1

Enlarging the scope of phonologization*

LARRY M. HYMAN

"… the original cause for the emergence of all alternants is always purely anthropophonic"

Baudouin de Courtenay (1895 [1972a: 184])

1.1 Introduction

It is hard to remember a time, if ever, when phonologists were not interested in the relation between synchrony and diachrony. From the very founding of the discipline, a constant, if not always central issue has been the question of how phonology comes into being. As can be seen in the above quotation from Baudouin de Courtenay, the strategy has usually been to derive phonological structure from phonetic substance. The following list of movements dating from the early generative period provides a partial phonological backdrop of the wide-ranging views and interest in the relation between synchrony and diachrony, on the one hand, and phonetics and phonology, on the other:

(1) a. classical generative phonology (Chomsky and Halle 1968)
b. diachronic generative phonology (Kiparsky 1965, 1968; King 1969)
c. natural phonology (Stampe 1972, Donegan and Stampe 1979)
d. natural generative phonology (Vennemann 1972a, b, 1974; Hooper 1976a)

* Earlier versions of this chapter were presented at the Symposium on Phonologization at the University of Chicago, the UC Berkeley, the Laboratoire Dynamique du Langage (Lyon), MIT, SOAS, and the University of Toronto. I would like to thank the audiences there, and especially my colleagues, Andrew Garrett, Sharon Inkelas, and Keith Johnson, for their input and helpful discussions of the concepts in this chapter. Thanks also to Paul Newman and Russell Schuh for discussions on Chadic.

e. variation and sound change in progress (Labov 1971; Labov et al. 1972)

f. phonetic explanations of phonological patterning and sound change (Ohala 1974, 1981; Thurgood and Javkin 1975; Hombert, Ohala and Ewan 1979)

g. intrinsic vs. extrinsic variations in speech (Wang and Fillmore 1961; Chen 1970; Mohr 1971)

For some of the above scholars the discovery of phonetic and/or diachronic motivations of recurrent phonological structures entailed the rejection of some or all of the basic tenets of classical generative phonology, as represented by Chomsky and Halle's (1968) *Sound Pattern of English* (*SPE*). As a generative phonologist, I found myself conflicted between a commitment to the structuralist approach to phonology, as reflected in the Prague School (e.g. Trubetzkoy 1939; Martinet 1960) and in *SPE*, and a desire to 'explain' this structure in terms of its phonetic and historical underpinnings. The resolution I opted for was to focus on the process of *phonologization*, which is concerned not only with these underpinnings, but also with what happens to phonetic properties once they become phonological. Thus, although resembling Jakobson's (1931) term *phonologization* (*Phonologisierung*), which is better translated as *phonemicization* (whereby an already phonological property changes from allophonic to phonemic), I intended the term to refer to the change of a phonetic property into a phonological one. Definitions of phonologization from this period include the following:

> A universal phonetic tendency is said to become 'phonologized' when language-specific reference must be made to it, as in a phonological rule. (Hyman 1972: 170)

> *phonologization*, whereby a phonetic process becomes phonological.… (Hyman 1975: 171)

> … what begins as an intrinsic byproduct of something, predicted by universal phonetic principles, ends up unpredictable, and hence, extrinsic. (Hyman 1976: 408)

As opposed to Jakobson's term, which referred to the development of contrasts, my specific interest was in the development of allophony. However, as seen in the last quotation above, I explicitly referred to Wang and Fillmore's intrinsic vs. extrinsic terminology, which they identify as follows:

> … in most phonetic discussion, it is useful to distinguish those secondary cues which reflect the speech habits of a particular community from those which reflect the structure of the speech mechanism in general. The former is called extrinsic and the latter, intrinsic. (Wang and Fillmore 1961: 130)

Since a clear distinction was not always made at the time between allophonic variations which might be captured by phonological rule and language-specific phonetics, the two were often lumped together. The result is a potential ambiguity, depending on whether one makes a distinction between allophonics and language-specific phonetics and, if so, whether the latter is identified as 'phonology' or as phonetics.

I have two goals in this chapter. First, I wish to explore the above notion of phonologization further, specifically addressing the role of contrast in the phonologization process. Second, I wish to show how phonologization fits into the overall scheme of the genesis and evolution of grammar. Extending the concept of phonologization to a wider range of phonological phenomena, I shall propose that it be explicitly considered as a branch of grammaticalization or what Hopper (1987: 148) refers to as 'movements toward structure'.

1.2 Phonologization and contrast

As stated in section 1.1, discussions of phonologization have focused on intrinsic phonetic variations which tend to become extrinsic and phonological. The most transparent of these concern cases of what Cohn (1998: 30) refers to as phonetics and phonology 'doublets'. Processes such as those listed in (2) may be phonetic in one language, but phonological in another:

(2)	*process*	*subsequent developments (incl. loss of trigger)*			
	a. lengthening before voiced Cs:	/ab/	→	[a:b]	(> a:p)
	b. palatalization:	/ki/	→	[k^{y}i]	(> či, ši, tsi, si)
	c. high vowel frication:	/ku/	→	[k^{h}u]	(> k^{x}u, k^{f}u, p^{f}u, fu)
	d. anticipatory nasalization:	/an/	→	[ãn]	(> $ã^{N}$, ã:, ã)
	e. umlaut, metaphony:	/aCi/	→	[æCi]	(> ɛCi, ɛCə, ɛC)
	f. tonogenesis from coda:	/aʔ/	→	[áʔ]	(> á)
	g. tonogenesis from phonation:	/a̰^{ʔ}/	→	[à^{ʔ}]	(> à)
	h. tonal bifurcation from onset:	/bá/	→	[bă]	(>pă)

In order for there to be a phenomenon of phonologization and such doublets, it is of course necessary to recognize a difference between phonetics and phonology. Some of the characterizations of phonetics vs. phonology by those who assume a difference (e.g. Cohn 1998, 2007; Keating 1996; Keyser and Stevens 2001; Kingston 2007; Pierrehumbert 1990; Stevens and Keyser 1989, etc.) are presented in (3).

(3)	*phonetics*		*phonology*
	gradient	>	categorical
	continuous	>	discrete, quantal
	quantitative	>	qualitative
	physical	>	symbolic
	analog	>	digital
	semantic	>	syntactic

As seen, phonetics and phonology can have very different properties. As one proponent of the distinction puts it, 'The relationship of phonology to phonetics is

profoundly affected by the fact that it involves disparate representations.' (Pierrehumbert 1990: 378). While most of the above descriptors are well-known and straightforward, others are intended as analogies, e.g. analog vs. digital, semantic vs. syntactic (Pierrehumbert 1990). It should be noted that the phonetics–phonology relationship is not one of universal vs. language-specific, since much of phonetics is itself language-specific (cf. below).

Two diagnostics were proposed for determining that phonologization has occurred: (i) A phonetic effect is exaggerated beyond what can be considered universal. (ii) A 'categorical' rule of phonology must refer to the phonologized property. As an example of the first diagnostic, the vowel length difference in English words such as *bat* [bæt] and *bad* [bæ:d] exceeds any intrinsic tendency for vowel duration to vary as a function of the voicing of a following consonant (Chen 1970). Another example comes from the intrinsic pitch-lowering effect of voiced obstruents which produces the so-called 'depressor consonant' effects in many tone languages: 'Tonal depression in Nguni languages has become phonologized. This means that there is no longer a transparent phonetic explanation for it, and secondly that the phonetic effect has been exaggerated.' (Traill 1990: 166).

