big dogs vs. small dogs, brown dogs vs. spotted dogs, well-behaved dogs vs. yapping dogs. Children, presumably, do not systematically explore these possible correlations between form and meaning, any more than the field linguist would test each of the myriad hypothetical senses of *gavagai*, in Quine's (1960) well-known example.[9] As Bloomfield stated, the learner would need to be apprised of the fact that the various pronunciations are, indeed, 'the same in form', as well as being 'the same in meaning'.

There are, to be sure, certain circumstances in which a learner might be explicitly alerted to the fact that different pronunciations count as 'the same', while other pronunciations are 'not the same', as, for example, when second language learners are being trained on the discrimination of minimal pairs (*ship* vs. *sheep*, and the like). The extent of this practice with children acquiring their native language is probably quite limited, and is likely to be restricted, in any case, to older children perceived to be suffering from delayed development. (We should bear in mind, also, that languages are acquired in all manner of socio-cultural settings. Whether or not children are coached in matters of pronunciation, they all – barring pathological cases – end up with adult mastery of the ambient language.) One possibility might be that learners themselves 'discover' the existence of minimal pairs, by noting, for example, that the pronunciations of *coat* refer to one kind of entity, while the pronunciations of *goat* refer to a quite different kind of entity. The need to make the conceptual distinction would therefore trigger awareness of the corresponding phonological categories. Some researchers have indeed suggested some such mechanism of phoneme acquisition (Werker & Tees, 1984).[10]

There are, however, a number of theoretical and empirical problems associated with the view that phoneme categories emerge on the back of minimal pairs. In the first place, while the existence of minimal pairs might be diagnostic of phoneme categories, it must fail as a definition of the phoneme. In English, there are scarcely any minimal pairs contrasting [ʃ], and [ʒ], or [θ] and [ð], yet we would still want to regard these sounds as belonging to different phonemes of English.[11] Moreover, the existence of minimal pairs will be largely a matter of the size of a person's lexicon. For young children, with very small vocabularies, minimal pairs, for any pair of

candidate sounds, are vanishingly rare. Caselli et al. (1995) list the first 50 words produced and understood by both English-speaking and Italian children. The English lists contain no minimal pairs, while the Italian lists contain only *nonna* 'granny' and *nanna* 'sleep'. Even more telling is the fact that by the age of 1, children are already well on their way to perceiving the ambient language 'phonemically' (Jusczyk, 1997), that is, they are categorizing the ambient speech sounds in line with the phonological structure of the language they are to acquire. At this stage, children have scarcely learned any words of their language at all, so cannot be relying on lexical contrasts. Once again, we are forced to the conclusion that the supervised learning paradigm – where learners have the task of categorizing labeled stimuli – simply fails to apply.

The role of supervised learning (or, rather, its absence) turns up in connection with yet another issue in language acquisition research, namely, the problem of negative evidence (Bowerman, 1988; Pinker, 1984). In supervised category learning, learners receive feedback on whether their classification of a stimulus is correct or not. Yet when it comes to the learning of the syntactic structures of their language, children are rarely given information on which of their utterances are grammatically ill-formed. Caretakers may comment on the factual correctness of a child's utterance, or on its stylistic or pragmatic appropriateness, but rarely, or not systematically, on its grammatical properties. A question that has much concerned researchers in language acquisition, therefore, is how a child comes to 'unlearn' the generalizations which give rise to utterances such as *They didn't wented*, or *He said me no*. It clearly will not do to say that the learner comes to regard these expressions as ungrammatical because they are never encountered in the input. Many things that speakers say are unique creations, never before encountered, but are not, for that reason, to be rejected as ungrammatical. One factor that seems to be involved is the child's working assumption that languages avoid synonymy (Clark, 1987). The learner comes to regard her own utterances as ill-formed to the extent that they are pre-empted by alternative wordings encountered in the input (Tomasello, 2003). Whatever the plausibility of this account, it is clear that learners must work out the properties of syntactic constructions largely on the basis of the input, its properties, and their analysis of it, not from explicit instruction or feedback on grammaticality.

The above considerations all point in the same direction, namely, that the supervised learning paradigm may not be applicable to first-language acquisition. Words do not come tagged with their semantic and phonological categories, nor is information provided on which utterances count as 'the same' in form and in meaning. I will not, in the following, pursue the question of the learning of semantic categories. With respect to phonological categories, however, there are grounds for taking seriously the reality of unsupervised learning, exactly as the structuralist insistence on the separation of levels entailed.

IV. UNSUPERVISED LEARNING OF PHONOLOGICAL CATEGORIES

Categorization has been a major research topic in cognitive psychology; for a review of the by now voluminous literature, see Murphy (2002). Surveying this literature, one is struck by the fact that the vast bulk of the research has been in the supervised learning tradition, employing procedures that in the psychological literature are commonly referred to as 'category formation' experiments. The term may actually be something of a misnomer, since the categories in question have already been formed, namely, by the experimenter; the subject's task would therefore be more accurately described as one of problem solving rather than category formation (Fodor 1980). The subject, that is, has to work out the criteria by which certain stimuli have been put into a certain category, whereas other stimuli have not. Much of this research has been conducted on the example of visually presented stimuli; in comparison, the categorization of (non-linguistic) auditory stimuli has been neglected (but see Lotto, 2002). There is, however, a modest tradition of concept formation experiments conducted on the example of phonological categories (Jaeger, 1980, 1986; Jaeger & Ohala, 1984; Mompeán, 2002; Weitzman, 1992).

As mentioned, surprisingly little research has been conducted by cognitive psychologists on unsupervised learning, or 'category construction', as Murphy (2002: 126) calls it, in contradistinction to 'category formation'. The little research that Murphy reports suggests that the categories that subjects spontaneously construct in such experiments are quite different from the categories that they normally operate with. There is a tendency, namely, for subjects to seize on a single dimension of the stimuli, such as their size, or colour, and to group them accordingly

(Murphy, 2002: 128). The complex, multi-dimensional, and probabilistic categories enshrined in the lexicons of human languages rarely emerge.

Further perspectives on supervised and unsupervised learning are provided by the computational modelling of learning, especially in artificial neural networks (McLeod *et al*., 1998). (It is, in fact, from the computational literature that I have taken the terms 'supervised' and 'unsupervised'). Consider a typical connectionist set-up. An array of input nodes is linked, possibly via one or more sets of hidden nodes, with an output array. Initially, the nodes are connected by randomly assigned connection weights. An input is presented, and the system's output is compared with the 'desired' output. The connection weights are then adjusted so as to decrease the system's error. The cycle is repeated – typically, many thousands of times – with each input being matched with a desired output. Eventually, the connection weights stabilize, and the system may be able to give the 'correct' output even for new inputs which it has never before encountered. One of the earliest and best-known applications of this procedure to language learning is Rumelhart & McClelland's (1986) account of the training of a network to produce past tense forms of English verbs (for an update, see Plunkett, 1995). The procedure, it will be appreciated, rather closely models the psychologists' category formation experiments. Thus in the psychologists' experiments, we might suppose that at first subjects cannot make head or tail of the array of stimuli that they are presented with, and, like the neural network, give random responses. After repeated trials, in which feedback is provided, they increasingly come up with the 'correct' classification of the stimuli.

