Language as a social institution: Why phonemes and words do not have explicit psychological form

Robert Port
Indiana University
April 27, 2010
Draft for Ecological Psychology

Abstract

It is proposed that a language, in a rich, high-dimensional form, is part of the cultural environment of the child learner. A language is the product of a community of speakers who develop its phonological, lexical and phrasal patterns over many generations. The language emerges from the joint behavior of many agents in the community acting as a complex adaptive system. Its form only roughly approximates the low-dimensional structures that our traditional phonology highlights. Those who study spoken language have attempted to approach it as an internal knowledge structure, rather than as a communal institution or set of conventions for coordination of activity. We also find it very difficult to avoid being deceived into seeing language in the form employed by our writing system, as letters, words and sentences. But our writing system is a further set of conventions that approximate the high-dimensional spoken language in a consistent and regularized graphical form.

Language seems to pose a problem for ecological psychology. Language is a domain where it seems very difficult to argue that linguistic structures like words are only in the environment since our intuitions are strong that words are internal, mental symbols of some kind. The claim that words and phonemes are internal symbols with an arbitrary link to meaning is very persuasive to most scientists – except perhaps those few who have a theoretical commitment to studying information in the environment. I propose a new way to think about language that I believe resolves these difficulties. For a newborn, language is clearly just part of its environment – something it hears and may have a special interest in. The child must learn to use the language, but may not need to represent it explicitly. The low-dimensional patterns of language (such as the ones we represent in our orthography) belong to the community of speakers; individual speakers merely imitate patterns they hear. Given the massive amount of variation, actual language use requires the rich dimensionality of speech spectra over time for perception to be successful, and speakers exert fine control over many subtle aspects of their performance. There is apparently no way to separate linguistic from nonlinguistic information. So speakers must store and deploy linguistic material using a relatively rich set of descriptive features (Port and Leary, 2005).

In recent years, the Distributed Language Group has blossomed, endorsing ideas that seem completely compatible with the characterization of language presented here. The term ‘distributed’ emphasizes that linguistic structures are not located in each individual, but are distributed across a population (Cowley, this issue). The DLG community (see
other papers in this volume) emphasizes that language cannot be separated from the rest of human behavior without severe distortion. The minimal case of language use is not a transcription of a sentence out of context, but two or more people in conversation on some topic. Language is just one aspect of the intense interpersonal coordination exhibited by humans. In fact, I claim no fundamental distinction can be made between linguistic conventions and other cultural conventions such as culture-specific gestures and facial expressions (Harris, 1981). Our professional linguistic definition of 'language' is narrowed to describe largely the aspects of human vocal interactions that happen to be preserved by typical alphabetical orthographies. Thus hand and body gestures and facial expressions are outside conventional linguistics – even though many expressions in language are tightly coupled with specific facial and manual gestures. One thinks of American expressions and gestures like: “I don’t know” (spoken with a shrug, raised eyebrows and the lip corners turned down), “Just a teensy-weensy bit more cake, please” (with finger and thumb demonstrating how little), “You told him what?” (with brows furrowed and head pushed forward), “Stop! Stop! Stop!” (with hands raised, palms forward, fingers splayed). Intonation is an unavoidable accompaniment of consonants and vowels, but even intonation seems to be an awkward borderline case that is studied by only a few linguists. The conventions in all these domains are acquired by very similar kinds of imitation learning: of linguistic expressions, tone of voice, gestures, idioms, constructions (Goldberg, 2006; Wray and Perkins, 2000), pronunciations (Labov et al., 2006), intonation patterns, patterns of semantic interpretation (Bolinger 1976), gestures of the hand, rhetorical constructions and body gestures like shrugs. Children each discover some way to talk appropriately and to recognize and use more and more linguistic patterns throughout their lives. Individual speakers will surely find idiosyncratic solutions to the problem of using a language. The idiosyncrasies of personal linguistic conventions, due partly to different patterns of exposure to the corpus, is one of the main reasons why linguists should focus their attention on language as distributed across a community and abandon the notion of an ‘idealized speaker-hearer’ whose cognitive representations we traditionally seek to model (Chomsky, 1965). One will not generally find linguistic generalizations represented physically within any speaker, although they can be found in a corpus of linguistic behaviors. If one wanted to study the detailed representational mechanisms of some single speaker, it might be possible, but then one is not likely to find the kind of generalizations that linguists are interested in: generalizations like the inventory of sound patterns used by most speakers, the word constructions, the syntactic generalizations – in other words, the ways that people usually talk.

Every speaker is a language-production and perception machine that is different from every other speaker. Nevertheless, it still seems intuitively persuasive to us that the code for each linguistic item should be the same from speaker to speaker. In fact every speaker’s ‘code,’ i.e., stored speech chunks with associated interpretations, will surely be different. The abstract linguistic units – phones, phonemes, words, phrases, etc. – can be best defined only at the group level and, even then, only approximately. If this sounds like nonsense to the reader, consider that Complex Adaptive Systems are now partially understood (Holland, 1995). Each human community is able to develop a culture over time which tends to continually adapt to its environment by slowly changing and
improving technologies for cooperative behavior, sustenance, defense, training of the young, etc. (Richerson and Boyd, 2005). Of course, language is one of the most important parts of a culture. From the perspective of a young person, the language they hear is just one aspect of the cultural environment they were born into.

I will first review the traditional view of speech and language, criticizing it as I go, then encourage the reader to consider how literacy might influence our linguistic intuitions. Finally I will sketch my proposal that a language is essentially a kind of social institution, something created by a community of speakers over generations and is not separately represented in the memory of each speaker.