The second diagnostic can also be illustrated via the effects of depressor consonants in Ikalanga (Hyman and Mathangwane 1998: 197, 204). As seen in (4a), when the L–L noun *cìthù* 'thing' is followed by L–H *cìícó* 'your sg.' there is no tone change:

(4) a. [cì-thù cìì-có] 'your thing'
b. [cì-pó cı̰ì-có] 'your gift'
L H L H
c. [z^{v}ì-thù z^{v}ìì-zó] 'your things'
d. [z^{v}ì-pó z^{v}ı̰ì-zó] 'your gifts'
L H L H

In (4b), however, the H of the L–H noun *cì-pó* 'gift' spreads onto the pronoun, producing a HL–H sequence. In (4c), the corresponding plural of (4a), there again is no tone change, as expected, since the input is a L–L + L–H sequence. In (4d), the plural and tonal correspondent to (4b), we do expect the H of *-pó* to spread onto the plural prefix *z^{v}ı̰ì-*, as it did in the singular in (4b). However, this does not occur, because the voiced obstruent [z^{v}] belongs to the class of depressor consonants which block H tone spreading in Ikalanga. Since the depressor effect must be referred to by a categorical phonological rule (H tone spreading) the second diagnostic has been met. As is well known to Africanist tonologists, there is a tug-of-war between the natural tendency for tone to spread vs. the intrinsic effects of consonants on pitch:

> Since L–H and H–L tend to become L–LH and H–HL as a natural horizontal assimilation [tone spreading], it can now be observed that the natural tendency of tones to assimilate sometimes encounters obstacles from intervening consonants. Voiceless obstruents are adverse to L-spreading, and voiced obstruents are adverse to H-spreading. The inherent properties of consonants and tones are thus often in conflict with one another. In some languages (e.g. Nupe, Ngizim, Ewe, Zulu), the consonants win out, and tone spreading occurs only when the

consonants are favorably disposed to it. In other languages (e.g. Yoruba, Gwari), the tones win out, as tone spreading takes place regardless of the disposition of intervening consonants. (Hyman 1973: 165–6)

In the terms of Archangeli and Pulleyblank (1994: 211), voiced obstruents are 'antagonistic' to H tone spreading, while other consonants are 'sympathetic'.

Two questions concerning what phonologization was (is) supposed to be are: (i) Does 'intrinsic' mean unavoidable, i.e. 'universally present', or 'universal tendency'? (ii) Does phonologization require that the phonetic feature of the trigger be contrastive? As mentioned earlier, it is widely accepted that one must distinguish between universal and language-specific phonetics (Keating 1988, 1990; Cohn 1993; Kingston and Diehl 1994, etc.). What this means is that there are two diachronic reanalyses which need to be recognized, as in (5):

(5)

a.		b.		c.
universal phonetics	>	language-specific phonetics	>	phonology
('automatic')		('speaker-controlled')		('structured')

First, a perhaps unavoidable universal phonetic property takes on a language-specific form which cannot be said to be strictly automatic or mechanical. The result is still phonetic in the sense of (3), e.g. it may still be gradient rather than categorical. The second diachronic reanalysis occurs when the language-specific property becomes phonological in the traditional sense, i.e. structured, categorical.

This brings us to the question: What does it mean to be 'phonological'? This will determine where 'phonology' begins in (5). For some, anything language-specific, hence (5b), is phonology by definition: '…any rule, gradient or binary, phonologized or categorical, to the extent that it appears in the grammar is fully phonological' (Hajek 1997: 16). The generative approach is to view phonology as a module of grammar. However, there is a notoriously fuzzy boundary between postlexical phonology (Kiparsky 1982) and phonetic implementation (Pierrehumbert 1980): 'The fact that it is difficult to draw a line follows in part from the conception of phonologization (Hyman 1976), whereby over time low-level phonetic details are enhanced to become phonological patterns' (Cohn 2006: 30). Even some of the basic distinctions in (3) have come under scrutiny. Cohn (2006) and Chitoran and Cohn (2009) consider the possibility of categorical phonetics and gradient phonology, while Silverman (2006a: 214) apparently considers all of phonology to be gradient:

… there is no such thing as 'phonologization': at the proper level of description, *all* phonological patterns are sound changes in progress, as they are *all* gradiently and variably implemented, and they are *all* ever-changing…gradience and variation are the very stuff of phonology and sound change…

If the boundary between phonetics and phonology is elusive, perhaps one can less ambiguously characterize phonologization in terms of contrastiveness, the hallmark

of structuralist phonology. Here the central question is: What does it mean to be 'contrastive'? As summarized in (6), the term has been used to refer to different levels of representation and to different domains:

(6) a. *contrastive at what level?*
morphophonemic (URs)
phonemic
phonetic

b. *contrastive within what domain?*
within morphemes
within words (or at stem or word boundaries)
across words (or at phrase or utterance boundaries)

Even if we limit ourselves to the quest for minimal pairs, hence words, it is still necessary to distinguish between underlying and surface contrasts. Many of the examples of phonologization discussed in the 1970s concerned the 'redundant' effects of contrastive features, e.g. [voice] in the following two examples:

(7)	*voice contrast*		*redundant effect*		*contrastive effect*
a.	/bæt/, /bæd/	→	[bæt], [bæ:d]	>	[bæt], [bæ:t]
b.	/pá/, /bá/	→	[pá], [bă]	>	[pá], [pă]

(7a) concerns the oft-reported vowel length difference observed before voiced vs. voiceless stops in English (see Purnell et al. 2005 for updated findings and more subtle discussion). Since vowels are also longer before fricatives and sonorants, e.g. *gas* [gæ:s], *man* [mæ:n], the process appears to be one of shortening before voiceless stops (House 1961). Be that as it may, the durational differences are first phonologized and then potentially phonemicized by final devoicing, as seen in the outputs. Conceptualized this way, the underlying voice contrast would correspond to a surface length contrast in English.

The second case, (7b), has been much discussed in both the phonologization and tonogenesis literature. Here we start with a H tone on syllables whose obstruent onset differs in voicing. As seen, the intrinsic lowering effect of voicing on f_0 is first phonologized to create a rising tone on [bă], whose consonant subsequently undergoes devoicing. The result is a 'tonal bifurcation' whereby the rising tone becomes phonemic.

Much of the work on phonologization concerns such cases of *re-* or *transphonologization* of contrasts (Jakobson 1931; Hagège and Haudricourt 1978). There are at least two possible interpretations of the voicing effects on duration and f_0. The first is that the phonologizations in (7) represent an enhancement of phonetic voicing. The second is that they instead enhance the phonological [voice] CONTRAST. The latter view of phonologization is explicitly adopted by a number of researchers:

> … because no other articulation is likely to produce the F_0 depression as an automatic byproduct, the depression must itself be a product of an independently controlled articulation, whose purpose is to enhance the [voice] contrast. (Kingston and Diehl 1994: 425)

Enhancement of the type we are considering here can be considered as a form of 'fine-tuning' of a basic phonological contrast. (Keyser and Stevens 2001: 287)

While it is possible to view such 'redundant' effects of voicing as enhancements which provide additional cues of the voicing contrast, the question is whether this strengthens vs. weakens the contrasting feature, here [±voice]. It is quite striking how allophonic variations such as in (7) often lead to the loss of the original contrast. In fact, some have seen transphonologization as having the purpose of maintaining a contrast which is being threatened:

la transphonologisation: une opposition ayant valeur distinctive est menacée de suppression; elle se maintient par déplacement d'un des deux termes, ou de l'opposition entière, un trait pertinent continuant, de toute manière, à distinguer ces termes (Hagège and Haudricourt 1978: 75)

On the other hand, phonologization need not imply transphonologization:

I will use the term phonologization throughout to mean specifically the innovation of changes to phonological representations, whether these result in neutralization of contrasts or not. (Barnes 2006: 16)

However, is phonologization always motivated by contrastiveness? In the present context the question is: What can contrastive [voice] do that phonetic voicing can't? This question will be further examined in section 1.2.1.

