The units of speech – A bio-cultural perspective

Björn Lindblom

Department of Linguistics, Stockholm University

lindblom@ling.su.se

Abstract

Humans have large vocabularies and learn new words all of their lives. From infancy to adolescence, children acquire lexical entries at a remarkably fast rate. These accomplishments are linked to the uniquely human method of coding information: the combinatorial use of discrete entities such as the phoneme. This method creates the conditions for open-ended lexical systems thereby contributing to the singular expressive power of human languages. After reviewing a broad range of experimental evidence on speech and non-speech we conclude that a strong case can be made for the claim that the hallmark of true phonology - the combinatorial structure of sound patterns – has behavioural origins.

Keywords: vocabulary learning; world's sound systems; perceptual contrast; size principle; syllable; phonotactic patterns; coarticulation; discrete units; positional control; motor equivalence; kinematic constancies; cognitive growth; vocal imitation; recombination; origin of phonemes.

Goals

Human language is a product of biology and culture. This truism is readily acknowledged in contemporary linguistics and forms an informal part of the ongoing academic narrative about language. However, when it comes to the more ambitious task of proposing formalized and explanatory theories of language, the bio-cultural perspective is curiously bypassed, even explicitly down-played. A case in point is the influential agenda persuasively promoted by Chomsky and embraced by many leading linguists. Anderson's "Why phonology isn't 'natural'" is a well-known attempt to spell out what this program entails for the study of sound patterns (1981).

There are several reasons for the lack of interest in language as behaviour in linguistics. An important one is the fact that the field continues to be under the sway of 'structuralism' which has been, and still is, firmly anchored in the belief that *'the classical Saussurean assumption of the logical priority of the study of langue (and the generative grammars that describe it) seems quite inescapable"* (Chomsky 1964:52).

Undeniably, this school of thought has been a huge academic success sociologically thereby significantly raising the scientific image of the discipline.

The focus of the present contribution is on the units of speech such as syllables and phonemes. The limitations of the structuralist approach become abundantly clear when we ask: "Where do these units come from?" In the prevailing paradigm that is an unasked question. Nonetheless, with respect to the phoneme, there is a possible answer: "It comes from phonemic analysis". In other words, it comes from observed linguistic facts filtered through an operational procedure aimed at identifying phonetic differences that change meanings. The output of the method is an abstract unit stripped of its intrinsic

phonetic content and specified by a set of attributes or 'features' (as derived from the phonetic differences).

The shortcomings of the ruling paradigm are also exposed in the context of applications. Knowledge about language is needed in many areas facing complex educational, clinical and technological problems. Those problems are more likely to be linked to aspects of language use (language learning and teaching, speech and language pathologies, man-machine interaction etc.) than to language as an 'autonomous' (performance-independent) object.

The goal of the present paper is to demonstrate the feasibility of a non-structuralist approach to the basic units of speech. Accordingly we reverse structuralist priorities starting from substance (performance-based factors) in a quest for possible behavioural causal factors underlying sound structure formation.

Some facts in need of explanation

Human vocabularies are large and open-ended

Measurements of vocabulary size provide an important datum about the language learning process.

Figure 1 presents findings from studies of American English and Swedish children (Miller 1977, 1991 and Edlund 1957). Straight lines were used to connect the points obtained by pooling the two data sets. The resulting pattern is s-shaped suggesting that learning has its most rapid course between 8 and 12 and then slows down and reaches 60,000 root words at 17 years - a conservative estimate for an average America high school graduate (Miller 1991).

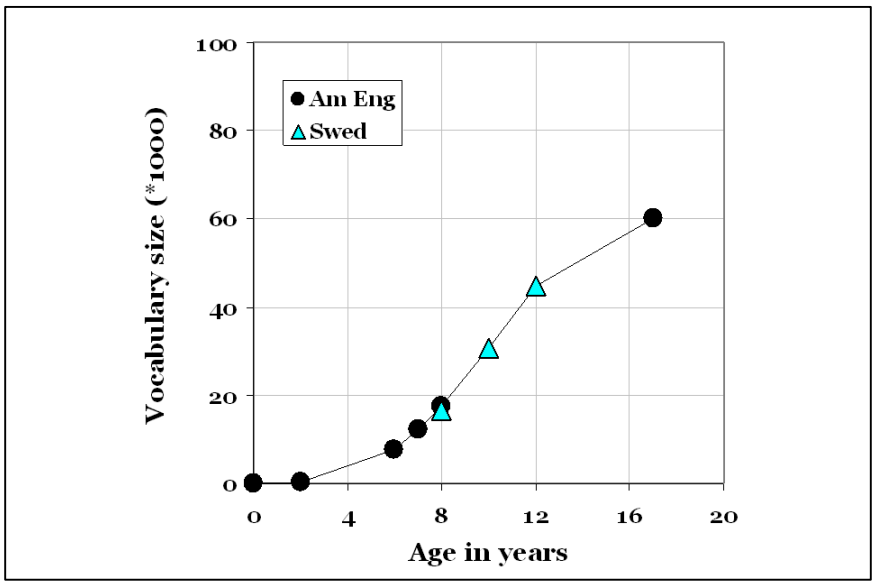

Figure 1 Average American English and Swedish data on vocabulary growth as a function of age.

This observation is in agreement with a well-known fact: Word learning continues throughout adult life. The learning curve would thus be expected to continue to rise above the 60,000 mark after the 17th birthday, albeit more slowly. 'Open' word classes, such as nouns, adjectives and verbs continue to receive new input, whereas 'closed' classes, grammatical morphemes, tend to remain fixed in size (Teleman 2012).

Vocabulary growth is open-ended. As word formation rules remain productive, there is no fixed upper limit.

Word learning is fast and unsupervised

To shed some light on numbers like those in Figure 1, Miller chose to describe his findings in terms of learning rate as measured in 'words per day'. We repeated this exercise for the data of Figure 1 by using adjacent observation points. For each pair we divided the increase in number of words by the number of days elapsed. For instance, from 8 to 10 years old there is an increase from 16630 to 30810 Swedish words. The average number of words learned per day then becomes (30810-16630)/((10-8)*365)= 19.4.

Table 1 Rate of word learning expressed in words per day.

age	words per day	language	source
10	19.4	Swedish	Edlund 1957
12	19.2	Swedish	Edlund 1957
age	**words per day**	**language**	**source**
6	5.1	Am English	Miller 1977
7	12.6	Am English	Miller 1977
8	14.2	Am English	Miller 1977
17	12.9	Am English	Miller 1991

In reflecting on these numbers we should bear in mind that they are averages. They are not to be taken literally. They do not imply that learning say ten words takes only one day. Words probably vary with respect to how long it takes to learn them. Initially there may be a large heterogeneous set in which individual items are at more or less advanced stages, some requiring more practice and exposure, others on the brink of complete mastery. Then an organizing principle emerges and, from then on, the words-per-day measure starts to return large average numbers.