A. Language as a mental code and its predictions

The standard idea about the structure of language, at least during the past century and perhaps for a millennium, is that it consists of discrete sound units composed into discrete words which are, in turn, composed into sentences. All mainstream linguistic theories of the 20th C, from Saussure (1916) to Chomsky (1965), are constructed on variants of this idea. Linguists have differed in how the set of phones or phonemes was to be determined and what their specific formal properties are, but practically all agreed that a small inventory of consonant and vowel tokens (less than 50 or so), best represented graphically using letters, are the basic units of psychological representation for language. It is usually assumed they are discrete, nonoverlapping (i.e., serially ordered) and abstract “speech sounds” that are independent of their neighboring context (that is, the same across contexts). Words (or morphemes) are the next larger psychological units which, in turn, are combined into realtime utterances whose basic structure consists of phrases and sentences spelled from phones or phonemes. If valid, all such theories should have many simple, testable predictions about the details of speech production and perception. Unfortunately, almost none of these predictions are supported by experimental data (Port and Leary, 2005; Lisker & Abramson, 1971; Sampson, 1977).

For example, if words are spelled from a discrete alphabet of modest size, then we should expect human speech to exhibit at least these five characteristics:

1. **Phonetic jumps.** Speech variation and historical sound changes should exhibit jumps from one sound type to a neighboring sound type, e.g., a \([t]\) might change to a \([d]\) or an \([i]\) to and \([I]\) (or in smaller steps if a larger phonetic alphabet were employed). One expects some noise to blur things, of course, but it has long been known that nothing resembling universal phonetic discreteness is found and that almost all of the important parameters of speech (e.g., vowel quality, place of articulation, voice-onset time, vocal pitch, etc.) vary continuously, both within individual speakers (depending on context) and as differences in target location between related dialects.

2. **Segment-shaped physical correlates.** There should be directly observable physical correlates of the boundaries or spaces between the segmental units separating each consonant and vowel (like the ones some may hear phenomenally when listening to speech). But it has been known since the first speech waveform displays were made in
the early 20th C that letter-sized units of speech cannot be observed in the speech pressure wave, even when frequency is plotted against time on a sheet of paper, as in a sound spectrogram. Frequently, it is very difficult, just looking at a spectrogram, even to determine how many letters it would take to transcribe a stretch of speech, much less determine the location of anything that could be called a ‘boundary’ between phones or phonemes (Joos, 1948). The so-called segments we hear phenomenally (i.e., the consonants and vowels) do not at all resemble the intervals of homogenous wave type that are clearly visible on a sound spectrogram. The natural segmentation of the speech wave corresponds to abrupt changes in articulation, e.g., vocal tract closure, changes in glottal aperture, tongue or lip gestures, etc. (Fant, 1973). But nothing has the properties we expect at a ”boundary between segments.” In fact, it seems that the acoustic or auditory transitions between neighboring steady-states are the most informative part of the auditory information.

(3) Physical invariance. It has generally been assumed that there must be acoustic invariants corresponding to the letters of phonemic or phonetic transcription (Halle, 1985), but research has failed to find acoustic patterns with those properties in most cases. If the letters of the phonetic alphabet comprise a valid psychological model, then each phoneme (or segmental speech sound) should exhibit the same acoustic pattern each time the same letter is used in a transcription (invariance property) and the physical acoustic correlates of each symbol should have the same serial position as the symbols themselves (the linearity property). These constraints simply assure that the symbols can be reliably identified, ordered and produced. If such sound units are to count as psychological ”symbols,” then they should have invariance and linearity. Of course, these properties accurately describe letters (since letters have sufficient invariance of shape that we nearly always correctly identify them), but do they describe speech? Chomsky and Miller (1963) acknowledged that the phonemic level units of languages do not satisfy the ‘invariance’ and ‘linearity’ conditions. For example, stop consonants differing in place of articulation (e.g., /b, d, g/) are differentiated from each other by acoustic cues whose form varies depending on both the preceding and following vowel (Liberman, et al., 1952; 1957; 1968; Fant, 1973) (violating invariance). And the timing cues for [+ Voice] vs. [– Voice] in, e.g., edger-etcher or slumber-slumper, are known to extend over at least the entire coda of a syllable (e.g., Klatt, 1976; Port & Leary, 2005; Hawkins & Nguyen, 2004) (thus violating linearity). Chomsky and Miller apparently took for granted that a phonetic alphabet could be provided that would exhibit the properties of the letters of their phonetic transcriptions. Apparently, they believed that the psychological phonetic units (resembling a phonetic alphabet) would surely eventually find satisfactory physical definitions that would ground their theory (see Pisoni, 1997). But data have continued to show the impossibility of satisfying linearity or invariance of phonetic symbols for describing speech in English.

Our perceptual experience of speech resembles articulation much more closely than acoustics (Fowler, 1986; Wilson, 2004), but the context-sensitive acoustic cues are still essential to provide spectro-temporal templates in order to identify these complex auditory patterns. Thus there may be nothing in the physical sound (as opposed to the articulation) to justify or explain our hearing the same consonant in Dee as in do. Still
the physical sound alone is quite sufficient to support perception of consonants and vowels that are invariant across contexts.