Unsupervised learning in artificial neural networks involves the automatic recognition of patterns and regularities in the input. Several aspects of linguistic structure have been subjected to this kind of procedure. Thus, Goldsmith (2001) proposes a heuristic for automatic morpheme segmentation, while other aspects of linguistic analysis are addressed in Broeder & Murre (2000). For example, for Gillis, Daelmans, and Durieux (2000), the issue is the learnability of word stress rules on the basis of syllable structure and segmental features, while for Shillcock et al. (2000) the problem is to identify words from a phonemic transcription of connected speech (from which, of course, the word spaces had been removed). A common technique in unsupervised learning involves the use of clustering algorithms (Manning & Schütze, 1999). Each stimulus is defined as

a point in multi-dimensional space, and inputs cluster according to their relative closeness, with each new stimuli being 'categorized' in terms of the cluster it gets associated with. One of the best-known unsupervised procedures is the self-organizing maps of Kohonen (1982); a more sophisticated model has been developed by Kasabov (2002). Employing Kasabov's ECOS (= 'Evolving Connectionist Systems') model, Morales & Taylor (2005) found that the unsupervised learning of small vocabularies, the sole input to the system being digitized 'signatures' of different pronunciations of the words, turned out to be remarkably robust at the testing phrase, that is, in correctly classifying new pronunciations of the words.

Under what circumstances might unsupervised learning take place in human subjects? One condition would be that the stimuli naturally cluster into so many categories. It might be the case, for example, that different sets of features co-occur in distinct sets of stimuli, or that a continuously varying feature has frequency-of-occurrence values that are bi-modally distributed over the stimuli. In such cases, the categories might be said to be 'in the world', in that the relevant categories can be identified in terms of feature correlations or feature maxima.

It goes without saying that the learner has to be able to perceive the features in question. Consider, in addition, the possibility the learner may be innately predisposed to respond to certain features, or to certain dimensions of the stimuli. In this case, the emerging categories would be a function of the system's perceptual mechanism, rather than feature correlation in the world. We can illustrate the issues on the categorization of colour. On the one hand, it could be argued that the colour solid represents a three-dimensional array of all possible colours (the three dimensions being hue, saturation, and brightness), with no natural boundaries or lines of segmentation. The colour solid does not naturally divide into so many categories. This aspect must be counterbalanced by the fact that not all the possible colours occur equally frequently in the environment. Regions in the colour space which dominate in the environment might therefore be good candidates for emergent categories. Research into the linguistic encoding of colour, however, has shown that different languages around the world tend to select their colour categories from a universal set of focal colours (Berlin & Kay, 1969). The focal colours are those which the human visual system is specifically attuned to respond to, such as red and green, blue and yellow in the first instance, and admixtures of these, such as orange, pink, and so on. The

focal colours are the ones that tend to be lexicalized first, in spite of the fact that they may occur relatively infrequently in nature (Taylor, 2003b)

In light of the above remarks, let us now return to the learning of phonological categories. One of the most intensively studied features of the acoustic-phonetic signal is the role of Voice Onset Time (VOT) in the differentiation of different kinds of stops, such as voiced vs. voiceless, unaspirated vs. aspirated (Liberman *et al*., 1958; Lisker, 1978).[12] It has also been established that prelinguistic infants are highly sensitive to differences in the VOT continuum (Eimas *et al*., 1971). Some scholars, including Eimas, have suggested that this fact alone may be sufficient to trigger the formation of the respective categories; there would, therefore, be grounds to claim that the categorization of stop consonants is driven by innate properties of the human perceptual mechanism. Complicating the situation, however, is the fact that different languages exploit the VOT dimension in different ways. To the extent that VOT defines language-specific categories, these categories presumably have to be learned from experience. But even within a single language, it may be inappropriate to refer to *the* VOT values which differentiate the different categories of stops. VOT depends on many variables, such as the place of articulation of the stop (VOT values for bilabials are, on the whole, shorter than for velars; Lisker & Abramson, 1964), the prosodic properties of the syllable, i.e. whether stressed and foot-initial, or unstressed, the overall speech rate, and whether in utterance-initial position, and so on. These variations are subtle and numerous, and native proficiency in a language requires that they be learned (Pierrehumbert, 2003).

Leaving aside these various sources of variation, let us consider the simplified case of stops in syllable- and foot- initial position, that is, in the onset position of stressed syllables, the kind of sounds, namely, that have been so intensively studied in the experimental literature over the past decades. Imagine two hypothetical languages, in which foot-initial VOT values between, say, -50 and +50ms, occur with more or less equal frequency. One language places the boundary between voiced and voiceless stops around +5 ms, the other places the boundary between voiceless unaspirated and voiceless aspirated stops around +25 ms. It will be apparent that the unsupervised learning of the respective categories will be all but impossible. The learner would need the information that in the one language, VOT-values of +10 and +40 count as 'the same',

in the other language they count as 'different'. Unsupervised learning, however, *would* be feasible, if VOT values were bimodally distributed, clustering, for example, around +5 and +30. Frequency distribution of the stimuli would therefore naturally divide the stimuli into two categories. As it happens, VOT values in natural languages (in a given prosodic position) do indeed tend to be distributed in this way (Lisker & Abramson, 1964).