Having said that, we still have reason to marvel at the speed at which this development unfolds.

The combinatorial nature of lexical packing

In short, vocabularies are large and open-ended. Normal children learn them fast and without explicit teaching. How do they do that?

Part of an answer must be linked to the fact that, unlike animal communication systems, the contents of human vocabularies is not a set of holistic Gestalt patterns but is built from a small number of discrete units (phonemes) that are combined to form bigger units such as syllables, morphemes and words. Combinatorial coding is an extremely efficient method for specifying large amounts of information and the organizing principle of lexical inventories is the recombination of phonemes.

Phonemic coding is one of the keys to the size of human vocabularies.

How do children find these basic units? The success of normal children on this task is a highly non-trivial achievement since the physical manifestations of the alleged units are never context-free: In the massively variable signals of natural speech they do not come neatly labelled (Perkell & Klatt 1986). This is all the more surprising when we contemplate that, not only do the units have to be discrete, they also need to be context-free to facilitate recombination effectively.

For the remainder of this chapter we shall try to pinpoint various behavioural skills and factors whose interaction may contribute to the unsupervised formation of large, open-ended and phonemic

vocabularies. In so far as they are successful such projects should provide a launching pad for future studies in developmental phonetics.

Some properties of the world's sound systems

Universals and design principles

An examination of the phonetic properties of the world's sound systems offers a useful perspective on the child's learning task. Standard sources like the UPSID database (Maddieson 1984) and the web-based World Atlas of Language Structures (Dryer & Haspelmath 2011) contain phonetic data from a large number of languages. Although these sources give only a sample of all languages spoken around the globe, the selections are made in an effort to make them representative with respect to both geographic dispersion and historical distance.

For the development-oriented phonetician the interest of typological comparisons derives from the possibility that the side-by-side investigation of a really remote and historically unrelated systems will reveal universal patterns and design principles that could provide clues as to the nature of the learning task and how children go about accomplishing that task.

The taxonomy of phonetic properties

Ladefoged & Maddieson's 1996 publication is so far no doubt the most comprehensive account of the articulatory and acoustic properties of the world's vowels and consonants. It summarizes lab research and field work on phonetic contrasts in over 400 languages. In keeping with the taxonomic tradition in phonetics, L & M describe the consonants in terms of their place and manner of articulation. To specify 'place of articulation', they distinguish 17 categories on the basis of two criteria (i) the region of the target place (from labial, dental, … to glottal) and (ii) the articulatory structure that is positioned (their Table 2.1). For example, for targets at the 'dental', 'alveolar' or 'post-alveolar' regions, it is possible to use the tip or the blade as the articulator. As a result their system technically allows for two alveolar 'places': one 'apical alveolar' (involving the tongue tip) and another one 'laminal alveolar' (involving the tongue blade).

Manners are divided up into stops, nasals, fricatives, laterals, 'r-sounds' and clicks. In the case of stops special attention is paid to the state of the glottis (voiced and voiceless), airstream mechanisms (ejectives & implosives) and the coordination of glottal and articulatory activity (pre- & post-aspiration). Nasals present variants linked to the relative timing of oral and velar movements (prenasalized stops, prestopped nasals). Fricatives occur along the entire place dimension and are especially numerous in the dental – postalveolar region since the shape of the tip-blade can be varied in subtle ways. The latter property is also characteristic of laterals and r-sounds. Flaps and taps illustrate the nimble dynamic capabilities of the tongue tip. Clicks show five types: bilabial, dental alveolar, palatal and lateral that can be combined with various glottal actions (voiced, voiceless, breathy, and aspirated).

Vowels are classified in terms 'tongue height', front-back position and lip positions. Additional dimensions include nasalization, advanced tongue root, pharyngealization, frication and retroflection. Phonation types include modal voice, voiceless, creaky voice, stiff voice, breathy voice and slack voice.

Many articulations analyzed by L & M involve a single articulatory movement combined with glottal activity and control of velar opening. In addition somewhat more complex configurations can occur: double closures (labial-dorsals) and secondary modifications (e g labialization, velarization, palatalization, pharyngealization).

On the basis of surveys such as the L & M project and the UPSID database it is possible to make a list of the segment types so far documented. For instance, we can ask: How many different types of /p/? The answer given by UPSID (1984) is: 1. voiceless; 2. long voiceless; 3. palatalized voiceless; 4. voiceless aspirated; 5. long voiceless aspirated; 6. labialized velarized voiceless aspirated; 7. palatalized voiceless aspirated; 8. voiceless pre-aspirated; 9. voiceless with breathy release; and 10. laryngealized voiceless. Moving on to a search through the entire UPSID database we find 556 different consonant segment types and 210 different vowel segment types.

What do these lists tell us? Do they represent hotspots in phonetic space from which languages draw their subsets of contrast?

Systemic constraints

One account of phonetic hotspots is given by the Quantal Theory of Speech (Stevens 1989). It was developed from noting that certain regions in acoustic space are more stable and insensitive to articulatory imprecision than others and then suggesting that languages seek these regions out in the interest of making the acoustic correlates of phonetic units, if not invariant, so at least less variable.

Another way of thinking about finite phonetic inventories starts from assuming that speech perception is not committed to signal invariance. It assumes that correct recognition and comprehension of a spoken word depends on both signal contents and listener knowledge and that the speaker adapts in the short-term to the listener's changing needs for signal information. The acoustic signal produced by such a talker would show variations in the explicitness of the physical cues. Pronunciation would sometimes be clear and rich, sometimes casual and reduced. Therefore the invariance of a given phonetic form would not be expected to be found in the physical signal but only at the level of interaction between signal and stored knowledge that is at the level of comprehension. This reasoning is known as the H&H (hyper&hypo) argument (Lindblom 1990)

According to such a perspective, there are no hotspots. Phonetic values are drawn from a continuous space of possible vocalizations (Catford's anthropophonic space) and unwittingly shaped by speakers, listeners and learners and a tacit demand for robust and reliable communication. Selections are constrained by those factors but are in principle made without limit rather than from a finite set of discrete a priori resources.

One example of such a performance-based criterion is 'contrast', a central concept in phonological analysis. To signal different meanings, words need to be perceptually distinct. Distinctiveness is not an absolute phonetic property. It is a systemic notion. 'Distinctive' means being different from all other elements in a system. The following simplified model of how speech is perceived helps us clarify that point.