(4) **Speech timing based on letter-sized segments.** If words are fully specified by their transcription, then phones and phonemes (i.e., letter-like representations of speech) should support timing relations that are expressible simply in terms of the number and serial order of segments. (Thus, for example, a language could use both /ata/ and /atta/ to exhibit a “consonant length” distinction.) Yet, many other subtleties of speech timing are exploited in languages of the world (e.g., Lisker and Abramson, 1971; Klatt, 1976; Hawkins and Nguyen, 2004; Port and Leary, 2005). Studies of speech acoustics and speech perception converge on the view that speakers produce consistent subtle variations in the timing of speech gestures, and listeners make use of much temporal detail for speech perception (Klatt, 1976; van Gelder and Port, 1995). So, from the viewpoint of what speakers listen to and what they control in speech production, any alphabetical representation will be so impoverished that it cannot support either real-life speech perception nor expressive and fluent speech production.

(5) **Abstract linguistic memory.** The hypothesis that words are represented in memory using an abstract, segmented alphabet predicts that memories for linguistic material should be speaker-independent and lack any timing detail. Thus, when a speaker (even an illiterate one) listens to an auditorily presented wordlist, the speech should be remembered in an abstract linguistic code using something that is roughly isomorphic with a transcription in a cognitive phonetic alphabet. Of course, speaker-specific information might be stored as well, but speaker idiosyncrasies cannot be part of the word representation itself. Thus, speaker properties could only have an association with the linguistic representation of the words. But in recognition memory experiments using lists of spoken words, it has been shown that listeners are better at recognizing that a word is repeated when the voice is identical in both occurrences than if the voice is different (e.g., Goldinger, 1996; Palmeri et al., 1993). One possible interpretation is that a fairly detailed auditory representation is routinely stored, so with more features that match between the new utterance and the detailed record of the earlier one, the same-voice repetition is more likely to be noticed as a repetition. But also, on the traditional view, we should be able to make it more difficult to associate the voice with the word if the associations were confused by the use of many different voices for the list instead of just two. If participants tried to remember the voice to improve recognition of the word, then they should be able to do that much better with only two voices than with 20. Indeed, to whatever degree the participants might be guessing their answer, the 2-voice condition should exhibit higher scores than the 20-voice condition. Yet the performance benefit for the same voice when words are repeated is the same no matter how many voices were used. This surprising result implies that speakers must routinely store auditory speech material in some rich and detailed form that includes speaker-specific properties. This result is compatible with something along the lines of a rich ‘exemplar’ memory, as many experiments on vision have found evidence for (Nosofsky, 1986).

(6) **Speech errors.** Another prediction is that when speech errors occur, they should show evidence that the segment (or a vector of segmental distinctive features) is
the unit that leads to the error. In fact, the speech error distributions should resemble the
distribution of typing errors, since typing employs discrete context-invariant serial tokens
as well. Thus, we should expect frequent errors like (a) reversals of immediately adjacent
segments (e.g., /tɛnt/ for intended /tent/) but fewer cases of switches between distant
phonemes, as in ‘Mardon me, padam’ for ‘Pardon me, madam.’ But the latter is a far
more common type of error that switching adjacent segments. We should find (b) the
complete replacement of one phone by another phone and (c) no evidence that
phonotactic constraints of the language influence the errors (since it is the segments that
are supposedly the confusable units) and (d) no evidence of attempts by speakers to
simultaneously produce several competing gesture components of the phones. In fact,
order reversals of adjacent phones are extremely rare and there is recent evidence from
speech kinematics that speakers often attempt to produce two different gestures at the
same time and that the independent units of speech are time-extended gestures rather than
segments (Shattuck-Hufnagel and Klatt, 1979; Goldstein, et al, 2007). Since the
traditional speech error databases relied on auditory (impressionistic) phonetic
transcription, so many details of articulation were overlooked. Thus, speech errors do not
appear to provide much support for segments as the units of speech production as
opposed to continuous-time phonetic gestures.

Notice that the evidence relevant to these 6 predictions in each case supports greater
richness (relative to an alphabet-based notation) of whatever representational methods
speakers employ to remember language. Speech is not stored in memory in an abstract
code, it is stored using as much detail and richness as the speaker can achieve. Rich
temporal and spectral detail is apparently what is required to do human-like speech
perception. In parallel, for talkers to achieve expressive natural performance, fine-
grained continuous control of the speaker’s vocal apparatus is also demanded.

Many of these arguments have been elaborated elsewhere (Port et al, 1995; Port and
Leary 2005: Port 2007, 2009). But in all six of these expectations, the data fail to support
the traditional view. Indeed, one fails to find any cognitive role for abstract, segment-
sized units that are invariant across speakers, across speaking rates or across variation in
neighboring contexts. The likelihood that speech relies on a relatively rich and detailed
representational scheme pushes all conventional linguistic representations based on letter-
like units right out of the picture. Consonant and vowel-sized units appear to be
irrelevant to both speech production and perception. They are only relevant for
representing some aspects of spoken language in an efficient graphical form.

Where did we go wrong?
Yet, of course, this conclusion is very difficult to accept for modern psychologists and
linguists. We have all resisted drawing this inference for well over a half century now.
Why is the inference that there is no real-time psychological role for phones or phonemes
so difficult to entertain? One reason that has been suggested is: why would speakers
ignore the availability of an efficient, low bit-rate code like the alphabet when it is so
obvious and readily available? The answer to this is, of course, that an alphabet-like
representation is not readily available – not until literacy skills are developed (Rayner et
al, 2001; Ziegler and Goswami, 2005). It may seem completely transparent and obvious
to us that *tap* has 3 `speech sounds’ and the *trap* has 4, or that the `same vowel sound’ occurs in *bat* as in *ban*, or that *spot* is pronounced with [p], not a [b]. However this way of listening to and thinking about speech sounds is not obvious at all to most six-year-olds, or to adults with no alphabet training (Morais, et al, 1979; Read et al, 1986). The most important reason we have not drawn the inference that segments are irrelevant is, in my view, simply that our intuitions about language overwhelmingly testify that a word `consists of’ letter-like parts. Words are not composed from letter-like units – neither acoustically, articulatorily, nor in memory.