The possibility that distributional properties of the input might drive category formation was investigated by Kornai (1998) with respect to the vowel formant data reported in Peterson & Barney (1952). Peterson & Barney took measurements of the first three formants of 10 American English vowels each of which was spoken twice by 76 talkers (men, women, and children). A first glance at a graph plotting the formant data for all the vowel tokens gives the impression of a broad swath of values, with few natural boundaries. Even so, as Kornai observes, the formant data present a picture very different from a random set of dots in 2- or 3-dimensional space. Kornai reports, in fact, that automatic clustering procedures were able to assign the formant values to 10 categories, whose central values corresponded rather closely with those of the 10 intended vowels.[13]

Could first language learners exploit distributional facts in the input to bootstrap the learning of phoneme categories? There is some evidence to suggest that they could. With respect to their sensitivity to statistical properties of the input, it has been demonstrated that prelinguistic children, when presented with strings of nonsense syllables, are able to utilize statistical information in order to identify recurring patterns of syllables as 'words' (Saffran *et al*., 1996). Support also comes from Maye & Gerken (2000) and Maye et al. (2002), who exposed learners to a range of stop-vowel stimuli which, in one condition, were unimodally distributed in terms of their frequency of presentation, and, in the other condition, bimodally distributed.[14] When tested, learners in the latter condition (infants as well as adults) responded in a way suggesting that they had constructed two categories, whereas learners on the first condition did not. As Maye & Gerken (2000: 530) remark, it is as if listeners "maintain some sort of mental histogram", tracking the frequency of occurrence of acoustic patterns they had encountered. Anderson et al. (2003) make a similar point, hypothesizing that the sequence in which phonological categories are acquired is driven by input frequencies.[15]

It is therefore entirely plausible that the phonetic categories distinctive of a particular language (such as the aspirated vs. unaspirated stops, or the various vowel categories) could be 'seeded' during the first years of life by the statistical properties of the input. Further exposure to the language will, of course, be needed in order to sharpen and refine these categories (Bohn 2000). There is evidence that this process may continue until well into the school years (Hazan & Barrett, 2000). Some additional aspects of this process are mentioned below.

(i) Although a single dimension (such as VOT for the stop consonants, or formant frequencies for the vowels) may be sufficient to seed the respective categories, further exposure may enrich the category representations through the accretion of correlated properties. While VOT has been shown to be reliable cue for different kinds of stop consonants, VOT is not the only dimension differentiating the syllable-onset stops in English (Lisker, 1978). The intensity and spectral properties of the burst, the rate of change of formant transitions, and even the pitch of the ensuing vowel tend to correlate with the aspirated/unaspirated distinction, thus providing additional, though possibly redundant, cues for the characterization and differentiation of the respective categories. For vowels, an additional differentiating aspect is variations in duration (Peterson & Lehiste, 1960), and even differences in inherent pitch. Thus, all other things being equal, the duration of the vowel in *sad* [sæd] is likely to be greater than the duration of the vowel in *said* [sɛd].[16]

(ii) As acquisition progresses, the categories will become subject to internal organization. Members of the same category will come to be perceived as increasingly similar, while perceptual differences between neighbouring categories are increased. Kuhl (1991) in this context speaks of the 'perceptual magnet effect' – outlying members of a category tend to be 'drawn in' towards its prototypical centre. Thus, speakers become increasingly desensitized to differences between stimuli belonging to the same category, but readily discriminate stimuli which lie just on either side of a category boundary. These constitute the well-studied phenomenon of categorical perception, defined, by Harnad (2003) as a situation where "perceived within-category

differences are compressed and/or between-category differences are separated, relative to some baseline of comparison".

(iii) The increasing size of the learner's lexicon may also be a factor in phonological development (Beckman & Edwards, 2000). One might suppose that the ability to discriminate categories of sounds will entail that learners integrate these categories into their mental representations of words. Pater et al. (2004), however, report that infants who are able to discriminate *pin* and *bin*, *bin* and *din*, etc., were unable to associate these syllables with meaning differences in a word learning experiment. They explain this seemingly paradoxical finding in terms of the additional processing demands of word learning, involving the association of the acoustic stimuli with referential meaning. In early stages of language acquisition, therefore, words may well be represented in terms of their gross acoustic properties. As word learning gets under way, and the child's lexicon increases in size, more accurate lexical storage will be necessary. This will not only strengthen the mental representation of the phonological categories, it will also reinforce their differentiating potential.

(iv) A further factor in the acquisition of phonological categories is the various 'knowledge effects' that come into play. I address this issue in the next section.

V. KNOWLEDGE EFFECTS

I have given a tentative account of how the phonetic shapes of words, such as *coat* and *goat*, *ship* and *sheep*, might plausibly be learned in an unsupervised learning situation. This account, however, does not equate to the learning of phonemes, as these are traditionally understood. What our hypothetical learner will have acquired are allophones, or "phonetic equivalence categories" (Maye & Gerken, 2000: 532), that is, categories which comprise sounds which occur in particular phonological positions. The key characteristic of phonemes is "equivalence across contexts" (Pierrehumbert, 2003: 118). The phoneme, namely, is the level of representation at which *coat* and *goat*, *lack* and *lag*, *anchor* and *anger*, *bicker* and *bigger* differ with respect to the very same

contrast, namely, /k/ vs. /g/. The words contain the 'same' sounds, albeit in different syllable and prosodic positions. Unsupervised learning might result in the acquisition of the properties of each of the above words, yet not deliver the insight that *coat* and *goat* differ in the same way as *lack* and *lag*. On what basis, therefore, can we say that *coat*, *lack*, *anchor*, and *bicker* all contain the 'same sound', namely /k/?

The standard structuralist answer to this question was that different sounds belong to a single phoneme category because of their similarity and their interchangeability. Referring to Z. Harris's (1951: 20) statement that "[i]t is empirically discoverable that in all languages which have been described we can find some part of one utterance which will be similar to a part of some other utterance", Hoijer (1958, cited in Heitner, 2005: 20) comments:

> 'Similar' here means not physically identical but substitutable without obtaining a change in response from the native speakers who hear the utterance before and after the substitution: e.g., the last part of 'He's in' is substitutable for the last part of 'That's my pin' (Hoijer, 1958: 573).

Drawing on the structuralist tradition, Quine (1987: 150) gave the following account:

> Two distinguishable sounds belong to the same phoneme, for a given language, if switching them does not change the meaning of any expression in that language: such is the ordinary uncritical definition of the phoneme (Quine, 1987: 150).

Quine immediately modifies this in an attempt to exclude the controversial reference to 'meaning':

> But meaning is a frail reed; surely the phonemes, the very building blocks of the language, are firmer than that. They are indeed, despite occasional misgivings to the point. There is an easy behavioral criterion of sameness of phoneme that presupposes no general notion of sameness of meaning. Two sounds belong to the same phoneme if substitution of one for the other does not affect a speaker's disposition to assent to any sentence (pp. 150-151).