When a given word is spoken, it is processed by the listener's auditory system and then interacts with a set of reference patterns stored in memory. The interaction involves the activation of all the items of this mental lexicon in a direct and parallel manner. The degree of activation depends on the similarity of input and reference pattern. The spoken word "resonates with" what is in storage (Shepard 1984). The resonance model suggests that it is the biggest response that determines the identity of the percept.

Two points: Languages are well advised to keep the phonetic shapes of syllables and words distinct because that facilitates the recognition process. Also there is the possibility of listener "expectations" controlling, in a top-down fashion, the sensitivity threshold of the reference elements thereby influencing the processing and thus the distinctiveness of the signal.

The signal is not solely responsible for successful lexical access.

Computer simulations of vowels systems (Liljencrants & Lindblom 1972) lend support to claiming that "*.. a vowel dispersion theory correctly captures a principle governing the distribution of vowels in*

natural languages" (Disner 1984). The findings suggest that [i e a o u] - the world's most common vowel inventory (Maddieson 1984) – is favoured, not because they individually score high in stability and have certain absolute phonetic values, but because they are sufficiently dispersed in the available space. As a system, they achieve workable robustness in the face of signal degradation and low signal-to-noise conditions.

Contrast is a systemic notion which means that the distinctiveness of a segment, syllable or word depends on how its physical attributes contrast with other competing elements. The cohort constituting the competition is not fixed but varies from moment to moment. The typological evidence strongly suggests that perceptual contrast plays a significant role in the design of both vowel and consonant systems.

The Size Principle

Phonetic systems also bear the marks of productions demands. Ease of articulation is an intuitively appealing explanatory principle but it still lacks a rigorous definition. Nevertheless, there is some indirect evidence that sound systems are shaped by it.

One approach to quantifying this notion was proposed by Lindblom & Maddieson (1988) who performed a sorting of the UPSID 1984 vowels and consonants into three categories: Basic, Elaborated and Complex.

The method was to compare all segments with the respect to multi-tasking - in other words, and to make estimates of the number of articulatory tasks required by any given speech sound. Segments with 'secondary articulations' and 'double articulations' were classified as Elaborated along with those with 'derived' properties: implosives (enhanced voicing), ejectives (reinforcement of release), affricates (extended frication in stop releases), voicing (non-spontaneous) in fricatives and segments involving extreme displacements (retroflexion, subapicals, pharyngeals) and modified default timing patterns (prenasalization, aspiration, prestopping).

Segments having multiple (two or more) extra articulations were labelled Complex. When all single and multiple processes had been identified a small set of Basic segments remained: [p t k ʔ b d g f s h m n ŋ l r w j], all segments involving movement of a single articulation.

When the distribution of Basic, Elaborated and Complex UPSID segments is plotted as a function of inventory size, a systematic pattern emerges: Small systems have mainly Basic articulations. Moderately large inventories have Basic and Elaborated segments. The biggest systems have all three types. Increasing system size entails greater articulatory complexity.

The probabilities of randomly selecting [iᶻ ẽ a̰ o̰ uˤ], or [i e a o u], from the 200 vowel UPSID segment list are equally small, vanishingly so. The Size Principle states that phonetic inventories are products of a tug of war between perceptual contrast and pronounceability. The first of the above systems uses secondary articulations (apicalization, nasalization, breathy voice, laryngealization, and pharyngealization) and could arguably make all elaborated pairs more contrastive than the corresponding basic ones. However, the typological data show beyond all doubt that [i e a o u] is the most frequent system. It appears justified to interpret this to mean that the perceptual advantage achieved by adding the secondary articulations is not sufficient to motivate their extra articulatory 'cost'.

That is the essence of the balance described by the Size Principle which is honoured by both vowel and consonant systems.

'Ease of articulation'

Speech uses a bio-mechanical system with numerous degrees of freedom. Given the rich articulatory and acoustic possibilities there are many dimensions that could be invoked to enhance phonetic contrasts. The typological data indicate that those opportunities are balanced by articulatory factors and left largely

unexplored (cf comments on [i e a o u) above). How do speakers go about applying an 'ease of articulation' criterion? Two brief comments.

The 'minimum variance theory' of movement was presented to account for arm and eye movements (Harris & Wolpert 1998). It is based on finding that neural signals tend to be corrupted by noise whose variance increases with the amplitude of the signal (Jones et al (2002). In the presence of such signal-dependent noise, the motor system was observed to adopt a precision criterion that minimizes the variance of the error in reaching the movement endpoint (Hamilton & Wolpert 2002).

Of relevance to the topic of ease of articulation is the fact that high precision implies low-amplitude signals. That strategy also means lower metabolic energy consumption. In other words, the energy cost of the movement is not explicitly monitored but emerges as a by-product of the control.

Both production and learning stand to gain from physiological energy minimization. In speech development, a 'minimum variance' process will bias the child's search for the phonetic forms of the ambient language in the direction of articulations of low complexity. By staying 'easy to produce' a sound system would supplement the child's attempt to imitate what she hears and facilitate the spontaneous production of ambient patterns (cf canonical babbling). An approach of '*Easy-way-sounds-OK*' would contribute to boot-strapping the phonetic development.

On the origin of phonetic units

The syllable

The patterning of the world's speech sounds cannot be investigated without reference to the syllable. Where does this unit come from? The answer given below is inspired by MacNeilage's account of the beginnings of speech (2008): Syllables have their roots in mechanisms developed a long time before human language appeared.

During the past century, attempts were made to specify vowels in terms of their jaw angle, or "Kieferwinkel" (Menzerath & Lacerda 1933). However, classical phoneticians soon lost interest in this parameter when it was pointed out that it is perfectly possible to produce normal sounding vowels with a pencil between the teeth. Since the talker's acoustic output is directly linked to how the tongue shapes the cavities of the vocal tract, a certain lack of interest in the jaw still persists in contemporary phonetics and phonology.

However, the jaw is on its way back. One reason is that, although the tongue clearly shapes the vocal tract, it does so in synergy with the jaw. Jaw movement is a normal attribute of natural speech and its actions have significant articulatory and acoustic consequences (Maeda 1990). So fixed-jaw experiments do not make the jaw irrelevant. Rather it helps highlight the compensatory abilities of the speech motor system.

In fact it turns out that the mandible is central to an insightful description of speech processes (MacNeilage 2008). To reinforce that point we will briefly review two cross-linguistic topics on which jaw movement throws explanatory light: (i) the preferred phonotactic patterns across languages; (ii) the origin of coarticulation.

Phonotactic patterns

Phonotactics is the description of the consonant sequences that are permissible before and after the syllable nucleus (the vowel). The sequencing in clusters reveals a tendency for consonants to order themselves according to their degree of vowel adherence - a quantitative and distributionally based number proposed by Sigurd (1965).