Let us examine a transcription more carefully. A written word like *tomato* is usually pronounced by me using the IPA phonetic alphabet as approximately:

\[ t^{h}ame^{o}i^{o} \].

That is, the initial [t] has an aspiration interval after the /t/ closure (suggested by the [h] superscript); the second and third vowel are diphthongs (defined by their movement in the vowel space as suggested by the superscript vowel symbols) and the second orthographic T is almost invariably pronounced as a tap or flap (with a flick of the tongue tip similar to a Spanish /r/) before an unstressed vowel (spelled phonetically as [ʃ] here) in my dialect. Notice that this transcription using superscripts preserves the same 6 segments as the orthographic spelling, but it expands the symbol inventory with additional letters so that single letters (with diacritic superscripts) can multiply articulatory states and articulatory motions. This way it suggests some of the dynamical gestures speakers make when pronouncing the word. Although many different pronunciations are possible for this orthographic word, they can all be approximated (we hope) by using some configuration of an expanded phonetic alphabet. Both the orthographic representation and the technical phonetic alphabet representation seem highly intuitive and at least approximately correct. But, since the evidence in the previous section strongly suggests speakers do not use such a representation, could there be something that biases us to insist on using letters as the basic components of words? The answer is yes.

It may be difficult for us to recall, but every person reading this page spent several hours a week for many years learning to read and to refine their reading skills. It is surely naïve to imagine that all this focused mental effort using an alphabet over several decades would have no consequences for our intuitions about the nature of language. Yet we linguists have never paid much attention to the possibility of biased intuitions in our interpretations of speech. It seems likely we tended to overlook this potential problem due to our lifelong training using the alphabet in our orthography. Only a few linguists have considered that our orthographic alphabet skills might, in fact, tend to dominate our linguistic understanding of speech (see Faber, 1992; Linell, 2005; Öhman, 2002). Chomsky, on the other hand, insisted that we should trust our intuitions completely on matters of grammar and phonology and that doing linguistics is primarily a matter of explicating our intuitions about linguistic structure and finding formal notations for them (Chomsky,1965). It is very easy for us to think about language using an alphabet and it is very difficult to think about the sounds of speech in any terms other than alphabetical ones. But that does not mean the alphabet is the real structure.
But alphabetical writing is a technology which achieved roughly its modern form only 3000 years ago. The earliest Greek alphabet might have been created by a Greek with Phoenician education who wanted to apply Phoenician writing techniques to Greek (or perhaps by an educated Phoenician who also spoke Greek). The Phoenician alphabet itself was the culmination of 4-5 thousand years of earlier record keeping technologies in the middle east that were gradually getting easy enough to learn and convenient enough to use (Fischer, 2005). One major consequence of the development of literacy in the middle east was the growth of the institution of schooling for teaching literacy to children. Schools have been a gradually increasing part of life in literate nations ever since (Olson, 1994). Parents in some communities begin teaching literacy to children as young as 2 by playing with alphabet blocks.

Alphabetical writing is certainly very useful, but letters are artifacts. It is difficult for children to learn to interpret letter sequences as syllables and to write down syllables as letter sequences, so we start teaching our children as young as possible. Of course, all of us who have had years of reading experience find the interpretation of continuous speech as a sequence of letter-like units to be trivially easy and completely natural. One might even ask: how could we expect to hear speech in any other way given all the practice we have had with letters?

Does this mean that phones and phonemes are illusions? Not at all. It just means they are not necessary participants in the realtime processing of language, either on input or output. What is illusory here is the conviction that phones and phonemes must play a functional role in any real-time cognition involving language simply because our conscious thoughts about language are vividly letter-based. Phoneme-like sound patterns do, at least, play a functional role for the community and possibly for individual speakers. They are regularities or symmetries in the speech corpus of a community which speakers learn to reproduce in their own speech (Newport and Aslin, 2004). How these approximations to a low bit-rate code provide benefit to the community and how they benefit individual speakers are important questions requiring further research (although there are various attempts in print to spell out the benefits of phonology and grammar, e.g., Studdert-Kennedy, 2003).

B. Language as Culture

My proposal is that we need to revise our thinking about language from the bottom up. Our mistake has been to assume that to study language is to study a form of “knowledge,” to study the representations in memory that make speech possible. Instead, a language is a kind of social institution, that is, a partially structured system of conventions created by a community of speakers and refined over generations. It is a technology developed by a community for coordination of behavior. Thus it is inherently distributed over space and time, and represented differently in its real-time behavioral details in the brain of each speaker. Other institutional technologies include religious practices, an orthographic system, an educational system and a system of community government, plus food producing technologies like hunting, farming, fishing etc. with all of their accumulated knowledge, tools, manual skills and social conventions. All of these
systems and more comprise the *culture* of a community. All exhibit ratchet-like accumulation of knowledge and skills over generations that facilitate survival and reproduction in the ecological niche of the community (Tomasello, 1999; Richerson & Boyd, 2005).