The claims made here are open to question on several counts. Consider, first, the issue of substitutability. Just as Pike (1947) queried whether linguists of his time really did pursue phonemic analysis without any reference to meaning, we can also ask ourselves whether anybody ever did perform the substitution tests. Nowadays, since the advent of digital signal processing, it is a relatively simple matter to cross-splice parts of recorded utterances. In earlier times, the experiment would have involved (literally) cutting up lengths of magnetic tape and sticking the bits together in a different sequence – a messy and time-consuming process at best, and prone to all kinds of errors and misjudgements. If such substitution experiments had been performed, the responses of native speakers might not at all have corroborated the phonemic analyses that the investigator was trying to validate. For example, if one cross-splices the initial /h/ sounds of *who* and *heat*, the resulting forms do not at all sound like *who* and *heat*, or even like English words at all. Or consider the initial and final consonants of a word like *tot*. If the final 't' were to be glottalized – a rather frequent pronunciation in many accents – interchanging the initial and final segments would, if anything, produce a word roughly transcribable as [ʔɒt^{h}], and heard as something like *ott*. Again, the two 't's cannot reasonably be said to be substitutable. When linguists, whether professional like Hoijer, or amateur like Quine, made statements about 'substitution' as the basis of phonemic analysis, we are dealing, I suspect, with armchair experimentation, intended to give a spurious air of scientific grounding to the enterprise.[17]

The claim that sounds are assigned to the same phoneme category on the basis of their phonetic similarity also does not hold up to scrutiny. As mentioned, the initial segments of *who* and *heat* – which, in terms of their articulation, are voiceless anticipations of the following vowels, phonetically [u̥] and [i̥] – do not sound at all similar, when excised from their context. (One would not, for example, want to claim that whispered versions of [u] and [i] are 'similar', and for this reason assign them to the same phoneme category). Likewise, there is little acoustic similarity between a glottal stop and an aspirated [t^{h}], in the above-mentioned pronunciation of *tot*. Conversely, the unstressed [ɪ] in *classify*, if voiceless (which it might well be), is essentially the same sound as the initial /h/ of *hit* (Pierrehumbert, 2003: 129), yet no one, presumably, would want to claim that the sounds are members of the same phoneme.

We need to look elsewhere for the source of intuitions about "equivalence across contexts". The place to look, I suggest, is the 'knowledge base' of categories (cf. Mompeán, 2004). In a seminal paper, Murphy & Medin (1985) posed the question why some groupings of objects are "informative, useful, and efficient", whereas others are "vague, absurd, or useless" (p. 289). They dismiss as simplistic the view that entities cohere in a category on the basis of their similarity; after all, some kind of similarity can be perceived in any grouping of objects. Rather, category coherence is a function of some 'underlying principle', or 'theory', which 'explains' why the entities should be grouped together, e.g. in terms of encyclopaedic knowledge of the domain, presumptions about causal connections, or the role of the entities within scripts and scenarios. Rather than entities being categorized on the basis of their similarity, it is the 'theory' relating the entities that makes them seem similar (p. 291). The intuitions of native speakers (and of linguists) that *who* and *heat* begin with similar-sounding segments would be the consequence of phonemic categorization, not its cause.

With regard to phonetic segments, an important piece of knowledge concerns how these sounds are made – which articulators are involved, manner of airflow, and so forth. Thus, for English, the initial and final segments of *tot* both involve alveolar closure with no accompanying vocal fold vibration, even though the acoustic effects of the articulation are very different for onset and coda consonants. Knowledge of the articulation could therefore support the grouping of the initial and final consonants into a single category. Jusczyk, in fact, has argued that a major impetus for the emergence of phoneme categories could well be the need for the learner to coordinate perception and production:[18]

> From the standpoint of word recognition, there is no need of an ability to detect the similarity in the initial portions of the words "big," "beet," "bop," and "bun." Nor is there any particular need for the speech perception system to extract any similarity between the way that the word "park" begins and the way that "tip" ends (although this ability is critical for learning to read English). However, in order to produce, and reproduce, any of these items correctly on another occasion, it may be helpful to take note of any similarities in the articulatory gestures that are required to produce these (Jusczyk, 1997: 205).

These remarks are relevant also for non-canonical articulations of stops. For example, stops in other than onset position of stressed syllables might not achieve full closure. *Upper* may be pronounced as [ˈʌɸə], *bigger* as [ˈbɪɤə]. Here, the gesture towards closure is made, but is not fully achieved. Knowledge of the articulations may thus support the intuition that [ɸ] and [ɤ] are members of the /p/ and /g/ categories, respectively. Or consider the fact that in many accents of English, a coda /t/ (under certain prosodic conditions) is typically glottalized, that is, the alveolar closure is made simultaneously with a glottal closure [ʔ͡t] If, furthermore, glottal closure should momentarily precede alveolar closure, there may well be no trace of the alveolar closure in the acoustic signal. The alveolar gesture may nevertheless be present, and could give rise to the intuition that the final [ʔ] is still a kind of /t/. Examples of this kind of 'hidden' articulation (that is, articulations with no auditory consequences) are documented in Browman & Goldstein (1992).

A second source of knowledge concerns dialectal and stylistic variants. There is no such thing as the perfectly homogeneous linguistic community of Chomsky's (1965: 1) idealization. Even leaving aside dialectal variation, each speaker commands a range of stylistic varieties, and comes into contact with many different speaking styles. Observing that *cat* is variably pronounced [kæt^{s}], [kæt^{h}], [kæt$^{\neg}$], [kæʔ͡t], and [kæʔ], the learner may come to group these different coda sounds as different kinds of /t/. Similarly, observing that in slow, careful pronunciation, *city* has a medial [t], whereas in rapid speech it has the flap [ɾ], the flap may again be assimilated to the /t/ category, in spite of its phonetic distinctiveness.

A third influence would be knowledge of the orthography. The flap in *city* might well be identical in articulation to the flap in *ready*. Knowledge of how these words are spelled, however, could cause the first to be categorized as a kind of 't', the latter as a kind of 'd'. Knowledge of morphological relations might also come into play. The perception of the flap in *madder* as being an example of /d/ rather than /t/ could be a consequence of the fact that the speaker knows that *madder* is derived from *mad*. Both these issues are extensively discussed in Mompeán (2004).

I have outlined some of the factors which might contribute to the emergence of the phoneme concept. The possibility still remains, however, that phoneme categories might be inventions of analyzing linguists, which play no role in the mental representations of linguistically naïve speakers. As Jaeger (1980: 233) put it, "even the most basic or self-evident" claims of theoretical linguistics need to be subjected to empirical investigation. A number of scholars, including Jaeger (1980) and Mompeán (2002: Experiment 3), have indeed attempted to demonstrate the psychological reality of the phoneme concept, with encouraging results. Thus, Jaeger found evidence that subjects classified various allophones and positional variants of /k/ into a single category, while Mompeán reported analogous findings for the allophones of /p/. It should be borne in mind, however, that both researchers employed a concept formation paradigm,[19] in the supervised learning tradition, as described above. It cannot, therefore, be ruled out that the experimental subjects were simply able to solve the categorization puzzle that they had been presented with, with no implications that the subjects had prior mental representations of categories, nor, even less, that the categories played a role in the subjects' day-to-day linguistic performance.[20] On the other hand, the fact that all 9 of Jaeger's subjects, and all 20 of Mompeán's, were able to form the categories to criterion,[21] would suggest that the subjects were indeed tapping into their mental representations of the respective categories, rather than constructing ad hoc categories in response to the experimental tasks.