1.2.1 *Voiced prenasalized consonants and tone*

Recall that we are concerned in determining if it is only contrastive [±voice] which may trigger phonologization. As a hypothetical test case, consider a language which has /t, k, b, d, g/, but no /p/. As seen in (8a), we begin with CV inputs with H tone:

(8)	*input*	*phonologized*	*transphonologized*
a.	tá, ká	tá, ká	tá, ká
b.	dá, gá	dă, gă	tă, kă
c.	bá	bă ?	pă ?

In (8b) these H tones become rising after [d] and [g], a phonologization which could be seen as an enhancement either of phonetic voicing or of their contrast with /t/ and /k/. The real question is what would happen in (8c), where /b/ is phonetically voiced, but does not contrast with /p/. Would the redundant voicing of [b] have an f_0 effect, as shown, or would this phonologization be blocked because there is no contrast with [p]? The phonological enhancement theories of Kingston and Diehl (1994) and Keyser and Stevens (2001) would need to be tweaked by some notion of phonetic analogy (Vennemann 1972a) if (8c) does develop the rising tone. On the other hand, (8c) seems to be allowed, if not predicted, by Ohala's (1981, 1992, 1993b) theory of sound

change, which involves a reinterpretation of the phonetic signal, as well as Kiparsky's (1995: 656) 'priming effect': 'Redundant features are likely to be phonologized if the language's phonological representations have a class node to host them'. That is, the intrinsic f_0 effect of voiced obstruents is most likely to become phonologized in languages which already have a tonal contrast (Matisoff 1973; Svantesson 1989).[1]

While the above example and discussion are hypothetical, a real test case can be derived from the following characteristic effects of 'depressor consonants' in African tone systems:

(9) a. trigger
   i. lowering of H or L
   ii. conversion of H to LH or L
   iii. delinking of H (esp. if followed by H)

b. block
   i. raising of H or L
   ii. H tone spreading (cf. (4d))
   iii. H tone plateauing

To account for the relation between consonant types and tone in synchronic phonologies, Halle and Stevens (1971) and Halle (1972) proposed the following distinctive feature analysis, where [stiff] = stiff vocal cords and [slack] = slack vocal cords:

(10)

	tones			*voiceless obstruents*	*sonorants*	*voiced obstruents*
	H	M	L	p t k f s	m n l w y	b d g v z
stiff	+	−	−	+	−	−
slack	−	−	+	−	−	+

As seen, both H tone and voiceless obstruents are [+stiff, –slack], while L tone and voiced obstruents are [–stiff, +slack]. Both M tone and sonorants are [–stiff, –slack]. Like vowels, sonorant consonants readily accept any tone, while obstruents have the tonal affinities indicated above. While these features are often assumed to this day, there are additional complications, as noted in the observations in (11).

(11) a. The above three-way distinction is not sufficient for tone (there can be a fourth or fifth contrastive pitch level).

b. The above three-way distinction is not complete for consonants (Hombert 1978), e.g:

   i. implosives are often pitch-raisers, hence expected to pattern with voiceless obstruents
   ii. breathiness and creak are typically pitch lowerers; aspiration is more complex.

[1] Nick Clements has brought Ewe to my attention, where /b/ and /ɖ/ are depressors even though /p/ occurs only in borrowings, and there is no voiceless counterpart to /ɖ/ at all (see Clements 2005).

c. While the 'best' pitch depressors are fully or breathy voiced obstruents, and although the phonetics of voice is complex (Kingston and Diehl 1994), depressor consonants readily become unvoiced, e.g. in Nguni (Schachter 1976; Traill 1990; Downing 2009).

d. Prenasalized voiced stops [mb, nd, ŋg] are sometimes depressors, sometimes not.

It is the observation in (11d) which potentially bears the question with which we are concerned: Is it phonetic voicing or enhancement of CONTRASTIVE [voice] that causes depressor effects? The following quotations show that there is a widespread belief that the voicing on depressor consonants is necessarily contrastive:

… F_0 will only vary with the presence of voicing in stops that contrast for [voice]…. (Kingston and Diehl 1994: 436)

Since implosives and prenasalized stops are not contrastively voiced [in Suma], they are assumed to be unspecified for the feature [voice] and, therefore, naturally excluded from the depressor consonant group. (Bradshaw 1995: 263)

Il convient de souligner que seules les consonnes *phonologiquement* sonores—c'est-à-dire s'opposant à des sourdes de même point et mode d'articulation—exercent un effet d'abaissement [in Yulu], ce qui n'est jamais le cas des consonnes *phonétiquement* sonores des séries glottalisée (partiellement), prénasalisée, nasale continue et vibrante. Cet état de fait prouve, s'il en est besoin, la pertinence d'une approche phonologique des unités articulatoires. (Boyeldieu 2009: 199n; emphasis my own)

While Bradshaw and Boyeldieu assume that implosives fail to lower pitch because they are non-contrastively voiced, the prevalent view has been that rapid lowering of the larynx and tensing of the vocal chords provide quite adequate phonetic explanations for why implosives tend to pattern with voiceless obstruents and H tone.[2] On the other hand, the ambivalent behavior of the voiced prenasalized stops [mb, nd, ŋg], which are sometimes depressors and sometimes not, is indeed puzzling. The question is whether their ambivalence has anything to do with contrastiveness.

As a practicing structuralist phonologist, my initial hypothesis was that /mb, nd, ŋg/ would function as depressor consonants only in languages where they contrast with /mp, nt, ŋk/. In order to test this hypothesis, I examined the relatively small group of African tone languages which have both depressor consonant effects and voiced prenasalized consonants (ND), whether contrastive with their voiceless counterparts (NT) or not. The results are presented in the following table:

[2] More recently, Tang (2008) has argued that the tonal effects of implosives can pattern with those of voiceless obstruents, voiced obstruents, or sonorants in different languages. While implosives do not contrast in voicing in these languages, it is yet to be determined to what extent these differences can be attributed to differences in phonetic production. The same conclusion will be reached with respect to voiced prenasalized stops.

(12)

	ND contrasts with NT	ND doesn't contrast with NT
ND = depressor	Nguni*	Lamang, Musey, Ngizim, Ouldeme, Podoko, Mbuko
ND ≠ depressor		Bole, Geji, Miya, Zar; Yulu, Suma, Mijikenda*

(*Nguni includes Swati, Zulu, Ndebele, Xhosa; Mijikenda includes Giryama, Digo, Kauma, Rihe.)

As seen, three of the four logical combinations of the two properties ([±contrast], [±depressor]), were found. Setting aside borrowings (see below), the only languages with a /NT, ND/ contrast were the Bantu Nguni languages of Southern Africa. Of the remaining languages, all of those in the upper right quadrant are Chadic, as are the first four languages of the lower right quadrant. Yulu is Central Sudanic, Suma is Ubangian, and the Mijikenda languages are Bantu.