A human community can be viewed as a **complex adaptive system** (Holland, 1995). Such a system supports the emergence of patterns and structures (i.e., symmetries of pattern) of many kinds on many spatial and temporal scales. We should think of a language, then, as one part of the culture of a community (Beckner, et al, 2009). It includes a socially maintained inventory of continuous-time speech fragments that facilitate coordinated action (as suggested by Arbib, 2005). Some of the fragments of English are [bə, də, gə] (as in *between, deposit, garage*), [bəl] (*buy*), [baːl siʃə omarə] (*Byebye, see you tomorrow*), etc. Note that minimal segments must have some extension in continuous time. The concept of discrete-time tokens with no temporal extent (exhibited by letters and supposedly exhibited by phones and phonemes) is a technical invention, a cognitive blend (Fauconnier and Turner, 2003) that merges aspects of speech sound with aspects of letters of the alphabet. Although I have described the English fragments here using two different alphabets (phonetic and orthographic), no alphabetical description is adequate. The relevant fragments are actually learned as synchronized continuous-time patterns of acoustics, articulation and somatosensation by speakers. These patterns in the speech of any particular community tend to exhibit many symmetries. That is, some articulatory and auditory properties of one fragment resemble properties of other fragments (i.e., *park* resembles *ark* and *Parr* which resemble *are* and *Pa*, etc.). These overlaps often create the impression (if one were motivated to develop the shortest possible list of fragments) that each linguistic usage is “composed” from a set of nested units that linguists loosely call phones or phonemes. But, in fact, the set of components can never be fixed and can never be timeless (like phones and phonemes). Speakers have much richer memories for language than are implied by any alphabetical descriptions (whether orthographic or phonetic). And speakers have production skills that allow control over an almost unlimited set of subtle articulatory properties. Since speakers remember considerable rich detail about speech, alphabets are clearly inadequate for just about any purpose (except for writing, of course, which is very useful if you can get the training).

*Homo sapiens* seems to have found ways to allow human communities to specialize and to develop complex cultures that create a vast range of different technologies appropriate to the environmental niche of each community. Clearly, intense human sociability contributed to the development of cultures and language (Tomasello, 1999; Frith & Frith, 2010; Herrmann, et al., 2007) along with the ability to learn statistical regularities after presentation of complex patterns (Newport and Aslin, 2004). Human communities are complex systems capable of adapting their culture over generations (e.g., Beckner, et al, 2009; Smith et al, 2003; Smith and Kirby, 2008; Port, 2007; Hruschka et al., 2009; Schoenemann, 1999; Richerson and Boyd, 2005). These social systems emerge in communities of many agents who interact with each other in complex ways.

Does this story imply then that, contrary to the standard view in linguistics (Chomsky, 1965; Pinker, 1994), humans have no specialized physiological or neural adaptations for
language? After all, I am claiming that actual languages are emergent structures created by human communities not by individual human brains. But surely, in my view, there are many evolutionary adaptations that were necessary for the rapid acquisition and fluent use of language. However those specializations should not be expected to look specifically linguistic. That is, they will not include any specific phonemes or distinctive phonetic features nor will they include grammatical parts of speech nor specifications for what constitutes a sentence. Instead, we should expect the specializations to relate to the perceptual and motor skills necessary for speech and to our proclivity to assign almost any aspect of our experience to one or another category depending on the conventions of our community. They are probably also related to the social skills of human infants and adults, such as the ability of human infants to direct attention to some object or event also attended to by their caregiver – what Tomasello (1999) calls “joint intentionality”. These behavioral, neural, physiological and anatomical specializations were presumably selected for over the past half million years or so and give us brains and bodies suitable for the emergence of practical and effective human languages. Surely, the languages that emerge among Homo sapiens can only be those that are compatible with the neural, physical and physiological properties exhibited by modern humans (as argued by Christiansen and Chater, 2008).

Some cultural categories: Toward a new linguistics.

One important aspect of language as an inventory of conventions for behavior coordination is categorization. It is important not to confuse categories with formal tokens, as linguists and others have done for long time. Formal tokens, like the letters [a-z, A-Z] and digits [0-9], are discretely different from each other and can be assumed to be recognized and produced with near-perfect accuracy (see Haugeland, 1985). The assumption that spoken language is constructed from such tokens has led to the speculation that spoken language might be a formal system (Chomsky, 1965; Chomsky and Miller, 1963). A category is a psychological grouping that is learned from one’s community – a set of things considered the same by the community – no matter what the reason may be for calling them the same. (I avoid using the term ‘symbol’ since it is used in so many different ways as to be very confusing in this discussion.) Cultural conventions chop up the world into parts, many of which have names or verbal labels. A big part of learning one’s culture and language is learning the things and events that are recognized by the society along with their simple or complex verbal labels (Heft, 2007; Hodges and Baron, 2007; Sloutsky, 2003; Evans and Levinson, 2009).

Apparently there are many cultural categories shaping our behavior that do not have words assigned to them by the popular culture, but may have been explicated by some modern academic subculture. Examples of these are the various categories that linguists have discovered and found to be relevant to linguistic behavior in specific languages (e.g., phoneme, ±voice, noun, verb, conjunction, mora, etc.). Academic linguists have sometimes proposed (universal) formal-token status for these linguistic categories (Chomsky, 1965; Chomsky and Halle, 1968), but it is far more plausible that each culture uses somewhat different grammatical and phonological categories (Croft, 2000; Evans and Levinson, 2009). It is important to keep in mind that, as socially supported
categories, the specifications will vary from language to language, and, in fine detail, from speaker to speaker (Bybee, 2007). The speech products of people may fall into various linguistic categories, but the speaker usually does not have explicit knowledge of what any of the linguistic categories are (unless the speaker has the benefit of literacy). The categories of everyday life, e.g., the names of animals, plants, cultural artifacts, social roles, etc., are conventions whose transmission depends on skilled language use and on social interaction. Categories may be defined by a rule (e.g., definition of a square), by physical or functional similarity (e.g., wheel, chair, stove, etc.) or by any other means, including a simple list of partially arbitrary members (Smith & Medin, 1981; Glushko et al., 2008; Murphy, 2002). The regularities of a language are exhibited in the corpus of speech that each child is exposed to (Tomasello, 1999), and the speaker adapts to the language regularities in some idiosyncratic way depending on the personal history of exposure to languages, dialects and various speakers.