This measure (VAM) is illustrated in Figure 2 with data on initial and final 2-member consonant clusters from Sigurd (1965:49, 74). It is defined as the difference between the number of clusters that have a given consonant in the vowel adjacent position and the number of clusters that have that segment in the non-adjacent position. A positive or negative VAM indicates correspondingly strong or weak vowel adherence. VAM is calculated in the same way for initial and final clusters.

Figure 2 shows VAM values for initial and final 2-member clusters plotted against each other. It can be seen that the measure is high for [r] and [l] and intermediate for nasals, sonorants and stops and lowest for [s].

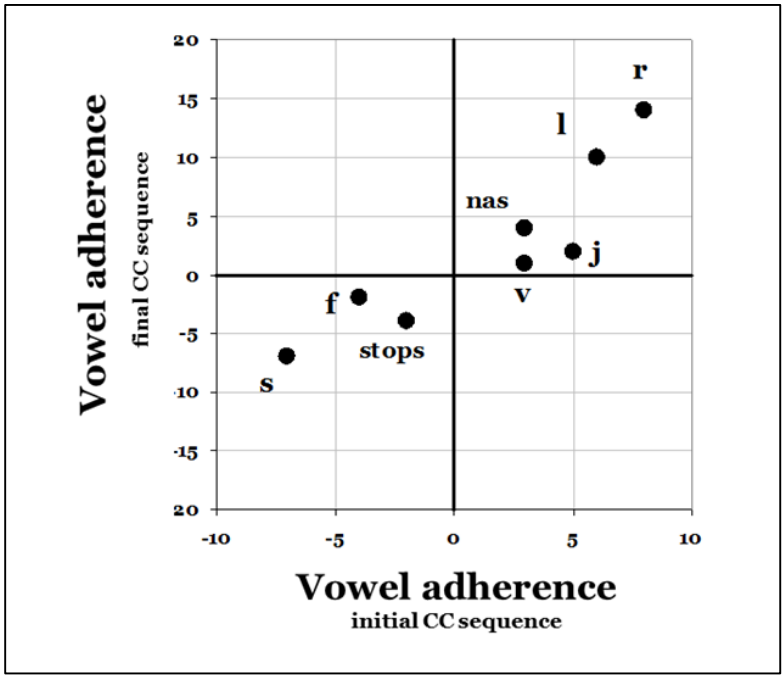

Figure 2 Comparison of vowel adherence values for Swedish CCV and VCC clusters. Abscissa: Initial clusters. Ordinate: Final clusters. Source: Sigurd (1965).

The values for initial and final clusters are similar. This correlation implies that vowel adherence is symmetrical and thus gives rise to mirror image effects; For instance, [skr] occurs initially but not finally, whereas the reverse sequence [rks] is only found finally but not initially. A great many languages conform to the vowel adherence effect.

The following experimental results (Lindblom 1983) give an idea of its phonetic basis.

Subjects were asked to produce symmetrical V1CV2 sequences in which the Swedish fricatives, stops, nasals and liquids occurred in the C slot and a range of vowels with different degrees of opening were used for the V. The movement of the jaw was tracked using the Stockholm University Movetrack system (Branderud 1985). The recorded traces were examined with respect to the position of the jaw in the consonant as well as in the surrounding vowels. As expected the curves all conformed to an open-close-open pattern. The closed positions for the consonants were influenced by the surrounding vowels, most strongly so, between open vowels. The comments below concentrate on the findings for the [’ɑː_a] context which are summarized in Figure 3.

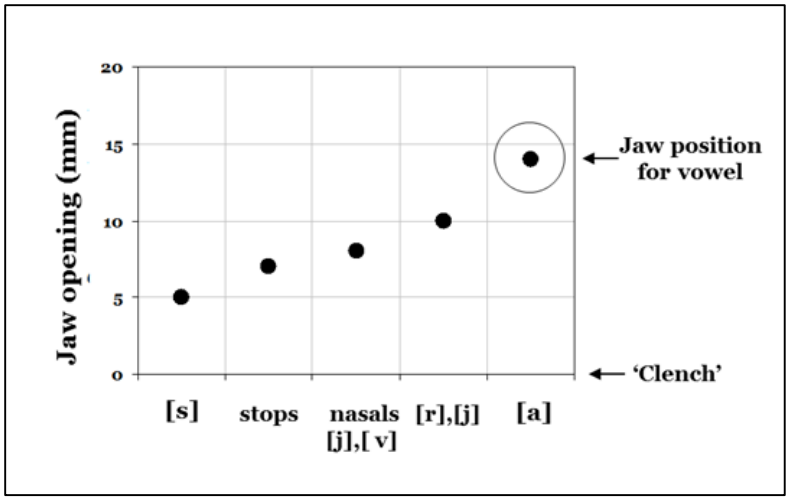

Figure 3 Average jaw opening (in mm relative to clench) for phonetic segments arranged in order of increasing vowel adherence.

As can be seen the tightest closure is observed for [s]. The most open consonant segments are [r] and [l]. Arranging the measurements in increasing order of jaw opening we obtain [s], stops, nasals, sonorants, [r] and [l] – a series that matches the VAM scores perfectly. In these clusters consonants are ordered according to their compatibility with the vowel.

Turbulent noise production in [s] requires air hitting a perpendicular surface with great velocity. Stevens (1998) describes how that happens: "*... The constriction is adjusted so that the airstream emerging from the constriction impinges on the lower incisors*." Conclusion: [s] requires a high jaw.

According to articulatory modeling studies (Lindblom & Sundberg 1971), increasing the jaw opening while keeping the tongue shape fixed will result in a backward movement of the tongue body. Thereby the size of the pharynx cavity is reduced which facilitates making the low and posterior constriction of [a]. Accordingly the natural way to say [a] is with a lowered jaw and to let the tongue and the jaw work in synergy. Conclusion: [a] is an open vowel.

These results suggest that the phonetic basis of vowel adherence is compliance with jaw openness.

Coarticulation in deep time

Further evidence for the jaw's explanatory role comes from segment durations in clusters. A general trend is that the larger the cluster the more the cluster segments are shortened (Haggard 1973). The effect is stylized in Figure 4. The notation is borrowed from music to represent the timing in three Swedish words: [so:r], [spo:r] and [spro:k] (sows, track, language).

Let us compare the three initial [s] segments. In [so:r], its duration is represented by the quarter note. In [spo:r] and [spro:k] it shortens to one half (quaver note) and one third (triplet) of that value respectively. A single jaw trace is drawn to suggest that the jaw is the 'pace maker' that provides the 'beat' common to the three words. Admittedly this is a stylization of the facts but one of sufficient realism for the point we are making.