From (12) we conclude the following: (i) If /ND/ contrasts with /NT/, /ND/ will have the same f_0 effects as /D/. (ii) If ND does not contrast with NT, ND may have the same f_0 effects as /T/ or /D/. As mentioned, the first group consists solely of the Nguni languages, e.g. Swati:

> In all cases [in Swati], the prenasalized counterparts of depressor consonants are themselves depressor consonants, while the prenasalized counterparts of non-depressor consonants are themselves nondepressors. (Schachter 1976: 213)

It may be relevant to note that the Nguni languages have a rule of postnasal deaspiration ($NT^h \rightarrow NT$). The alternations in (13) illustrate the application of this rule in Ndebele (Pelling 1971; Galen Sibanda, pers. comm.):

(13)	a.	u-p^{h}ondo	'horn'	pl.	im-pondo	cf.	impisi	'hyena'
		u-p^{h}awu	'sign, mark'	pl.	im-pawu		imbizi	'pot, pan'
	b.	u-t^{h}ango	'fence'	pl.	in-tango	cf.	intaba	'hill, mountain'
		u-t^{h}ungo	'rafter'	pl.	in-tungo		indaba	'matter, news'
	c.	u-k^{h}uni	'firewood'	pl.	iŋ-kuni	cf.	iŋkalo	'waists, hill passes'
		u-k^{h}alo	'waist'	pl.	iŋ-kalo		iŋgalo	'arm'

As seen in the forms to the right, this distributional constraint produces (near) minimal pairs involving unaspirated [mp, nt, ŋk] vs. voiced [mb, nd, ŋg]. The latter's depressor effect on tone may therefore be a welcome cue for the voicing contrast. It is interesting to note in this context that a much larger group of Bantu languages have a rule of postnasal aspiration ($NT \rightarrow NT^h$), e.g. Mwiini, Zigula, Pokomo, Pare, Shambala, Ngulu, Bondei, Namwanga, Chichewa. This process may then lead to the

transphonologization of aspiration (NT → T^h), as in Swahili, Yaa, Giryama, Digo, Yaka, Cokwe, Makua, and Venda. As a result, the Mijikenda languages Giryama and Digo have a surface contrast between T^h and ND consonants, the latter of which are not depressors.

Turning to the languages in the right-hand column of (12), where ND does not contrast with NT, it should be noted that the difference between the two groups of languages cannot be attributed to the nature of the tonal property in question. Quite comparable tonal proceses occur in languages which treat ND differently, e.g. register lowering after ND in Podoko, but not in Yulu, blocking of H tone spreading by ND in Ngizim, but not in Bole or Zar.

The question is how to explain the inconsistent depressor status of ND when voicing is non-contrastive. We will mention four potential accounts. The first is to seek an explanation in phonetic terms: NDs may have slightly different phonetic properties in languages where they function as depressors vs. those languages in which they function as non-depressors. Perhaps ND is fully voiced in one language, but partially devoiced in another. Or perhaps there are slightly different phonations associated with ND in the different languages. Another phonetic difference could be in the timing of the nasal vs. oral portions of the unit: depressor NDs might have a longer D phase than non-depressor NDs. Since Cohn and Riehl (2008) have recently argued that there is no phonetic difference between a prenasalized stop (^{N}D) and a post-stopped nasal (N^D), pointing out that the D phase is universally short, this does not seem likely—nor is there any motivation for recognizing monosegmental ^{N}D vs. bisegmental ND. In the absence of instrumental evidence, speculations on phonetic differences are simply that.

A second approach is to seek an explanation in the history of the different languages. For instance, perhaps ND behaves as a depressor when it derives from *D, perhaps as 'hypervoicing' (Iverson and Salmons 1996), but as a non-depressor when it derives from *N via partial denasalization (Wetzels 2007). Although such sources have been documented in Mexico, Amazonia, New Guinea, and other parts of the world, the history is less clear in Chadic, which we have seen to be inconsistent in how it treats ND and tone. A different kind of history might be one involving analogy: Perhaps languages with depressor ND have (or had) processes by which D and ND were morphophonemically related, which then caused the pitch-lowering effect of D to extend to ND. Perhaps this relationship was missing in the other languages, which may instead have had a relation between N and ND. Like the first two accounts, this one also is speculative in the absence of historical evidence.

A third strategy is to recognize ND as a separate category from the three consonant types distinguished in (10). Perhaps the high-to-low hierarchy of consonant–f_0 relations should be T >> N >> ND >> D (where N = sonorants) with languages drawing the depressor line in different places, as in (14).

(14)

	H	M		L
ND = depressor	T	N	ND	D
ND ≠ depressor	T	N	ND	D

The problem is that we do not know what the intrinsic effects of ND on f_0 really are. The hierarchy in (14) suggests that ND has more of a depressor effect than N, but less than D. We don't really know this other than from the phonological facts, which are inconsistent. What is needed are instrumental studies of ND in languages which have not phonologized depressor consonant effects. We need to do this both for languages which have a phonetic NT/ND contrast, e.g. Luganda, and which don't, e.g. Kinande—ultimately establishing what the intrinsic effects of ND are expected to be even in non-tone languages.

The fourth and last account seeks an explanation in terms of contrast, but in the absence of /NT/ suggests that it is a different contrast that is being enhanced: /ND/ vs. /N/. Languages which treat ND as a depressor do so to distinguish it further from N. Particularly if the oral phase is minimal, there could be perceptual confusion between ND and N, and hence transphonologization via the tone of the following vowel. Such has happened in Masa, a Chadic language closely related to Musey. While /H/ tone can occur after any consonant, there is a (near-) predictability of L vs. M tones as in (15) (Caïtucoli 1978: 77):

(15)

	initial root segments	*tone*
a.	b, d, g, v, z, ž, ɮ, ɦ	L
b.	p, t, k, f, s, č, ɬ, h, ɓ, ɗ, l, r, w, y, a, e, i, o, u	M
c.	m, n, ŋ	L, M

As seen in (15a), L tone appears after a voiced obstruent, while M tone appears if the root-initial segment is a voiceless obstruent, an implosive, or an oral sonorant, including vowels. While several Chadic languages have similar distributions of L and M tones, the originality of Masa is that it has a L vs. M contrast after nasals. The reason, of course, is that there has been a sound change of **mb, *nd, *ŋg > m, n, ŋ* with the original contrast being transphonologized in terms of L vs. M pitch. Crucially, those roots which had historical *ND now have L tone, while those which began with *N have M tone. Since closely related Musey treats ND as a depressor (cf. (12)), we can be reasonably certain that the same was true in pre-Masa before the prenasalized consonants lost their oral release. While we cannot predict which nasals will be depressors, it is possible to say that contrastive [+voice] necessarily conditions L tone: 'Le ton moyen est incompatible avec les consonnes sonores ayant une correspondante sourde…' (Caïtucoli 1978: 77).

Transphonologization of an earlier ND vs. N contrast is not without parallel. As seen in (16), such a merger, either complete or in progress, has been transphonologized as a contrast in vowel nasalization in several Western Austronesian languages (Court 1970):

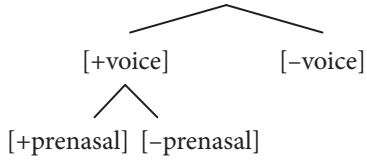
(16)		*NDV		*NV	
a.	Sea Dayak	[nãŋ$^{(g)}$a]	'to set up ladder'	[nãŋã]	'to straighten'
b.	Sundanese	[mãn^{d}i]	'to bathe'	[mãnĩ]	'very'
c.	Ulu Muar Malay	[mŋ$^{(g)}$oet]	'to twitch'	[mŋõẽʔ]	'to bellow'
d.	Měntu Land	[əm$^{(b)}$ak]	'gong stick'	[əmãk]	'sleeping mat'
	Dayak	[ɲĩn$^{(d)}$aʔ]	'to love'	[ɲĩnãʔ]	'snake (sp.)'