Linguistic Categories

Thus far I have tried to show that the representational abilities of speakers should be thought of as including distributions of rich multidimensional descriptions of a large set of heard utterances along with many abstracted multidimensional templates as well. These representations do not resemble their abstract, context-free, static, written form. Fragments of these concrete memories, however, are assigned to categories in many ways by speakers of the language, related to what we call semantics (e.g., singular/plural, etc.), phonology (e.g., ±voice, /b, d, g/) and syntax (e.g., grammatical categories). The traditional assumption that phonological patterns could be adequately represented in memory merely in terms of letter-like physical tokens, discretely ordered in time, was in part a description of some gross properties of the language but, at the same time, it was also a projection of the properties of orthographic technology onto our intuitions about human psychological processes.

A special trait of human communities is the tendency of cultures to categorize their world. One of the most important technologies created by human communities is spoken language. A language is shaped over generations with respect to the categorization provided by its “lexicon” and “grammar” as well as to the range of speech gestures and sound categories the speakers of a community employ (Hock and Joseph, 1996). But there will always be inherent uncertainty about exactly what the patterns are. The uncertainty stems from variation in the corpus to which each speaker is exposed and from such basic issues as exactly what community is under study and whose speech is representative of it. Of course, there are speech chunks at a large range of sizes, from syllable pronunciations to speech formulas to entire spoken “texts” (such as prayers or epics) (Wray and Perkins, 2000). But there are no universal criteria for how to distinguish word-sized pieces from common phrases, from idioms or from “styles of talking” about things. Our orthography has made partly arbitrary decisions about what counts as a word (and thus deserving separation from other words by a space) vs. a phrase or sentence but each speaker must find their own way to store memories of linguistic fragments and generate new utterances.
What are some of the structures of language that are created by communities of speakers? Although it is clearly an illustration of literacy bias, we tend to think, first, of a "lexicon," a list of word-like chunks. But there are also "grammatical regularities," illustrated by the structure of simple sentences or case markings on nouns and tense markings on verbs. The sentence is a fundamental concept for written language, but probably plays a role in the spoken language primarily only for speakers with extensive literacy training, such as professors, lawyers, preachers, newscasters, etc. Languages always seem to employ a restricted inventory of gestures and sounds. A problem for linguistics is that our literacy bias leads us to insist that syllables are composed from so-called "speech sounds," i.e., letter-sized units, the consonants and vowels. These hypothetical "sounds" (although, of course, they are not really physical sounds at all) are supposed to be just like letters except that they occur in real time as sound patterns. But it has been known for 50 years that speech sound is not divisible into countable units that line up with letters in any consistent way (Joos, 1948; Fant, 1973). The phonology of a language provides a continuous series of dynamic gestures which are partly serial and partly simultaneous (Browman and Goldstein, 1992; Goldstein and Fowler, 2003).

The importance of the notion of "restricted inventory" is that words in any specific language or dialect tend to resemble each other quite a bit such that, as noted above, for most speech fragments, other fragments can be found in the corpus that are very similar. To illustrate this, notice that sometimes pairs of words differ from each other in very similar ways. For example, the distinctions between the 6 English words in Set 1 below are fairly similar to the differences between the 6 words of Set 2. (An asterisk marks the forms that are not actual words in my dialect.) The orthographic notation here should be interpreted as standing for or pointing out various lexical categories of continuous speech gestures or their continuous-time acoustic description. As noted above, some acoustic differences will be the same between ban-pan and Bill-pill, although the formant transitions from the stops into the vowel (known to be essential bearers of information about stop place of articulation) will be different in the two lists. But the third set, where the stops occur in a different position in the syllable, presents more of a problem. The voicing distinction between [b] and [p], etc. is manifested quite differently between ban-pan (in initial position) as opposed to lab-lap (in syllable final position) since the vowels in the words (i.e., lexical categories) ending in [b,d,g] in Set 3 are quite a bit longer than the vowels in the words ending in [p,t,k] in Set 3 and the aspiration that distinguished ban from pan occurs nowhere.

<table>
<thead>
<tr>
<th>ban</th>
<th>pan</th>
<th>Bill</th>
<th>pill</th>
<th>lab</th>
<th>lap</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dan</td>
<td>tan</td>
<td>dill</td>
<td>till</td>
<td>lad</td>
<td>*lat</td>
</tr>
<tr>
<td>*gan</td>
<td>can</td>
<td>gill</td>
<td>kill</td>
<td>lag</td>
<td>lack</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Set 1</td>
<td>Set 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

But English orthography implicitly claims the same consonantal "speech sounds" are used in all 3 sets of words, even though research on the physical speech signals (as argued above) shows that speakers of the language actually store and make use of some detailed auditory form for all the words. Presumably languages gain some benefit from favoring word specifications involving this kind of partial similarity between syllable-initial and syllable-final contexts, since many languages have similar sets of consonants in syllable-initial and syllable-final position.
This is an example of symmetrical patterns created by a community of speakers. These words show many symmetries since, for example, the "voicing property" seems to be the same for all the words in the same column in the sets above and the place of articulation feature is nearly identical for all stops in the same row. Of course, this is true only sometimes since (a) the contrast between /t/ and /d/ is largely neutralized before an unstressed vowel, as in budding/butting (although not completely neutralized in some dialects, cf. Fox and Terbeek, 1977; Port and Crawford, 1989), and (b) the difference between /b,d,g/ and /p,t,k/ is not found after an /s/ (cf. spot, stow, Scot, but words like sbot, sdow, skot do not occur in English) although the claim of "identity" across contexts may be approximately true in many cases. So the very simple and discrete regularities and the neat serial order suggested by the alphabetical notation in the tables above apply only to letters or other graphic tokens on paper and not to the physical sound of speech. Actual phonologies are only approximately discrete and require far more degrees of freedom than letters do.