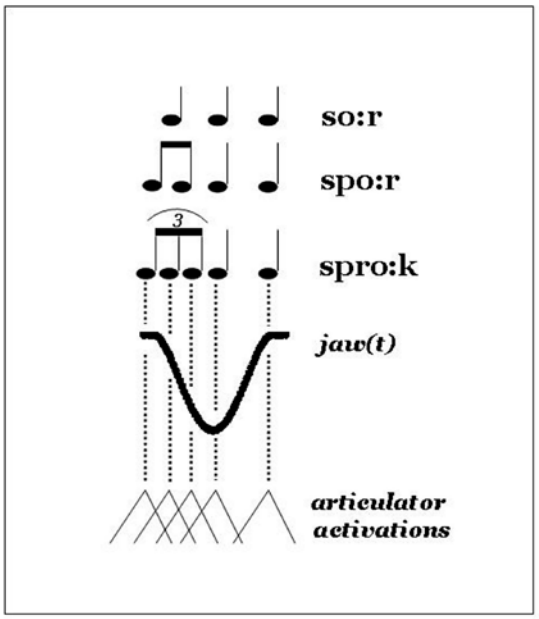

Figure 4 The role of the jaw/ syllable in coarticulation. In musical notation a stylized representation of the timing of the segments in three Swedish words. If a jaw trace common to the three words is assumed the initial consonants are forced to adjust their durations. Vertical lines connect the notes/segments of [spro:k] with the rising and falling lines on the bottom line. They represent the articulatory anticipation/de-activation for [s] [p] [r] etc. The syllable/jaw sets the tempo. Articulatory actions show overlap in time.

The jaw is the 'rhythm section'. The notes (one per segment) indicate the timing of the various articulatory movements. They are not instantaneous. Each one requires periods of anticipation and de-activation. That is indicated at the bottom of the diagram by the saw-tooth shapes of rising and falling lines. On-glides and off-glides take longer than the interval available between the notes. Consequently, the articulatory actions overlap in time which means that there is 'coarticulation'.

Coarticulation arises because of the priority of the basic close-open-close rhythm generated by the jaw. To correctly time the articulatory actions for the initial consonants, the on-glides and off-glides segments must be fitted into the intervals allotted by the jaw-based score. This is done by having the actions overlap, a solution that is facilitated by the use of alternate articulators such as the tongue blade for [s], lips for [p] and the tongue tip for [r].

The old beginnings of the syllable

The open-close alternation of the jaw certainly helps us understand better why there are phenomena such as vowel adherence and coarticulation. However, observations of swallowing and mastication indicate that the jaw is capable of making highly complex motions in precise coordination with tongue activity (Luschei & Goldberg 2011). The oscillation seen in speech represents only a small fraction of its full capacity. Why is its phonetic repertoire so limited and more or less confined to a single dimension?

It is often said that humans lack 'speech organs' proper. Speech is an 'overlaid function' shaped by chewing and swallowing (Kent 2004). Evolution is known to tinker with available resources and to prefer building on existing capacities. Such observations point us in the direction of those capacities for a better understanding of the jaw's speech movements.

The motor systems that became the "vocal" tract were not a tabula rasa but a sophisticated set of tools serving various functions, notably breathing and chewing. It does not appear far-fetched to assume that early speech was significantly moulded by the motor mechanisms already in place (Lund & Kolta 2006).

Breathing, chewing and locomotion exemplify a large class of movements that are cyclic and continuous. Many studies have shown that such behaviours involve central pattern generators (CPG:s) which produce rhythmically alternating motions (Grillner 2006). It appears plausible that the CPG system for chewing came to be used in speech, especially because chewing and speech require the coordination of very similar muscle sets. In speech, precision rather than power is called for. To meet that goal the network would have to be driven at a low-amplitude excitation level (Harris & Wolpert 1998) which - we might speculate - would result in an elementary open-close oscillation near the mandibular system's resonance frequency and which would satisfactorily meet precision demands and communicative needs.

Summarizing this section we make two points: (i) Evidence was presented for giving the open-close alternation of the jaw a key role in explaining preferred cross-linguistic phonotactic patterns and identifying the source of the spatial and temporal overlap of articulators known as coarticulation. Although reinforcing belief in the importance of the jaw, this result still left the question unanswered of where the open-close alternation comes from. For a reply we adopted a phylogenetic perspective concluding that the alternation is a natural product of using the motor mechanisms on which speech is an overlaid function - a pianissimo gently performed on the CPG keyboard of mastication. This natural motion captures the phonetic essence of the 'syllable'.

The discreteness of phonetic segments

Clues from the IPA.

The practice of describing speech sounds in terms of static articulatory attributes can be traced far back in phonetic history. The International Phonetic Association provides the world with the notational standard for the phonetic description of all languages. Its alphabet, revised in 2005, is presented in charts containing the current set of phonetic symbols and diacritics. Consonants are specified with respect to manner and place and voicing, vowels in terms of front-back, open-close and rounded-unrounded. For phonation types and secondary articulations diacritics are used.

Do the descriptive labels given to the dimensions of the IPA (i.e., place, manner, source etc. categories) reveal anything about how the different segments are produced by the motor system? Most of the classificatory terms refer to static states of vocal tract (such as 'voiced', 'voiceless', 'labial', 'dental', ' uvular', 'stop', 'nasal', 'fricative' …) implying a target-driven rather than gesturally based organization of speech production. For example, the slots along the place dimension could be interpreted as static positional goals for the motor system and the motions from and to those regions could be derived as secondary consequences. The patterns representing the acoustic targets of manner and source features could be similarly defined in terms of discrete steady-state vocal tract configurations.

Clues from experimental phonetics

The term 'target' has been widely used with reference to articulatory positions or acoustic patterns that a talker aims at but does not always reach (Stevens & House 1963). It continues to produce the most parsimonious descriptions of complex articulatory and acoustic data.

An example is the recent set of measurements of the articulatory and acoustic properties of bV, dV and gV syllables reported by Lindblom & Sussman (2011). The goal of the research was to predict the coarticulatory patterning of stop-vowel formant transitions by means of a computational articulatory model. The observed effects were used as a window on the underlying control processes. Good quantitative accuracy was obtained by using a single target configuration for each stop and each vowel and by generating the transition as a smooth trajectory with fixed shape and time constant and scaled to interpolate between the consonant and vowel targets. These results extend and reinforce Öhman's classical target-based account (Öhman 1967).