As seen, progressive vowel nasalization appears to set in before *ND completely loses its oral release, just as we can assume the depressor effect of ND to precede its simplification to N in Masa. Western Austronesian and Chadic are thus quite parallel, the difference being the feature that is chosen for the transphonologization. While Western Austronesian is sensitive to the nasal vs. oral release of N vs. ND, the contrast which is enhanced in Chadic is the sonorant vs. obstruent release of N vs. ND. As we have seen, it is the combination of obstruent and [+voice] that produces the pervasive f_0 lowering seen both in African tone systems and elsewhere. The problem, of course, is to show with certainty that the ND depressor languages in the upper right quadrant of (12) have a shaky ND vs. N contrast in need of reinforcement as against a more robust ND vs. N contrast in the languages of the lower right quadrant.

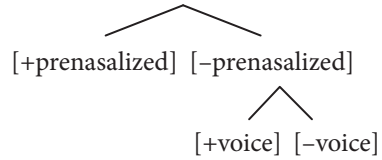
In (12) the different African languages were classified according to whether they have a contrast between /NT/ and /ND/. One complication concerns what to do about languages which have NT only in borrowings, e.g. Ikalanga *kámpá* 'a camp', *pénte* 'paint', *donkí* 'donkey'. In this language of the Shona group, inherited *mp, *nt, *$ŋk$ become p^{fi}, t^{fi}, fi, whereas in Shona proper *mp, *nt, *$ŋk$ > m^{fi}, n^{fi}, fi. In both languages the resulting consonants lower pitch, thereby illustrating that *NT can also develop into depressor consonants.

To summarize, we have seen that depressor NDs suggest that the effects of a non-contrastive [+voice] trigger may also be phonologized. As Sharon Inkelas (pers. comm.) has reminded me, this is reminiscent of the interaction of predictable postnasal voicing with Lyman's Law in Japanese (Ito, Mester, and Padgett 1995). As in the Japanese case, we are still faced with how to formalize the synchronic differences between the two groups of languages in the right-hand column of (12). This turns out not to be a problem, rather a case of having too many possibilities. First, since postnasal voicing is redundant, one could analyze the non-depressor vs. depressor difference as /NT/ vs. /ND/. Or, one could use different feature or feature-geometries for the two kinds of NDs, underspecification, or perhaps different contrast hierarchies (Dresher 2003, 2009; Mackenzie 2008), as in (17).

(17) a.



b.



As seen in (17a), the primary contrast is [±voice], which is further differentiated into [±prenasalized] (or whatever feature/representation distinguishes D and ND). In (17b), however, the first cut is [±prenasalized], and only [–prenasalized] consonants are further distinguished for [±voice]. If tone is sensitive to [+voice], ND consonants will be depressors in Ngizim, but non-depressors in Miya.

The issue of providing different underlying representations for the 'same' segment types in different languages is an old tradition, and it has come in handy in treating nasality (see Piggott 1992 and Rice 1993, for instance). In order for such a move to be compelling it must not appear circular or ad hoc, but rather have implications that hold through the language question. So far this has turned out to be a problem. Schuh (1998: 13), for example, treats the non-depressor NDs of Miya as [+sonorant], but recognizes that this poses a problem for one of his rules:

> … if the last consonant in a word is an obstruent, it must be followed by /ə/, whereas if the last consonant is a sonorant, nasal, it cannot…Here, prenasalized consonants pattern with obstruents (*gùmbə* 'gourd' vs. *gwágúm* 'dove').

While he proposes to account for the inconsistency by proposing that ND begins as a sonorant (hence a non-depressor) and ends as an obstruent (hence requiring schwa), it has already been pointed out that the same tonal process may occur in both types of languages in the right-hand column in (12). Since my interest here is in the nature and motivation of the phonologization process, I will leave further implementational issues to another time.[3]

1.2.2 *ATR harmony in Punu*

In the preceding subsection we have seen that it is possible for phonologization to be triggered by a non-contrastive feature. In this section I present a perhaps even more striking case of this involving ATR vowel harmony in Punu, a Bantu language

[3] Louis Goldstein has suggested to me that when the voicing of ND is non-contrastive, speakers need not invoke articulatory mechanisms that result in lowered pitch, whereas such mechanisms are unavoidable when there is a contrast with NT. It is significant that all of the examples cited by Lee (2008) involve depressor consonants whose voicing is contrastive. Most striking is Tsonga (Baumbach 1987), where NDs do not contrast with NT and are not depressors, but their contrastive breathy counterparts N̤D̤ are. In such a case, there is a disincentive for ND to exploit the gesture(s) which result in the lowering of f_0. Thanks to both Louis Goldstein and Maria-Josep Solé for helpful discussions of these matters.

spoken in Gabon. It is useful to distinguish two prototypes of vowel harmony (VH), each of which shows clear structure-dependency. The first is root-controlled VH (Clements 1981) whereby harmony expands out from a root vowel to affixes. This type of harmony is often bidirectional, feature-filling, and structure-preserving. The second prototype is non-root-controlled and is often referred to as 'metaphony' or 'Umlaut'. In this case VH is anticipatory, hence unidirectional. Suffixes can be triggers, while prefixes rarely, if ever, are (Hall and Hall 1980; Hyman 2002, 2008a; Krämer 2003). Prefix-triggered VH on a following vowel is rare or non-occurring because it is neither root-controlled nor anticipatory (Hyman 2002). Attempts to attribute VH to the phonologization of vowel coarticulation (Ohala 1994b; Beddor and Yavuz 1995; Przezdzieci 2005) must account for why VH is typically unbounded and word-delimited (cf. Barnes 2006: 197–200).

In this section we are concerned with non-contrastive ATR harmony in Punu (Kwenzi Mikala 1980; Fontaney 1980). In Punu the five vowels /i, ɛ, u, ɔ, a/ contrast within the first CV of a root, most of which are CVC-. Prefixes, suffixes, and non-initial root vowels are limited to /i, u, a/. Although /ɛ, ɔ/ are limited to the first syllable of a root, they become tense or [+ATR] in the following contexts (Kwenzi Mikala 1980: 9):

(18) a. /ɛ, ɔ/ → [e, o] / ___ C i
b. /ɛ, ɔ/ → [e, o] ~ [ɛ, ɔ] / ___ C u
c. /ɛ, ɔ/ → [ɛ, ɔ] / ___ C a

Other than occurring in occasional ideophones, the only other occurrences of [e] and [o] result from the fusion of /a+i/ and /a+u/, respectively, which succeed each other only in prefixes:

(19) a. /a-i-lab-i/ → [é-lǎb-ì] 'he sees' (-lab- 'see' is the root)
b. /a-u-lab-a/ → [ó-lǎb-ə̀] 'he will see'

Finally, an /a/ which occurs 'post-radically', i.e. after the first syllable of the root, is automatically realized as [ə].

With the above vowel processes established, the distribution of underlying and surface vowels can be summarized by position, as in (20):

(20)

	Prefixes	*Root*	*Suffixes/post-radical vowels*
Underlying:	/i, u, a/	/i, ɛ, u ɔ, a/	/i, u, a/
Surface:	[i, e, u, o, a]	[i, e, ɛ, u, o, ɔ, a]	[i, u, ə]

Since the fusions in (18) result in [e, o], not *[ɛ, ɔ], a feature such as [+tense] or [+ATR] can be assumed to be phonologically 'active' on /i, u/. In (21) I assume a privative feature analysis, where each of the features A, F, R, and O is phonologically active (Hyman 2002, 2003):

(21)

		underlying vowels					*derived vowels*		
		i	u	ɛ	ɔ	a	e	o	ə
ATR	A	x	x				x	x	
Front	F	x		x			x		
Round	R		x		x			x	
Open	O			x	x	x	x	x	

As seen, the postradical process /a/ → [ə] would be interpreted as the deletion of the Open feature (which technically yields [ɨ], from which Punu [ə] is non-distinct).