In this section, some arguments for viewing spoken language as a rough inventory of conventional speech patterns have been made. The apparent letter-like component parts of words are only approximately the same from context to context. Thus letters cannot be the actual components speakers rely on. They are simply categories of different items that are thought of as the same even though the degree of identity is highly variable. It is our expectation, as literate observers of language, that spoken language must employ a small alphabet just like the written language does. This has biased linguists to imagine that all those little details (e.g., the formant transitions, the small changes in vowel quality, the small differences in timing, etc.) must not really matter. We linguists thought: "Speakers can surely ignore all that messy detail and still produce and understand language. After all, the written language ignores those details. So we should be able to treat spoken language just like other formal systems." But speakers cannot ignore all that small stuff. We have been deceiving ourselves into thinking we could for over a century now. It is time to look at the data with clear eyes and then abandon all low-bitrate formal models for spoken language processing if we seek to understand human speech as it is produced and perceived. It is as at the communal level, in the social institution, where the low-dimensional description of language works – at least approximately.

Literacy.

Looking at human history in the long view, human communities probably first developed language roughly 200-150 thousand years ago (Mithen, 2006). The primary evidence is that this is when evidence of cultural flowering begins. It is when tools made of bone and human campsites began to show evidence of continuous cultural change and specializations for particular ecological niches. Improved language skills would have greatly facilitated the cumulative enrichment of cultures by a "ratchet effect" (Tomasello, 1999). Then much more recently, certain human communities developed a practical and easily taught alphabetical writing system (about 3 thousand years ago). The approximately low-dimensional statistical patterns in human speech, as described in the previous section, can be mapped reasonably well onto a short list of alphabet tokens.
Thus useful texts could be created requiring a minimum amount of learning of arbitrary relationships. Over the past 3k years, there has been some progress in making teaching literacy more effective, but the basic methods for teaching reading and writing have hardly changed: memorize the suggestive names for letters, then practice reading and writing short common words until the child `gets the hang of it.’ The alphabet literacy that we all share trains us to experience speech as a sequence of letters from a very small, fixed alphabet, and to consist of discrete words and sentences as well. The fit between the alphabet patterns and the real phonology of a language is good enough that with sufficient training, we have been happy to overlook all the problems (since we learned to just `get over’ the inconsistencies and the arbitrariness of spelling and writing while still very young). In the past century, formal mathematical grammars have taken the properties of orthographic letters (discreteness, serial order, etc.) and generalized them to create rich mathematical systems (Harrison, 1978) that led to the development of computers and to speculation that language might have a formal symbolic representation in the human brain (Saussure, 1916; Chomsky, 1965).

Any community provides its children with exposure to a fraction of the linguistic culture (somewhat different fractions for each speaker, of course, and for each speaker at any particular point in time). The linguistic culture of a community provides a massive set of conventional categorizations of the experience of that community. These include both unnamed categories and ones that have a lexical name or a description that requires only one or a few words to describe it. Young speakers learn `speech chunks’ by rote example at first (see Grossberg, 2000, 2003 for an explicit model that chunks speech) but gradually come to categorize speech the way their family members and neighbors do (see Werker and Tees, 1984 for results on the earliest stages of phonological acquisition) and learn to control their own body to produce speech that conforms to the conventions of their community. None of this requires that individuals produce or perceive speech using identical descriptors nor that they have any conscious knowledge of the categories they rely on.

How could it be that a speaker does not need to `know the language’ in order to use it? The key idea is that it is the community which undergoes self-organization and creates the language in historical time, not individual agents. The argument is analogous to the famous case of termites that follow simple behavioral biases about where to deposit little wads of dirt mixed with a pheromone. These biases happen to result in periodically spaced columns of a fixed height with a ceiling created above to support the next layer of columns (Kugler and Turvey, 1987). The termites themselves have no idea what they are doing. They are simply behaving instinctually, under guidance by genes as to where to deposit their wads of dirt. Yet regular and periodic physical structures are created for the benefit of the colony by animals with no representations of either periodic columns or layers. For the case of language as well, the community has become the relevant complex adaptive system, not the individual. (Of course, an individual agent may also be a complex adaptive system, but no individual human actually creates a conventional language.) The group creates structures (e.g., lexicons, phonologies, rituals, technologies, etc.) as conventions. The agents in the system are imitators as well as users of the community speech conventions. It is a serious mistake to insist that these
community structures could only exist if they are part of the psychological constitution of the agents.

So it seems highly likely that agents (i.e., speakers) differ from each other in the details of their representations (since there is nothing that can enforce identity). Still speakers in the same community will behave in ways that generally accord with the conventions followed by others. What the speaker must learn are appropriate behaviors, both perceptual and motor. The behaviors are, of course, tied to the micro-circuitry of their nervous system, their vocal-tract anatomy and their auditory and linguistic experience. The ```linguistic units``` of the community language (such as the ones we formalize crudely in English orthography and in linguists’ grammars) are not something native speakers have any use for in order to speak correctly.