A word of caution is necessary, however. English speakers who 'have' the relevant phoneme categories readily appreciate that *cat*, *tack*, and *act* contain the 'same' three sounds, arranged in different sequences. Indeed, the insight that words can be segmented into smaller units, and that these units recur in different words, would seem to be a prerequisite for mastery of an alphabetic writing system (Treiman & Baron, 1981), even though, as the example of *cat, tack*, and *act* shows, the correspondence between phonemes and letters is not always one-to-one. Continuing experience with an alphabetic writing system will only serve to strengthen and entrench the phoneme concept and its application to the words on the language. As Kornai (1998) observes, "to the extent that a … phoneme based writing system can easily be acquired and consistently used by any speaker of the language, the psychological reality of the units forming

the basis of the system becomes hard to deny". To be sure, the segmental phonemic analyses made by children learning the writing system may not always correspond with the analyses of professional linguists, nor even with the analyses that are enshrined in the writing system (Treiman, 1985). Neither is it the case that all children achieve the phonemic insight at the same time, and at the same rate. Even 'wrong' segmentation, though, still testifies to the implementation of some segmentation strategy, and could be taken as evidence that the phonemic insight is present. It may be, however, that some speakers *never* achieve the initial phonemic insight. Mattingly (1972) and Sampson (1985: 163) suggest that residual levels of illiteracy, even in societies with universal education, are due to the fact that a small minority of individuals fail to appreciate the phonemic structure of their language. While illiteracy, in a predominantly literate society, obviously impacts on a person's linguistic development in many ways (for example, by depriving them of exposure to literary styles and genres, and their associated syntactic and lexical properties), we probably should not conclude that the basic speaking and listening abilities of these individuals will be substantially impaired vis-à-vis that of their literate compatriots. Maye & Gerken (2000: 532) suggest that "phonetic equivalence categories … could plausibly be the *only* psychological correlates" (authors' emphasis) of the linguist's phonemes. In view of the experimental evidence cited above, as well as the fact of widespread literacy in alphabetic writing systems, this is probably an overly cautious view. On the other hand, knowledge of phonetic equivalence categories (i.e. positional allophones) could plausibly be *sufficient* for speaking and listening proficiency to be guaranteed.

VI. CONCLUDING REMARKS

As stated earlier in this paper, the phoneme concept is controversial. I have framed the above discussion around some of the controversies which were current during the heyday of Bloomfieldian structuralism, in the mid decades of the last century, concerning the criteria by which phonemes are to be identified. As is well-known, the advent of generative phonology, in the 1960's and 1970's, ushered in new controversies. Specifically, generative phonologists such as Postal (1968) and Chomsky & Halle (1968) queried the need for a phonemic level of

representation at all, proposing instead that a battery of ordered rules was able to generate the surface form of an utterance (roughly, the utterance in a narrow phonetic transcription) directly from a unique representation of each constituent morpheme, with no special theoretical status attaching to an intervening phonemic representation. Crucial to their argument for ignoring the phoneme was the fact that certain rules (e.g. of assimilation) sometime seemed to bypass the phonemic level altogether,[22] while others appeared to create surface contrasts (i.e., minimal pairs) which did not correspond with intuitions about a phonemic level.[23] Even so, as Schane (1971) observed, the output of one set of rules – the morphophonemic rules – did correspond, by and large, with what would earlier have been called a phonemic representation, while the phonetic rules corresponded, by and large, with what would have been regarded as phoneme realization rules.

The generative phonology approach, it will be observed, was strictly 'top-down', in the sense that details of surface pronunciations were derived from more abstract representations, rather than vice-versa. Generative phonology thus inverted the 'bottom-up' programme of the Bloomfieldians. From the perspective of the child acquiring an ambient language, a top-down approach can be viable only if one makes the gratuitous assumption that the abstract units are already available to the learner, namely, through genetic inheritance (Lindblom, 2000). This is a dubious proposition, if only because of the language-specificity of the more abstract categories (such as the phonemes). If we make – as I think we should – minimal assumptions concerning the learner's initial state, we are obliged to consider seriously the bottom-up perspective. This has been my aim in this paper.

Chomsky (1964) spoke disparagingly of the 'taxonomic phoneme'. Underlying the present account is, on the contrary, the view that phonemes are properly regarded as categories whose members are positional allophones; these latter in turn are also categories, whose members are encountered utterance events (cf. Nathan, 1986). Phonemes, therefore, take their place within a taxonomy of phonetic segments. The taxonomy need not, of course, stop at the phoneme. Phonemes might be grouped together in higher level, i.e. more 'schematic' categories, such as 'vowel' and 'consonant', with several intervening categories in between, such as 'obstruent', 'nasal', 'front vowel', 'short vowel', and so on. The very essence of the phoneme is, therefore, its

taxonomic status. If this view of the phoneme is accepted, then the extensive research on categorization and taxonomies that has been conducted by psychologists, and cognitive scientists more generally, becomes relevant to phonological theory. It becomes legitimate to enquire, for example, whether phoneme categories exhibit prototype effects, whether phonemes might be considered as 'basic level' categories within a taxonomic hierarchy, and what the distinctive properties of categories that are 'superordinate' and 'subordinate' to the basic level might be. These questions are touched on in Taylor (2002: to appear). A full discussion, however, must await a sequel to this paper.

NOTES

1. "Psychologically, setting aside its expression in words, our thought is simply a vague, shapeless mass. ... [W]ere it not for signs, we should be incapable of differentiating any two ideas in a clear and constant way. In itself, thought is like a swirling cloud, where no shape is intrinsically determinate. No ideas are established in advance, and nothing is distinct, before the introduction of linguistic structure." (Saussure/Harris, 1983: 155)

2. "The substance of sound is no more fixed or rigid than that of thought. It does not offer a ready-made mould, with shapes that thought must inevitably conform to. It is a malleable material which can be fashioned into separate parts in order to supply the signals which thought is in need of." (Saussure/Harris,1983: 155).

3. Strictly speaking, of course, the input to acquisition is not just auditory, but (in the case of sighted individuals) auditory-visual, in that the learner has access to visual information pertaining to the speaker's lip and jaw movements.

4. This statement leaves open the question of what constitutes a 'word' for purposes of perception, storage, and retrieval. The category comprises, in the first instance, word forms, such as *run, runs, running*, but also 'phonological words', such as *cuppa* [kʌpə] in *cup of tea*. Frequently occurring combinations, such as *all gone, bye-bye*, and *good-night*, might also have word-like status, at least for the young child. These issues, though important, are not strictly relevant to the point made in this paragraph.