The crucial point concerns the assimilation of /ɛ/ and /ɔ/ to [e] and [o] before /i/ and /u/. This clearly has to be viewed as a phonologization of the common tendency to tense mid vowels when they are followed by a high vowel in the next syllable. However, it can be observed from the feature specifications in (21) that the ATR feature, although active, is non-contrastive on the input vowels: Without ATR, /i/ and /u/ would still be distinct from /ɛ, ɔ, a/ in not having an Open feature. Thus, the tensing process involves the phonologization of a non-contrastive feature.

Recall from section 1.2.1 that we allowed for the possibility that non-contrastively voiced ND might exert a depressor effect by virtue of its contrast with plain nasals. It is hard to make a similar case for Punu. Since post-radical /i/ and /u/ contrast only with /a/, which is realized as [ə], there seems to be little, if any, need to enhance this highly redundant, minimal contrast. In fact, there are additional processes which further obscure post-radical vowels. The first two in (22a, b) concern R- and F-VH, while /a/-reduction is repeated in (22c).

(22) a. a, i → u / __ C u
b. a → i / __ C i
c. a → ə

	i	u	a
i	i-i	u̲-u	i-ə
u	u-i	u-u	u-ə
a	i̲-i	u̲-u	ə-ə

The rules in (22a, b) result in considerable loss of contrast. As seen in the distributions to the right, nine phonological inputs result in only six distinct outputs. What's worse, when /CɛC-aC-i/ and /CɛC-aC-u/ are realized as CeCiCi and CeCuCu, the input /a/ is no longer recoverable. The inescapable conclusion is that phonologization is not necessarily triggered by contrastiveness, nor does it necessarily lead to

transphonologization (cf. Blevins 2004: 43). While Punu may ultimately develop an underlying seven- (or eight-) vowel system, the mid-vowel ATR harmony appears to have been phonologized as a 'mere' articulatory convenience!

In the following section we will extend these findings to other phonological phenomena and then turn to their relation to grammaticalization in general.

1.3 Phonologization and grammaticalization

In section 2 it was established that phonologization is not necessarily dependent on contrastiveness. In this section I first compare this result with other types of phonologization and then suggest that phonologization should be viewed as one aspect of 'grammaticalization'.

1.3.1 *Other types of phonologization*

While most of the discussion has centered around the phonologization of phonetic processes, the terms 'phonologization' and 'grammaticalization' have both been invoked to refer to the activation of any formal property within a phonology. The question, then, is whether other phonologizations such as those listed in (23) are dependent on contrastiveness?

(23) a. distinctive features (redundant or contrastive)

> It will be generally assumed that the inventory of phonological features is identical to the inventory of phonetic features, and that languages implement these universal phonetic features in various linguistic ways. In other words, phonetic features can be 'phonologized'. (Hyman 1975: 58–9)

b. prosodic constituents: syllable, foot, phonological word

> We may interpret the existence of the prosodic unit 'syllable' as a grammaticalization of one of the planning units for the coordination of muscular gestures. [Re the foot:] …for each language this general rhythmic tendency is grammaticalized into particular phonological rules of foot construction. (Booij 1984: 274)
>
> Since stress has these intrinsic properties associated with it, it is not surprising to find languages phonologizing…these properties into rules of the language. Numerous cases of strengthening in stressed syllables and weakening in unstressed syllables are attested.…(Hyman 1975: 207–8; cf. Barnes 2006: ch. 2)

c. distributional constraints on morphemes, stems, words, ultimately templatic, e.g. the maximum CVCVCV 'prosodic stem' in Tiene, where C_2 must be coronal and C_3 must be labial or velar (Ellington 1977; Hyman 2010a)

d. demarcation: initiality-/finality-effects (Keating et al. 2003; Barnes 2006), final glottalization (Henton and Bladon 1988; Hyman 1988); also root-affix asymmetries, stem-initial prominence (Beckman 1997; Smith 2002)

e. intonation based on the 'grammaticalization' of three biological codes (Gussenhoven 2004: ch. 5)

f. 'boundary narrowing': pause > phrase > word; phonologization of prepausal effects, which can include final devoicing, debuccalization, glottalization, lengthening, 'nasal pause' (Aikhenvald 1996: 511–12), and loss

I would like to suggest that the 'pronunciation in isolation' form of a word is its lexical representation. At the pause…words may undergo phonetic modifications; in particular, final oral stops may become unreleased as in English and thereby lose their aspiration, and vocal cord vibration may cease early, leading to devoicing. Since they occur at the pause, and the ad-pausal variants are registered in the lexicon according to my proposal, these ad-pausal variants may next appear in connected speech and may cause or undergo further changes in their new context. (Vennemann 1974: 364)

Even a cursory glance over (23a–f) will reveal that contrastiveness is involved in some aspects of the above phonological issues, but not others. Thus, it has long been observed that syllable structure is never underlyingly contrastive—its very redundancy or predictability in fact kept syllable boundaries (and syllable constituents) out of early generative phonology:

One argument which has been raised against phonological syllables is that, unlike segments, the location of a syllable boundary within a morpheme can never be phonemic. That is, two morphemes such as /a$pla/ and /ap$la/ cannot differ only in their syllable structure.… Because syllable boundaries can be determined automatically from universal principles and language-specific facts about the segments contained in the syllables, generative phonologists have largely worked under the assumption that the syllable is unnecessary in phonology. (Hyman 1975: 192)

The syllable would thus appear to have more of an organizational function, rather than a contrastive one, also presumably the metrical foot and higher level prosodic domains. The phonologization of prepausal effects is perhaps less clear. It is tempting to interpret languages which insert prepausal glottal stops as having phonologized utterance-final creakiness, as in British English (Henton and Bladon 1988):

… final GS may be conditioned by a number of disparate factors from all parts of the grammar. Since the common denominator appears to be 'before pause in declarative utterances', it is tempting to conclude that such GS's result, historically, from the PHONOLOGIZATION of an intrinsic variation in the speech signal. In the case of prepausal vowels, the speaker is expected to cease voicing with the completion of the vowel. When GS is not present, this cessation is smooth, in many cases giving the impression of a final slight breathiness. On the other hand, when GS is present, the cessation of voicing is abrupt, giving the impression of a non-syllabic articulation, i.e. a final 'consonant'. (Hyman 1988: 124)

While some languages suspend the final glottal stop in questions, suggesting a contrastive function between declaratives and interrogatives, the situation can be much

more complex. Thus, in Dagbani (Gur; Ghana), a prepausal glottal stop is inserted if a complex set of conditions is met (Hyman 1988: 122):

(24)
a.	phonetic condition:	before pause
b.	pragmatic condition:	'declarative' utterance (i.e. not interrogative)
	plus either:	
c.	syntactic condition:	final word is within scope of negation
	or:	
d.	phonological condition:	after a short, stem-final vowel
e.	morphological condition:	final word is [–Noun]

In fact, final glottal stops do not always derive from prepausal phonologizations. In certain Akan and Guang languages to the south of Dagbani glottal stops transparently derive from apocope:

(25)
Akuapem/Asante	Fante		Chumburung	Gonja	
jírì	jîrʔ	'overflow'	wʊrɪ	ka-wʊlʔ	'skin'
hʊ̰mɪ̰	hʊ̰mʔ	'breathe'	kɨ-furi	ku-fulʔ	'moon'
tʊ̰nʊ̰	tʊ̰nʔ	'forge'	ɔ-narɪ	e-ɲinʔ	'man'
Akan (Schachter and Fromkin 1968: 204)			Guang (Snider 1986: 136)		

In Tikar, glottal stops are restricted to prepausal position (Jackson and Stanley 1977, Stanley 1991). As proposed in Hyman (2008b), these final glottal stops result from the debuccalization of coda **t* and **k* which are realized as glottal stops before a pause, but as Ø before a consonant. As part of the process, back vowels were fronted before **t*, while front vowels became backed before **k*, hence transphonologizing the F2 properties of the two coda consonants as per Thurgood and Javkin (1975).