The developmental implication of these two studies is that they reveal the end state of learning to produce and coarticulate stops correctly: For each place of articulation the child needs to find the unique context-independent target – the representation that sounds right and generates the coarticulation patterns of the input. The modelling demonstrated that, once the onset and endpoint were known, any given CV transition was fully specified. This further implies that, once the child gets the correct spatial targets set up, the articulatory and acoustic properties of the transition will unfold automatically as a result of the non-linear articulation/acoustics mapping and the general motor mechanisms of trajectory formation (Flash & Hogan 1984, Shadmehr & Wise 2005).

Clues from neurobiology

Many non-speech continuous motions produced by CPG networks also allow superposition of discrete positioning. An example is the special case of locomotion known as '*precision walking'*. It requires the visually guided coordination of the rhythmic CPG activity and the signals for the discrete movement generated by neurons in the motor cortex in precise phases of the movements (Grillner 2006, Ijspeert 2008).

A similar analysis is presented for reaching tasks. "The motor cortex and its corticospinal outflow are preferentially engaged when precise positioning of the limb is needed during locomotion and are also involved during reaching and active positioning of the hand near objects of interest" (Georgopoulos & Grillner 1989).

It does not seem far-fetched to fit speech into this framework. Paraphrasing the formulation we just used above, we could describe it as requiring auditorily (rather than visually) guided coordination of the rhythmic CPG activity (read: the syllabic rhythm of the jaw) and the signals for the discrete movement (the phoneme command) generated by neurons in the motor cortex in precise phases of the (jaw) movements.

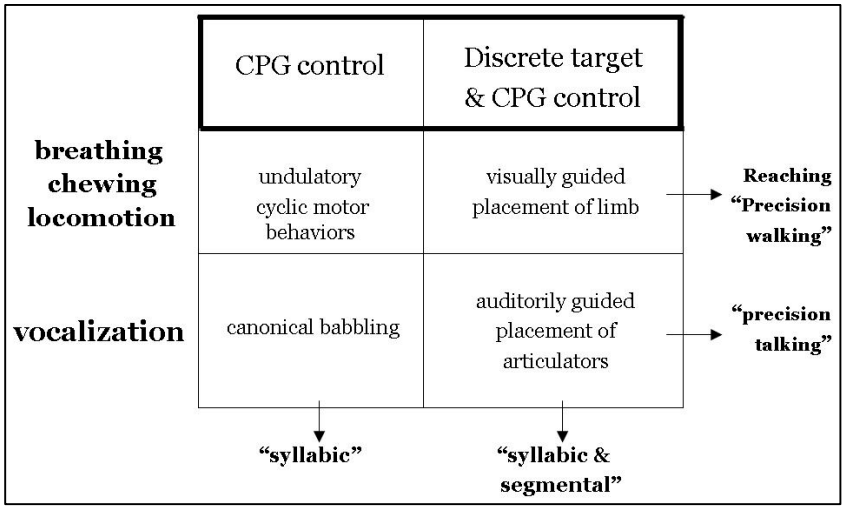

Figure 5 Top row describes non-speech motor mechanisms. Bottom row indicates how those mechanisms are paralleled in speech (from Lindblom & MacNeilage 2011).

Figure 4 highlights the possible parallels between speech and non-speech movements. In the top left cell examples of CPG-based rhythmic movements such as breathing, chewing and locomotion are exemplified. In the right hand cell we find locomotion that involves exact foot placement: '*precision walking'*.

The bottom cells extrapolate to speech. Canonical babbling is here assumed to be driven by CPG networks also serving chewing. Infants do not have sufficient control to produce genuine segments but gradually become able to coordinate the basic jaw movement with articulatory activity. Rather than just

letting the jaw close the lips as in canonical babbling, the child may intentionally reinforce that labial occlusion by active muscular action, or may, with time, successfully make it elsewhere using a different articulator. As this placement of articulators is generalized it represents a form of discrete target control - 'precision talking' - a close parallel to 'precision walking'.

We hypothesize positional control as the source of discreteness in phonemic units. It is a general process that is available to the child and would facilitate the launching of the discrete coding of speech patterns accordingly taking the role of an important boot-strapping function.

Vocal imitation and analysis-by synthesis

Children learn to link speech percepts from the ambient language to their own articulatory actions. The possibility of a neural account of vocal imitation is suggested by the discovery of mirror neurons (Gentulucci et al 1988, Rizzolatti et al 1988). These neurons which were first identified in the macaque's premotor cortex, discharge when the monkey manipulates objects and when it observes other monkeys or humans do the same. Neurons that respond to sound and to communicative or ingestive actions have also been identified (Ferrari et al 2003). In humans the existence of a cortical network with the properties of mirror neurons has been demonstrated (Cattaneo & Rizzolatti 2009, Pulvermüller & Fadiga 2010). Brain stimulation and imaging studies indicate increased activity in speech muscles when subjects listen to speech (Fadiga et al 2002; Watkins et al 2003, Wilson et al 2004).

This sound-to-movement mapping is obviously an important aspect of phonetic learning but there is an issue concerning its resolution. The question is whether its developmental role extends to all the fine details of speech signals. It appears reasonable to look also for other sources aiding children to emulate what they hear. For instance, consider ambient patterns that are simple for the child to produce. They may spuriously appear to be imitations but could in fact be produced spontaneously because of their sheer naturalness - cf the pseudo-syllabic nature of canonical babbling. In other words, if sound patterns keep their articulatory complexity low, chances are that, during his exploration of the phonetic space, the child might accidentally score quite a few hits (*'Easy way sounds OK'*).

The child's path to a phonemically organized vocabulary might be a case where such an (unintended) analysis-by-synthesis strategy would come into play and yield significant pay-offs.

Phonetic recombination

Motor tools

The insights we need to make better sense of phonetic recombination in children come once more from research on non-speech motor mechanisms. Parsimoniously, we assume that motorically speech is not special. Therefore what we learn from this domain should apply also to speech.

We highlight two aspects: (i) The organization of motor systems is output-oriented. Motor equivalence (Lashley 1951) exemplifies this type of control; (ii) Movement paths (as in point-to-point reaching) exhibit certain invariant kinematic properties including 'minimum jerk' and bell-shaped velocity profiles (Flash & Hogan 1984, Shadmehr & Wise 2005).

The conclusion from the first topic is that spatial goals can be reached from arbitrary onset points provided that onset and endpoint are located within the system's work space. From the second topic we learn that movement paths can be scaled according to the extent and duration of the task but their shape and time constant tend to be fixed across a large range of conditions.

With respect to speech development the implication is the following: When the child puts these capabilities into play they will enable her to make the appropriate transition between any two speech

sounds provided that she has mastered their target representations. In other words, she is ready to "recombine".