5. Analysis and segmentation does not, of course, entail that words will cease to be stored as wholes. To claim this, would be to fall foul of the 'rule-list fallacy' (Langacker, 1987). It is plausible, indeed likely, that words continue to be stored as phonological wholes at the same time as their phonological constituents are recognized (cf. Lachs *et al*., 2000).

6. Cf. Sapir's (1921: 56) often-cited remark: "In watching my Nootka interpreter write his language, I often had the curious feeling that he was transcribing an ideal flow of phonetic elements which he heard, inadequately from a purely objective standpoint, as the intention of the actual rumble of speech".

7. A couple of representative statements: Hockett (1942: 20) asserted that "no grammatical fact of any kind is used in making phonological analysis", while Bloch (1948: 5) declared: "we shall avoid all semantic and psychological criteria. The implication is, of course, that such criteria play no part, or at least need not play one, in the theoretical foundation of phonemics. … The basic assumptions that underlie phonemics, we believe, can be stated without any

mention of mind and meaning". Bloch did, however, concede that in practice a linguist *would* appeal to meaning, but only as a "shortcut".

8. For the Bloomfieldian structuralists, with their fiercely anti-mentalist stance, it would have been unscientific to attribute psychological reality to their analyses. Nowadays, largely as a result of the 'cognitive turn' initiated by Chomsky, we have few qualms about crossing the boundary between the subject matter of linguistics and the cognitive states and processes of a language user. A linguistic description, namely, is taken as a hypothesis about a speaker's mental representations, and a linguist's analytical procedures may be regarded as analogous to those of the child language learner.

9. Quine (1960) posed the question of how a field linguist, on observing the native to utter *gavagai* on seeing a rabbit, would establish that *gavagai* means 'rabbit'. *Gavagai* could mean many things, only some of which could be subject to empirical disconfirmation.

10. See, however, Werker (2003) for a more recent view of the matter.

11. Daniel Jones insisted that "it is incumbent on us to distinguish between what phonemes *are* and what they *do*" (Jones, 1973 [1957]: 28). Thus, the possibility of lexical contrasts (i.e. minimal pairs) should be seen as a consequence of the existence of phonemic categories, not their defining, or causal feature: "An important point to notice is that the phoneme is essentially a phonetic conception. The fact that certain sounds are used in a language for distinguishing the meanings of words doesn't enter into the definition of a phoneme. It would indeed be possible to group the sounds of a language into phonemes without knowing the meaning of any words" (Jones, 1929, quoted in Bloch,1948: 6). For a critique of the view that phonemes are inherently contrastive entities, see Berg (1993).

12. VOT is the duration, usually measured in milliseconds, between the release of a stop closure and the onset of voicing, typically diagnosed by the presence of periodicity in the wave form. A positive VOT value, e.g. +50, indicates that voicing sets in after the release; a negative value, e.g. -50, indicates that voicing sets in before the release.

13. The clustering experiment reported by Kornai did, however, specify the number of target clusters as 10. This would correspond to a situation in which a learner is informed about the number of vowel categories in a language, and is left to work out to which of the categories individual tokens are to be assigned.

14. The experiments were conducted with English speakers, and concerned the contrast between initial voiced stops (as in *die*) and unaspirated voiceless stops, such as occur, after an initial 's', in *sty*. The contrast is alien to the phonological system of English.

15. A reviewer takes issue with the notion of the statistical learning of phonetic categories, pointing out that "speakers are not tape recorders... they don't just record sound images and compare them." However, the results obtained by Maye and Gerken (2002) are very strong evidence that listeners do indeed record and compare even minute phonetic details of heard utterances; without some such mechanism, it is difficult to imagine how their results could be explained at all. Listeners' attention to, and retention of, fine acoustic-phonetic detail is also supported by research by Goldinger (1996) and by Lachs, McMichael and Pisoni (2000). Circumstantial evidence is the fact, noted by Pierrehumbert (2003, 120), that the properties of phonetic categories are in the main language-specific; consequently, these properties "must be learned by native speakers, because they have consequences for category boundaries in perception and because they must be accurately reproduced to achieve a native accent in production." It may be relevant, also, to recall that for Slobin (1985), one of the "operating principles" enabling language acquisition to take place, was: "Keep track of the frequency of occurrence of every unit and pattern that you store". The role of the frequency of occurrence in language acquisition has been reviewed by Ellis (2002).

16. Bohn and Flege (1990; 1992) report that second language learners of English – for whom the [æ]–[ɛ] and [i]–[ɪ] contrasts are notoriously difficult – may rely predominantly on durational differences, unlike native English speakers, who rely predominantly on spectral differences.

17. A more generous interpretation of the substitution test would be that the analyzing linguist would replace one phonetic symbol in a transcription by another and then try to articulate the result. The feasibility of this enterprise presupposes that phonetic symbols accurately represent the acoustic properties of the speech signal, which, in the case of the stop consonants, is at best questionable.

18. Barring pathological cases, speakers are also listeners, and vice versa. Thus, a speaker is continually presented with the auditory consequences of her own articulations. Without needing to subscribe to the now largely discredited motor theory, with its claim that speech sounds are perceived in terms of the articulations that produced them, we can suppose (as a reviewer has suggested) that listeners will be inclined to intuit the articulatory intentions of a speaker. This suggestion links up with a wider theme in the acquisition literature, namely, the view that language acquisition may be driven by the learner's ability to read the intentions of an interlocutor (Taylor, 2002: 67-8; Tomasello, 1999). The matter has been investigated mainly from the point of view of the learning of word meanings, rather than with respect to the learning of phonetic categories.

19 This, at least, is true of Jaeger's Experiment 2. Experiment 1 used a classical conditioning paradigm, in which the results from 10 out of 16 subjects had to be discarded.

20. Imagine, for example, a concept formation experiment, in which the concept to be acquired is defined by the features [two-syllable word] and [beginning with either /l/ or /s/]. The fact that some subjects might be able to form this category to criterion would not entitle us to infer that the category plays any role whatsoever in the subjects' mental representation of their language.

21. In contrast, 6 of Mompeán's 20 subjects (2002: Experiment 1) failed to form the category '(word-initial) consonant'. This finding could be interpreted to mean that the superordinate category 'consonant' is less available to consciousness than a basic level phoneme category such as /p/.