Concerning boundary narrowing, although Luganda must originally have shortened bimoraic long vowels before pause, present-day final vowel shortening is subject to a number of complex factors and no longer requires pause (Hyman and Katamba 1990). It seems that while contrast can become implicated in a phonologization process, it is typically not the driving force of the phenomena enumerated in (23). If the analysis of Punu in (21) is correct, even a redundant distinctive feature, e.g. ATR, may first become activated for allophonic effects and only later become contrastive.

While a phonetic motivation has been assumed for all of the phonologizations in sections 1.1 and 1.2, at least some of the phonological properties in (23) raise the question of whether phonetics is the only source of phonology, i.e. the only input to phonologization. At least three other sources of phonology have been proposed in the literature. First, phonology has been claimed to occasionally arise from frequency distributions:

> … it is possible for a phonological generalization to arise from frequency distributions in the lexicon rather than from pure coarticulation effects. However, the former type are much less

frequent, since the conditions for coarticulation effects are always present in spoken language. (Bybee 2001: 94–5)

Second, certain phonological properties have been said to derive from analogical processes:

… new phonemes can arise through morphophonemic analogy.… In all such cases…no new distinctive features are added.… morphophonemic genesis merely leads to a combination of distinctive features which had not previously been used. (Moulton 1967: 1405)

… phonetically unnatural patterns can also arise by analogical processes. Since they are phonetically unnatural, they do not have purely phonological origins, but reflect instead the generalization of fortuitous morphological patterns.… even the most regular morphophonological patterns may lack phonetic origins. (Garrett and Blevins 2009: 543)

Finally, phonological distributions and alternations can be due to borrowing. For example, many phonologists assume that English has a rule of velar softening responsible for such alternations as in *electric* vs. *electricity*, where the *k*∼*s* alternation is clearly borrowed from French.

The above three non-phonetic sources of phonology are of course more indirect and less frequent producers of phonology than phonetics. If 'phonologization' is interpreted literally as the creation or genesis of phonology, then all of the above can be referred to as such. However, most phonogenetic work has been concerned with the phonetics > phonology sense of the term with simultaneous focus on the structural codification and dephoneticization of possibly universal phonetic tendencies. I return to this dual notion of phonologization + dephoneticization in section 1.4.

1.3.2 *Phonologization as grammaticalization*

In this section I address the question of how phonologization fits into the overall scheme of grammar and grammar change. As indicated in (26), phonologization can be identified with the second of the four stages which Baudouin de Courtenay's (1895 [1972a: 197]) proposes for the development of an alternation:

(26)

1. embryonic alternation	intrinsic	takes conscious effort to perceive
2. neophonetic alternation or divergence	extrinsic, phonologized	minimal effort to perceive
3. paleophonetic or traditional alternation	phonemicized	*Phonologisierung* (Jakobson 1931)
4. psychophonetic alternation or correlation	morphologized, lexicalized	exceptional, arbitary

In the above I have provided Baudouin's terminology, the modern equivalences, and a few descriptive notes. Baudouin's insights are clearly mirrored in the work of Vennemann (1972a, b), Dressler (1976, 1985), Joseph and Janda (1988), and others on the rise-and-fall 'life cycle' of phonology, where the stages in (27) are distinguished:

(27) phonetic > phonologized > phonemicized > morphologized > lexicalized > LOSS

We have already discussed phonologization and phonemicization, the latter typically being the product of transphonologization. Morphologization refers to the loss of the phonological condition on an alternation, while lexicalization comes in when specific morphemes have to be marked as undergoing vs. not undergoing the alternation. As the alternation develops greater exceptionality, one arrives at a stage where there are only relics of the original rule, followed by its loss entirely.

While intended only to capture the natural history of phonological processes, the stages in (27) are strikingly similar to the stages of Givón's (1979) proposal for the rise and fall of syntax and morphology, which I slightly reword as in (28):

(28) pragmatic > syntactic > morphological > morphophonemic > lexical > LOSS

As seen, Givón was primarily concerned with the development of syntax from pragmatics, which he refers to as 'syntacticization'. Once a property has become syntactic, it can then become morphological, as when an original independent word becomes a concatenated affix, perhaps with phonological reduction or erosion. Givón's morphophonemic stage arises when the original source is obscured, ultimately producing a phonological alternation which is morphologically conditioned or morphologized. This alternation may then become lexicalized and lost as in (26).

While phonology plays a role in Givón's view of the rise and fall of grammar, he is mainly interested in the first three stages of (28), for which he had established the mantra, 'Today's morphology is yesterday's syntax' (Givón 1971: 413). In fact, the parallel in (29) is something that phonologists readily acknowledged during this period:

(29) Phonetics : phonology pragmatics : syntax

> … it is…very much part of the business of phonologists to look for 'phonetic explanation' of phonological phenomena.…just as when syntacticians look for pragmatic accounts of aspects of sentence structure, the reason is to determine what sorts of facts the linguistic system proper is not responsible for…(Anderson 1981: 497)

> Phonetics provides much of the substance of phonology, and pragmatics provides much of the substance of syntax. However, the ever-present phenomena of phonologization and grammaticalization cannot be explained by reference to the origin of the substance. (Hyman 1984: 83)

Two examples of the syntacticization of pragmatic tendencies concern the following subject–object asymmetries which, as pointed out in much of the literature of the time (e.g. in various papers in Li 1976), tend to have the properties indicated in (30):

(30)

subjects	*vs.*	*direct objects*
given (old)		new
presupposed		asserted (focused)
definite		indefinite
animate		inanimate
1st/2nd person		3rd person
actor		non-actor

The first example is the tendency for subjects to be definite. While in some languages the correlation between subjecthood and definiteness is statistical, in others it becomes a requirement imposed by the grammar. Looking at different discourse genres in English, Givón (1979: 52) reports the following counts of definite subjects and direct objects in declarative-affirmative-active clauses:

(31)

subject		*direct object*	
definite	*indefinite*	*definite*	*indefinite*
302 (91%)	33 (9%)	193 (56%)	156 (44%)

As seen, the skewing between definite and indefinite is dramatic in subject position, or, as Givón notes, 156/189 of the indefinite noun phases occur as direct object. What is important is the relationship between English, which tends to have definite subjects, vs. various Austronesian languages which REQUIRE the subject to be definite (Keenan 1976; Schachter 1976). Put differently, English is at the pragmatic/phonetic stage, while Malagasy and Tagalog are at the syntactic/phonological stage.

The second example concerns the tendency for the direct object site to double as a focus position: '…the basic position for the focused or emphasized constituent is that position which is filled by the object in a neutral sentence' (Harries-Delisle 1978: 464). While, again, this tends to hold pragmatically in discourse, SOV languages may syntacticize the immediate-before-verb (IBV) position and SVO languages the immediate-after-verb (IAV) position for focused elements. A case of the latter comes from Aghem (Watters 1979; Hyman 1984). The sentence in (32a) shows the 'neutral' word order S AUX V O ADV:

(32) a. énáʔ mɔ̀ zɨ̀ kɨ́-bɛ́ ↓nɛ̀ 'Inah ate fufu today'
Inah PAST$_1$ eat fufu today

b. énáʔ mɔ̀ zɨ̀ né ↓bɛ́ ↓kɔ́ 'Inah ate fufu TODAY'
Inah PAST$_1$ eat today fufu DET

c. à mɔ̀ zɨ̀ énáʔ bɛ́ ↓kɔ́ nɛ́ 'INAH ate fufu today'
ES PAST$_1$ eat Inah fufu DET today

Example (32b) shows that when informational or contrastive focus is placed on the adverb *né* 'today', it appears in the IAV position that would otherwise be occupied by the direct object. Similarly, when the subject is in focus in (32c), it too appears in the IAV position, with an expletive subject *à* holding its place. WH-elements also normally go in the IAV position, as expected, as do other constituents of the sentence, particularly when they are singled out for exclusive focused information.