Motor equivalence

The classical example of motor equivalence is handwriting (Lashley 1951). Letter shapes tend to stay constant whether one writes on a piece of paper, a blackboard or in wet sand making the strokes with the foot. Such examples show that motor control is output-oriented: There is constancy in the outcome although each time the motor system has to recruit different sets of muscles to get the job done. The motor implementation depends on the context but the goal stays the same.

The task of reaching for a glass in a specific location is carried out by different muscles depending on whether the glass is above, below, in front of or behind the person performing the action. Motor systems evolved to operate in a compensatory mode. They are set up to deal with the unforeseen. Motor equivalence refers to the ability of the system to move an effector (arm, foot, articulator …) from its current location A to an arbitrary position B within the work space.

Kinematic constancies of non-speech (& speech) movement paths

By parsing the continuous flow of speech into strings of targets, we treat it as a sequence of reaching gestures. A lot of research has been done on point-to-point reaching movements (Shadmehr & Wise 2005). A salient fact about such movements is that they are highly replicable from trial to trial and from subject to subject. The trajectory approximates a straight line. Its velocity profile is consistently a unimodal, bell-shaped curve. Those features are retained when position, extent and duration are varied. The big questions are how and why the brain arrives at a unique solution each time with the above characteristics. The arm-hand system has many degrees of freedom and subjects are certainly capable of using them to make various more "unnatural" motions. Accordingly there are strong indications that motor systems arrive at uniqueness by adopting criteria that limit the search space to a single optimal choice (Soechting and Flanders 1998). That is presumably also true for speech gestures.

The classic study by Flash & Hogan (1984) presents a model of point-to-point hand trajectories. They found that the computed movements matched observed movements with high accuracy when they minimized the derived curve's rate of change of acceleration - the so-called 'minimum jerk' criterion. They remarked that "*the predicted trajectory depends only on the initial and final positions of the hand. It is invariant under translations and rotations. Thus the shape of the predicted trajectory does not change with amplitude or duration of the movement, which merely serve to change the scale of the position and time axes, respectively*".

Ready to build an open-ended vocabulary

As her motor development progresses, the learner can use a) positional target control, b) her ability to produce motor equivalent movements and c) her ability to make minimum jerk transitions. These tools and skills will enable her to represent phonetic input as strings of context-free control objects (targets). It will also obviate the necessity to include details on the articulatory properties of the transitions. They will unfold automatically as a result of how motor mechanisms handle trajectory formation.

Assume a child who consistently uses [ba] for 'ball' and [dɛ] for 'daddy'. She often hears the words 'bed' and 'dog' but does not yet attempt to say them. However, since she is familiar with the stop and vowel targets of ball' and daddy', she is in principle ready to recombine them and generate the new forms [bɛ] for 'bed' and [da] for 'dog'.

The key idea of the present account is captured by the phrase "once the targets … have been established". If the target assumption is accepted, the construction of an open-ended lexicon by recombining context-free discrete units becomes less of a mystery. We have come a bit closer to being able to deal with the questions posed in the introduction by proposing that: Positional control paves the way for discrete units; Motor equivalence offers context-independence; Motor equivalence and minimum jerk movement paths produce the appropriate movement between any two contiguous targets.

Final piece of the puzzle: Where re-use comes from

If we test what we have so far by also asking "Where does re-use come from?" we realize that our sketch addresses that issue only in part. It would be fair to say that we now have an idea of how re-use gets *maintained* by each child. However, our account does not get to the ultimate source of the phenomenon. It leaves children's development of re-use to the ambient input. In other words, since all languages exhibit that feature, it is not surprising that it shows up also in the acquired lexicon of each individual. To answer more fully it will be necessary to seek an evolutionary answer.

Although we cannot make direct observations of the early steps taken by our ancestors towards phonetic re-use in vocabularies, it appears likely that they occurred in the first attempts at creating and memorizing new words. According to Donald (1991) a communicative 'mimetic' culture arose during the period of Homo erectus allowing individuals to share mental states and to represent reality in new and expanding ways. Man's growing cognitive capacity created a strong demand for fast and precise ways of communicating. In Donald's estimate spoken language appeared after the period of archaic Homo sapiens in response to such demands. An abundance of meanings had to be paired with their individual phonetic shapes – either by the invention of holistic signals, or by building on the vocal patterns already in use.

Evolution's preference for tinkering with existing resources makes the latter route more probable. New meanings had to sound different. Pre-linguistic vocalizations were subject to same production constraints that we have found in present-day speech and sound systems. When new things had found their names they had to be committed to memory. It is therefore likely that the available vocal patterns already formed a small subset of the 'anthropophonic' space, i.e., man's total sound production capacity (Catford 1982). The playing field was not level with respect to that total capacity. The phonetic search space would have been reduced by tacit preferences of 'easy to say', 'easy to hear' and 'easy to learn'. To come up with phonetic shapes for a large number of meanings, would involve making forced choices and occasionally re-using the discrete elements of the available resources.

After reaching the canonical babbling stage infants become better able to coordinate the basic jaw movement with articulatory activity. There then follows a period when the child explores its capabilities by abandoning the passive, default behaviour involving just the jaw and begins to vary its placement of articulators (variegated babbling). The moment when, stimulated by what she hears around her, the child intentionally reinforces the labial occlusion of the babble by active muscular action, or uses the tongue to make the closure elsewhere, she takes the first step towards discrete positional control - 'precision

talking'. As her exploration continues several more targets are born. It is not inconceivable that, at a distant point in time, human vocalizations took a similar path.

In terms of bit rate the information of a holistic vocal pattern is expensive. By definition its overlap with other patterns is minimal and therefore its specification takes a lot of bits. Modifying an existing item is cheaper since the overlapping parts come free of charge. A physiological basis can be found for this economy argument. In biology memory storage is associated with a biochemical 'cost' which derives from the energy metabolism of memory formation. Brains change physically as a result of learning. This change is activity-dependent. Active neurons contain more energy-rich substances. Should a demand for their activation arise (e.g., recall), active neurons have the 'fuel' to respond. Building up that capacity costs metabolic energy.

It is possible to measure the effect of learning on brain tissue. A substance used as an index of metabolic capacity is cytochrome oxidase. More active neurons have greater amounts of this enzyme. Experiments were run on rats trained to associate reward with an auditory stimulus. Histochemical analyses of brain tissue were then performed on experimental and control animals. The experimental group showed significantly increased amounts of cytochrome oxidase in the auditory neostriatum. The memory of the conditioning stimulus had modified the neurons activated by the task (Gonzales-Lima 1992).

The implication of this type of work is that patterns that contain more information and whose specification therefore requires more bits, are energetically more costly, and therefore they take longer to commit to memory. The new phonetic forms that would be especially favoured by the metabolic factors of memory formation would be the ones that conform to a '*minimal incremental storage'* criterion. Such considerations give some substance to the notion of 'easy to learn'.