**What is predicted if language is a social institution?**

If this story is on the right track and linguistic structures are essentially communal not personal, then many predictions seem to follow. The theory predicts at least the following. I believe all of them are supported by evidence, but I will point here to just one or two references in support of each.

1. **Variability between speakers.** Since linguistic categories have conventional specifications, individual agents should find a wide variety of alternative ways to implement the linguistic conventions in various situations. Small differences in pronunciation or intended meaning or in the use of grammatical conventions, since they appear in a rich realtime communicative context, should be easily tolerable. So speaker-to-speaker variation should be large if we look at the details (e.g., Docherty and Foulkes, 2000; Bybee, 2007; Labov et al., 2006). This is predicted because categories demand only approximate sameness while speakers have continuous control over many parameters of their productions.

2. **Within-speaker variation.** Any given speaker may implement the conventions differently depending on the linguistic context and the speaker’s evaluation of the communicative needs of the moment, such as the prevailing signal-noise ratio at each moment of speech (e.g., Lindblom, 1990). They may also gradually change their target pronunciations along some parameters during their lifetime for various reasons (Harrington, 2006; Sancier and Fowler, 1997). There is nothing to prevent small random variations or to deter gradual shifts in pronunciation targets in ways that reflect the speaker’s social situation at the moment.

3. **Temporal patterns.** Timing patterns should include many that cannot be well described using serial letter-like units. Durational ratios and periodic patterns, for example, should be possible, since these require some form of measurement of continuous time, and not mere counting of letter-like segments (e.g, Klatt, 1976; Port and Leary, 2005; Port, Dalby and O’Dell, 1987; Port, 2003). This is predicted since speech memory relies on representations that approximate continuous-time (not discrete-time) representations.
4. **Indexical information in speech memory.** Linguistic memory should store the speaker’s idiosyncratic voice information, rather than be abstracted away from the specific utterance (Palmeri, et al., 1993; Pisoni, 1997). This is predicted because auditory representations for words do not have invariant acoustic specifications but are hypothesized to be more concrete and rich in detail than linguists usually imagine. And the word-specifying information cannot be somehow factored out from the rest of the auditory properties of speech.

5. **Dialect variation and change over time.** Social groups that are distributed geographically tend to create subnetworks exhibiting greater or less intense intercommunication. These groups should exhibit slow change of pronunciation and other usages over time (e.g., Hock and Joseph, 1996; Labov, 2006; Bybee, 2007). This is predicted because linguistic features are simply conventional aspects of the culture of a community. Like all other cultural traits, they should exhibit gradual change over time. Variability is tolerated and small changes along various continua may be imitated by others depending on the details of each speaker’s communication network and the relative prestige of talkers. People talk primarily with people in a restricted group, so consistency at the local scale is what is most important for efficient language use.

6. **Computer speech recognition will not succeed by trying to recover discrete-time letter-like segments.** Speech recognition engineers should find it difficult to try to identify consonant and vowel segments as a step toward recognizing larger chunks of speech like words and phrases (e.g., Jelinek, 1997; Huckvale, 1997). Words as recognize and produced are not composed from simple segments or segmental features. Instead, they are defined by overlapping continuous gestures and acoustic trajectories. Static letter-like segments should be of very limited use for speech recognition (even though native speakers can be trained to use letters for reading and writing).

There seems to be plenty of observational and experimental support for every one of these predictions. On the other hand, we saw earlier that the predictions of the abstract, segment-based notion of speech found almost no supporting evidence.

C. **Conclusions.**

It seems that the radical story I am endorsing here is compatible with a great deal of the experimental data dealing with language and its psychological basis. It is also compatible with the ideas of the Distributed Language Group since they emphasize the tight social embedding of linguistic behavior as well as the vast differences between written and spoken language. Since we linguists and psychologists are all literate, we find it difficult to see that our thoughts and intuitions about language have been profoundly shaped by the reading and writing skills we worked so diligently to achieve in our youth. Almost all linguists have followed Saussure (1916) in claiming to be studying spoken language not written language. But the fact is that almost all modern linguists, like Saussure, never really escaped from letter-based characterizations of language. Audio recordings and sound spectrograms are rarely found in the linguistics classroom nor in most linguistic...
research. When we think of ‘words,’ ‘speech sounds’ and ‘sentences’ in our descriptions of language, we are importing the conventions of our writing system and trying to use them uncritically as hypotheses about psychological representations. Confusion about the distinction between the written language and the spoken language creates difficulties for many types of research (e.g., see Love’s, 2004, criticisms of Clark, 2001).

What I have tried to show here is that the neat code-like units of our normal conscious thoughts about language — that is, the letter-like ‘speech sounds,’ the discrete words, morphemes, prepositional phrases, full sentences, etc. — play at most a very tiny, specialized role in real-time conversational performance. The kind of speech patterns linguists are interested in only exist at the level of the community and reflect a kind of social institution that appears as the cultural environment for each child. Linguistics and the psychology of language should abandon attempts to describe the representations of abstract linguistic units stored in people’s heads — even though this has been the goal of these disciplines for at least a century. The patterns of a language that linguists are interested in simply do not exist in the form of abstract, formal, low-bitrate structures common across all speakers. Instead, our every utterance is a creative behavioral response to our experience, shaped by every possible facet of our lives, but also generally compatible with the conventions observable in the speech of others in our community. A language is, first of all, the set of linguistic conventions about speech shared by some community. For several millennia some communities have also had a system of orthographic conventions based partly on the phonetic value of letters. Of course, the letters and the orthography are part of a language too, but letters differ in profound ways from continuous speech and have seriously confused our attempts for the past century to understand human linguistic capabilities. The spoken language does not depend on letters of any sort.

References