The above examples are intended to illustrate the similarities involved in quite different domains when 'substance' becomes grammaticalized as 'form': The phonologization of phonetics and the syntacticization of pragmatics are exactly parallel. Interestingly, reinforcement of a paradigmatic contrast, which has been assumed in enhancement versions of phonologization and transphonologization, does not seem applicable here. When the grammar requires a subject to be definite or a focused element to appear in the superficial object slot, there is the suppression of a paradigmatic contrast in the one case (subjects no longer contrast in definiteness) vs. the establishment of a syntagmatic contrast in the other. (To simplify considerably, an element in the IAV is in a privileged position vis-à-vis other elements in the sentence. For recent statements on the IAV and focus in Aghem, see Hyman 2010b and Hyman and Polinsky 2009.)

Having established that phonologization bears resemblance to Givón's syntacticization, it seems reasonable to incorporate it under the general heading of grammaticalization. In (33) I have added phonologization at the bottom of the list of the common linguistic effects of grammaticalization presented by Heine et al. (1991: 213):

(33)	Semantic	Concrete meaning	>	Abstract Meaning
		Lexical content	>	Grammatical content
	Pragmatic	**Pragmatic function**	>	**Syntactic function**
		Low text frequency	>	High text frequency
	Morphological	Free form	>	Clitic
		Clitic	>	Bound form
		Compounding	>	Derivation
		Derivation	>	Inflection
	Phonological	Full form	>	Reduced form
		Reduced form	>	Loss in segmental status
	ADD:	**Phonetic substance**	>	**Phonological form**

Although Heine et al. see phonologization as an accompanying reduction or 'erosion' following on the heels of the other effects of grammaticalization, phonologization meets the literal definition of grammaticalization: Something which is not grammar (phonetics) BECOMES grammar (phonology). It seems appropriate, therefore, to recognize parallels such as in (29) and adopt phonologization as one of grammaticalization's 'movements toward structure' (Hopper 1987: 148).

1.4 Conclusion

In the preceding sections I have established that phonologization need not involve contrast, nor even be limited to cases where something phonetic becomes phonological. Taken literally to mean 'the processes by which phonology comes into being', phonologization becomes one branch of the more general phenomenon of grammaticalization: 'the processes by which grammar comes into being', i.e. Hopper's 'movements toward structure'. Unfortunately this is not the usual meaning of 'grammaticalization', which often refers to the historical development of grammatical morphemes: 'Grammaticalization consists in the increase of the range of a morpheme advancing from a lexical to a grammatical or from a less grammatical to a more grammatical status.' (Kuryłowicz 1965 [1972: 52] cited by Heine et al. 1991: 3). Thus, the linguistic effects of grammaticalization indicated above in (32) mostly have to do with what happens when a lexical morpheme (e.g. a word) becomes a grammatical morpheme (e.g. an enclitic or affix). In my use of the term, grammaticalization refers more generally to the development of any aspect or component of grammar (syntax, morphology, phonology).

This is but one of two terminological problems. The first is that there is no generally accepted term meaning 'conversion of substance to form'. While 'grammaticalization' would have been an excellent and transparent choice, it has been preempted for specific phenomena, namely, the creation of grammatical morphemes. Other terms I have heard are either inexplicit or awkward, e.g. codification, coding strategies, linguistification, grammatogenesis, movements toward structure. The second terminological problem is that terms such as phonologization, grammaticalization, syntacticization, lexicalization, etc. are potentially ambiguous, since they only indicate the end product, not the source. This issue arose in the discussion in section 1.3.1 of whether the possible development of phonology from non-phonetic sources should be included under phonologization. As has been pointed out by others, alternative terminology might instead refer to the source, hence dephoneticization, dephonologization, demorphologization, etc. (Dressler 1985; Janda 2003; Joseph and Janda 1988).

I would like therefore to conclude by making the following modest and totally impractical proposals: (i) We should create terms which indicate both the input and the output of the process. (ii) The input should be indicated by the prefix *de-* (indicating a change in status) or *re-* (indicating a restructuring with the same status). (iii) The output should be indicated by a prefix placed on the base *-grammaticalization* (or *-grammatogenesis*?). (iv) *Grammaticalization* should be taken to mean that the output is grammar, whether phonology, morphology, or syntax. With these proposals, a systematic terminology of a catalogue of different types of grammaticalization (in the broader sense) might look like (34).

(34)

		Input	*Output*	*Term*
a.	widespread	phonetics	phonology	dephonetic phonogrammaticalization
		phonology	phonology	rephonogrammaticalization
		lexical morpheme	grammatical morpheme	delexical morphogrammaticalization
		grammatical morpheme	grammatical morpheme	remorphogrammaticalization
		syntax	syntax	resyntactogrammaticalization
		pragmatics	syntax	depragmatic syntactogrammaticalization
b.	'sporadic'	grammatical morpheme	lexical morpheme	demorpho-lexicogrammaticalization
		grammatical morpheme	phonemic material	demorpho-phonogrammaticalization

Since the resulting terms are a bit clumsy, perhaps we would refer to them by three-letter codes: DPP, RPG, DLM, RMG, RSG, DPS, DML, and DMP.

Whatever one thinks about the terminological issue, I hope I have established the following:

(35) a. phonology is grammar; therefore:
b. phonologization is grammaticalization
c. as with other aspects of grammaticalization, one can have greater interest in…
   i. the beginning point (articulatory, perceptual, conceptual) to determine what is or isn't available for phonologization, how, and why (Hombert 1977; Moreton 2008a, b; Yu 2011)
   ii. the end point (phonology), e.g. how the structured version ultimately diverges from the phonetics
   iii. the diachronic correspondences between the beginning and end points
   iv. the logical or actual stages of the changes in input/output, their diffusion, social significance, etc.
d. there is overlap and unclarity as to where phonetics ends and phonology begins
e. however, there is a difference between phonetics (substance) and phonology (form), just as there is a difference between pragmatics (substance) and syntax (form)

Much of the interest in phonologization (and Heine et al.'s notion of grammaticalization) has been in determining the nature of the substance that underlies grammar. This has led certain scholars to seek ways of reducing phonology to phonetics and morphology/syntax to semantics and pragmatics. While no one can deny such relationships, establishing the sources of grammar is only part of the story. The rest has to do with why the intrinsic phonetic, semantic, and pragmatic properties do not *remain* intrinsic rather than becoming structured within the grammar. This in turn reduces to the question of why there is grammar at all. On the one hand grammar necessarily underspecifies the substantive sources: a language cannot provide a structural analogue for every aspect of phonetic naturalness, semantic transparency, or pragmatic coherence. What it does do is impose strictly formal linguistic structures which take over from where the extralinguistic sources leave off. A full account must therefore be concerned with both the beginning and endpoints of phonologization (and, more generally, grammaticalization), and ultimately recognize that phonologies/grammars have properties that are not reducible to the natural tendencies in speech and communication:

> … it is necessary to assume a considerable degree of independence between linguistic principles proper and the principles that obtain in those extralinguistic domains that appear to underlie them. (Anderson 1981: 496)

> … the concerns of Grammar… are not derivable from extragrammatical factors. (Hyman 1984: 71)

Or, as I like to put it, Grammar has a mind of its own.