With vocalizations under positional control in the repertoire, opportunities would have arisen for evolution to apply the re-use method. Once it had been introduced it would have been maintained and reinforced owing to the numerous advantages it offered: Its emergence can be seen as a case of selection-by-consequences.

Conclusion

The goal of this paper was to demonstrate the feasibility of a non-structuralist approach to the basic units of speech. Accordingly we reversed structuralist priorities starting from substance (performance-based factors) in a quest for possible behavioural causal factors underlying sound structure formation. We found that a strong case can be made for the claim that the combinatorial structure of sound patterns has behavioural origins.

References

Anderson S R (1981): "Why phonology isn't "natural"', *Linguistic Inquiry 12*:493-539.

Anderson S R (1985): *Phonology in the twentieth century*, Chicago: Chicago University Press.

Branderud P. (1985). "Movetrack - A movement tracking system". *Perilus IV*, Stockholm University, 20-29.

Catford J C (1982): *Fundamental problems in phonetics*, Indiana University Press: Bloomington.

Cattaneo & Rizzolatti (2009): "The mirror neuron system", *Arch Neurol 66(5)*:557-560.

Chomsky N (1964): "Current trends in linguistic theory" 50-118 in Fodor J A and Katz J J (eds): *The structure of language*, New York:Prentice-Hall.

Disner S (1984): "Insights on vowel spacing", 136-155 in Maddieson I (1984): *Patterns of sound*, Cambridge.

Donald M (1991): *Origins of the modern mind*, Cambridge, MA: Harvard University Press.

Dryer, M S & Haspelmath M (2011): *The World Atlas of Language Structures Online*, Munich: Max Planck Digital Library, Available online at http://wals.info/

Edlund S (1957): *Studier rörande ordförrådsutvecklingen hos barn i skolåldern* (Studies of vocabulary development in school children), Lund, Sweden, Gleerup.

Flash T & Hogan N (1984): "An organizing principle for a class of voluntary movements", *Journal of Neuroscience 5(7)*: 1688-1703.

Georgopoulos A P & Grillner S (1989): Visuomotor coordination in reaching and locomotion", *Science* 245:1209–1210.

Gonzales-Lima F (1992): "Brain imaging of auditory learning functions in rats: Studies with fluorodeoxyglucose autoradiography and cytochrome oxidase histochemistry", 39-109 in Gonzales-Lima F, Finkenstädt T & Sheich H (eds): *Advances in metabolic mapping techniques for brain imaging of behavioral and learning functions*, NATO ASI Series D:68, Dordrecht: Kluwer.

Grillner S (2006): "Biological pattern generation: The cellular and computational logic of networks in motion", *Neuron 52*:751–766.

Haggard M (1973): "Abbreviation of consonants in English pre- and post-vocalic clusters", *J of Phonetics* 1(1):9-24.

Hamilton A F & Wolpert D M (2002): "Controlling the statistics of action: Obstacle avoidance", *Neurophysiol 87*:2434–2440.

Harris C M & Wolpert D M (1998): "Signal-dependent noise determines motor planning", *Nature* 394:780-784.

International Phonetic Association (2005): http://www.langsci.ucl.ac.uk/ipa/

Ijspeert A J (2008): Central pattern generators for locomotion control in animals and robots: a review, *Neural Networks 21/4*:642-653.

Jespersen O (1926): *Lehrbuch der Phonetik*, Leipzig:Teubner.

Jones K E, Hamilton A F & Wolpert D M (2002): "Sources of signal-dependent noise during isometric force production", *Neurophysiol 88*:1533–1544.

Kent R D (2004): "Development, pathology and remediation of speech", in Slifka J, Manuel S & Matthies M (eds) (2004): *From sound to sense: 50+ years of discoveries in speech communication*, conference proceedings (CD-ROM format), Research Laboratory of Electronics, MIT, Cambridge, Mass.

Kroos C, Geumann A & Hoole P (1999): "Tongue–jaw trade-offs and naturally occurring perturbation", *J Acoust Soc Am 105(2)*:1355-1355.

Lashley K (1951): "The problem of serial order in behavior", 112-136 in Jeffress L A (ed): *Cerebral mechanisms in behavior*, Wiley:New York.

Lindblom B (1990): "Explaining phonetic variation: A sketch of the H&H theory", 403-439 in Hardcastle W J & Marchal A (eds): *Speech Production and Speech Modeling*, Dordrecht:Kluwer.

Lindblom B & MacNeilage P F (2011): "Coarticulation: A universal phonetic phenomenon with roots in deep time", *FONETIK 2011*, 41-44 in TMH - QPSR Vol. 51, KTH Stockholm, http://www.speech.kth.se/prod/publications/files/3588.pdf

Lindblom B & Sussman H M (2012): "Dissecting coarticulation: How locus equations happen", *J of Phonetics 40(1):*1–19.

Lund J P & Kolta A (2006): "Brainstem circuits that control mastication: Do they have anything to say during speech?", *J Communication Disorders 39(5)*:381–390.

Luschei E S & Goldberg L J (2011): "Neural mechanisms of mandibular control: Mastication and voluntary biting", *Comprehensive Physiology* 1237–1274.

Maddieson I (1984): *Patterns of sound*, Cambridge.

Maeda S (1990): "Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model", 131-149 in Hardcastle W J & Marchal A (eds): *Speech Production and Speech Modeling*, Dordrecht:Kluwer.

Menzerath P & Lacerda A (1933): Koartikulation, Steuerung und Lautabgrenzung, Berlin.

Miller G A (1977): Spontaneous apprentices: Children and language, New York: Seabury.

Miller G A (1991): *The science of words*, New York: Freeman.

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

Perkell J S & Klatt D H (1986): *Invariance and variability in speech processes*, Hillsdale, NJ: Erlbaum.

Shadmehr R & Wise S P (2005): *The computational neurobiology of reaching and pointing*, The MIT Press: Cambridge MA.

Shepard R N (1984): "Resonant kinematics of perceiving, imaging, thinking and dreaming", *Psychological Review 91(4)*: 417-447.

Sigurd B (1965): *Phonotactic structures in Swedish*, Lund, Sweden, Gleerup.

Stevens K N (1989): "On the quantal nature of speech," *J Phonetics 17*:3-46.

Stevens K N (1998): *Acoustic phon*etics, Cambridge MA: M.I.T. Press.

Stevens K N & House A S (1963): "Perturbation of vowel articulations by consonantal context. An acoustical study", *J Speech Hearing Res* 6:111-128.

Teleman U (2012): "Ordförråd", *Nationalencyklopedin*.