22. For example, nasal consonants typically assimilate to the place of articulation of a following obstruent. In words such as *link* [lɪŋk], *imp* [ɪmp], and *sent* [sɛnt], the assimilated nasals would be assigned, unproblematically, to the phonemes /ŋ/, /m/, and /n/, respectively. But in the case of *comfort* [kʌmfət] and *camphor* [kæmfə], it is by no means obvious to which phoneme the assimilated [m] should be assigned. In the first set of examples, assimilation determines the occurrence of different phonemes, in the second set, assimilation results in a sound whose phonemic status is uncertain. The two sets can be unified by assuming an underlying nasal segment, which receives its place feature through assimilation, thereby removing the need for a distinctive phonemic level of representation.

23 An example of this kind of spurious minimal pair is the contrast, in some dialects, between *cat* [kæt] and *can't* [kæ̃t].

REFERENCES

Anderson, J.. Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech*, 46, 155-182.

Aslin, R. N., Pisoni, D. B., Hennessy, B. L., & Perey, A. J. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development*, 52, 1135-1145.

Aslin, R. N., Jusczyk, P. W., & Pisoni, D. B. (1988). Speech and auditory processing during infancy: Constraints on and precursors to language. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology: Cognition, perception, and aanguage*, vol. 2. New York: Wiley, pp. 147-254.

Beckman, M. & J. Edwards. (2000). The ontogeny of phonological categories and the primacy of lexical learning in linguistic development. *Child Development*, 71, 240-249.

Beckman, M. & J. Pierrehumbert. (2000). Positions, probabilities, and levels of categorisation. Keynote address, *Eighth Australian International Conference on Speech Science and Technology*, Canberra, Dec. 4-7, 2000. Available at http://babel.ling.northwestern.edu/~jbp/SST2000.pdf

Berg, Th. (1993). The phoneme through a psycholinguist's looking-glass. *Theoretical Linguistic*s, 19, 39-76.

Berlin, B. & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley: University of California Press.

Bloch, B. (1946). A set of postulates for phonemic analysis. *Language*, 24, 3-46.

Bloomfield, J. (1933). *Language.* London: George Allen & Unwin.

Bohn, O.-S. (2000). Linguistic relativity in speech perception: An overview of the influence of language experience on the perception of speech sounds from infancy to adulthood. In S. Niemeier & R. Dirven (Eds.), *Evidence for linguistic relativity*. Amsterdam: J. Benjamins, pp. 1-28.

Bohn, O.-S. & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics*, 11, 303-328.

Bohn, O.-S., & Flege, J. E. (1992). The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition*, 14, 131-158.

Borges, J. L. (1964). Funes, the memorious. In D. A. Yates and J. E. Irby (Eds.), *Labyrinths. Selected stories & other writings*. New York: New Directions, pp. 59-66.

Bowerman, M. (1988). The "no negative evidence" problem: How do children avoid constructing an overgeneral grammar? In J. A. Hawkins (Ed.), *Explaining language universals*. Oxford: Blackwell, pp. 73-101.

Broeder, P. & Murre, J. (Eds.) (2000). *Models of language acquisition: Inductive and deductive approaches.* Oxford: Oxford University Press.

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

Brown, R. (1958). *Words and things*. Glencoe, Ill.: Free Press.

Caselli, M. C., Bates, E., Casadio, P., Fenson, J., Fenson, L., Sanderl, L., & Weir, J. (1995). A cross-linguistic study of early lexical development. *Cognitive Development*, 10, 159-199.

Chomsky, N. (1964). Current issues in linguistic theory. In J. A. Fodor & J. J. Katz (Eds.), *The structure of language*. Englewood Cliffs, NJ: Prentice-Hall, pp. 50-118.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Mass.: MIT Press.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Clark, E. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), *Mechanisms of language acquisition*. Hillsdale, NJ: Erlbaum, pp. 1-33.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, 171, 303-306.

Ellis, N. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143-188.

Fodor, J. A. (1980). The present status of the innateness controversy. In J. A. Fodor (Ed.), *Representations: Essays on the foundations of cognitive science*. Cambridge, MA: MIT Press, pp. 257-316.

Gillis, S., Daelemans, W., & Durieux, G. (2000). 'Lazy learning': Natural and machine learning of word stress. In P. Broeder & J. Murre (2000), *Models of language acquisition: Inductive and deductive approaches.* Oxford: Oxford University Press, pp. 6-99.

Gleitman, L. (1990). The structural sources of verb meaning. *Language Acquisition*, 1, 3-55

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27, 153-198.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1166-1183.

Harnad, S. (2003). Categorical perception. In *Encyclopedia of cognitive science*. London: Nature Publishing Group/Macmillan. At http://www.ecs.soton.ac.uk/~harnad/Temp/catperc.html

Harris, R. (1973). *Synonymy and linguistic analysis*. Oxford: Blackwell.

Harris, Z. (1951). *Methods in structural linguistics*. Chicago: University of Chicago Press.

Hazan, V. & Barrett, S. (2000). The development of phonemic categorization in children aged 6-12. *Journal of Phonetic*s, 28, 377-396.

Heitner, R. M. (2005). An odd couple: Chomsky and Quine on reducing the phoneme. *Language Sciences*, 27, 1-30.

Hockett, C. (1942). A system of descriptive phonology. *Language*, 18, 3-21.

Hoijer, H. (1958). Native reaction as a criterion in linguistic analysis. *Proceedings of the Eighth International Congress of Linguistics*, 573–591.

Hombert, J.-M. (1978). Consonant types, vowel quality, and tone. In V. Fromkin (Ed.), *Tone: A linguistic survey*. Orlando: Academic Press, pp. 77-111.

Jaeger, J. J. (1980). Testing the psychological reality of phonemes. *Language and Speech*, 23, 233-253.

Jaeger, J. J. (1986). Concept formation as a tool for linguistic research. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology*. Orlando: Academic Press, pp. 211-238.

Jaeger, J. J. & Ohala, J. J. (1984). On the structure of phonetic categories. *Berkeley Linguistics Society*, 10, 15-26.

Jones, D. (1929). Definition of a phoneme. *Le Maître Phonétique*, 43-44.

Jones, D. (1973). The history and meaning of the term "phoneme". In E. C. Fudge (Ed.), *Phonology*. Harmondsworth: Penguin, pp. 17-34.. First published 1957 in supplement to *Le maître phonétique.*

Joos, M. (Ed.). 1957. *Readings in linguistics*. Chicago: University of Chicago Press.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Kasabov, N. (2002). *Evolving connectionist systems: Methods and applications in bioinformatics, brain study and intelligent machines.* London, New York & Heidelberg: Springer.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43, 59-69.

Kornai, A. (1998). Analytic models in phonology. In J. Durand & B. Laks (Eds.), *The organization of phonology: Constraints, levels and representations*. Oxford: OUP. pp. 395-418. Also at http://people.mokk.bme.hu/~kornai/Papers/roy1.pdf

Krámský, J. (1957). *The phoneme: Introduction to the history and theories of a concept*. Munich: Wilhelm Fink.

Kuhl, P. (1987). Perception of speech and sound in early infancy. In P. Salapatek & L. Cohen (eds.), *Handbook of infant perception*, Vol. 2. New York: Academic Press, pp. 273-382.

Kuhl, P. (1991). Human adults and human children show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics*, 50, 93-107.

Lachs, L., McMichael, K., & Pisoni, D. B. (2000). Speech perception and implicit memory: Evidence for detailed episodic encoding of phonetic events. *Progress Report 24. Speech Research Laboratory*, Indiana University, Bloomington, pp. 149-167.

Langacker, R. W. (1987). *Foundations of cognitive grammar,* Vol. 1*: Theoretical prerequisites*. Stanford: Stanford University Press.

Lehiste, I. & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America*, 31, 428-435.

Liberman, A. M., Cooper, F. S, Shankweiler, D. P. & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, 74, 431-461.

Liberman, A. M., Delattre, P. D., & Cooper, F. S. (1958). Some cues for the distinction between voiced and unvoiced stops in initial position. *Language and Speech*, 1, 153-167.

Liberman, A. M. & Mattingly, I. G. (1985). The motor theory of speech perception revisited. *Cognition*, 21, 1-36.

Lindblom, B. (2000). Developmental origins of adult phonology: The interplay between phonetic emergents and the evolutionary adaptations of sound patterns. *Phonetica*, 57, 297-314.

Lisker, L. (1978). In qualified defense of VOT. *Language and Speech*, 21, 373-383.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops. *Word*, 20, 384-422.

Lotto, A. J. (2000). Language acquisition as complex category formation. *Phonetica*, 57, 189-196.

Luria, A. R. (1968). *The mind of a mnemonist*. Cambridge, MA: Harvard University Press

McLeod, P. Plunkett, K. & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes.* Oxford: Oxford University Press.

Manning, C. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Mattingly, I. (1972). Reading, the linguistic process, and linguistic awareness. In J. Kavanagh & I. Mattingly (Eds.), *Language by ear and by eye: Tthe relationship between speech and reading*. Cambridge, MA: MIT Press, pp. 133-147.

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. *Proceedings of the 24th Boston University Conference on Language Development*, 522-533

Maye, J., Werker, J. F. & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82, B101-B111.

Mompeán, J. A. (2002). *The categorisation of the sounds of English: Experimental evidence in phonology*. Unpublished Ph.D. dissertation, University of Murcia.

Mompeán, J. A. (2004). Category overlap and neutralization: The importance of speakers' classifications in phonology. *Cognitive Linguistics*, 15, 429-469

Morales, F. & Taylor, J. 2005. Learning and relative frequency. Unpublished manuscript, University of Otago.

Murphy. G. (2002). *The big book of concepts*. Cambridge, MA: MIT Press.

Murphy, G. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289-316.

Nathan, G. (1986). Phonemes as mental categories. *Berkeley Linguistics Society*, 12, 212-223.

Pater, J., Stager, C., & Werker, J. (2004). The perceptual acquisition of phonological contrasts. *Language*, 80, 384-402.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24, 175-184.

Peterson, G. & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, 32, 693-703.

Pierrehumbert, J. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speec*h, 46, 115-154

Pike, K. (1947). Grammatical prerequisites to phonemic analysis. *Word*, 3, 155-172.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Plunkett, K. (1995). Connectionist approaches to language acquisition. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language*. Oxford: Blackwell, pp. 36-72.

Postal, P. (1968). *Aspects of phonological theory*. New York: Harper & Row.

Potter, R. K., Kopp, G. A. & Green, H. C. (1947). *Visible speech*. New York: Van Nostrand.

Quine, W. v. O. (1960). *Word and object*. Cambridge, MA: MIT Press.

Quine, W. v. O. (1987). *Quiddities: An intermittently philosophical dictionary*. Cambridge, MA: Harvard University Press.

Rumelhart, D. & McClelland, J. (1986). On learning the past tenses of English verbs: Implicit rules or parallel distributed processing? In J. McClelland, D. Rumelhart, and the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 2. Cambridge, MA: MIT Press, pp. 216-271

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926-1928.

Sampson, G. (1985). *Writing systems: A linguistic introduction*. Stanford, CA: Stanford University Press.

Sapir, E. (1921). *Language: An introduction to the study of speech*. New York : Harcourt, Brace & World.

Saussure, F. de (1964). *Cours de linguistique générale*. Paris: Payot. First published 1916. Translated by R. Harris, as *Course in General Linguistics*. London: Duckworth (1983)

Schane, S. (1971). The phoneme revisited. *Language*, 47, 503-521.

Shillcock, R., Cairns, P., Chater, N., & Levy, J. (2000). Statistical and connectionist modeling of the development of speech segmentation. In P. Broeder & J. Murre (Eds.), *Models of language acquisition: Inductive and deductive approaches.* Oxford: Oxford University Press, pp. 103-120.

Slobin D. I. (1985). Cross-linguistic evidence for the language-making capacity. In D. I. Slobin (Ed.), *The cross-linguistic study of language acquisition.* Vol 2*: Theoretical issues*. Hillsdale, NJ: Erlbaum, pp. 1157-1249.

Taylor, J. R. (2002). *Cognitive grammar*. Oxford: Oxford University Press.

Taylor, J. R. (2003a). *Linguistic categorization*. Oxford: Oxford University Press. First edition: 1989.

Taylor, J. R. (2003b). Near synonyms as co-extensive categories: 'high' and 'tall' revisited. *Language Sciences*, 25, 263-284.

Taylor, J. R. (to appear). Prototypes in cognitive linguistics. In P. Robinson & N. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition*. Mahwah NJ: Lawrence Erlbaum.

Tomasello, M. (1999). *The cultural origins of human cognition*. Cambridge MA: Harvard University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Treiman, R. (1985). Phonemic awareness and spelling: Children's judgments do not always agree with adults. *Journal of Experimental Child Psychology*, 39, 182-201.

Treiman, R. & Baron, J. (1981). Segmental analysis ability: Development and relation to reading ability. In G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice*, Vol. 3. New York: Academic Press, pp. 159-198.

Vihman, M. M. (1996). *Phonological development: The origins of language in the child*. Oxford: Blackwell.

Weitzman, R. (1992). Vowel categorization and the critical band. *Language and Speech*, 35, 115-125.

Werker, J. (2003). The acquisition of language specific phonetic categories in infancy. *Proceedings of the 15th International Congress of Phonetic Sciences*, 21-26.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7, 49-63.

