

The OCP in the perception grammar

Paul Boersma

University of Amsterdam

paul.boersma@hum.uva.nl, <http://www.fon.hum.uva.nl/paul/>

December 29, 2000

An important concept in Optimality Theory (OT) is *input-output faithfulness*, the phonological similarity between underlying forms and surface forms.¹ McCarthy & Prince (1995) distinguish three main types of families of input-output faithfulness constraints: MAX-IO, which states that an element (e.g. segment or feature) that is present in the underlying form should have a corresponding element in the surface form; DEP-IO, which states that an element present in the surface form should have a correspondent in the underlying form; and IDENT-IO, which states that if an element in the underlying form and an element in the surface form correspond to each other, then these two elements should be identical. A remarkable (though inconspicuous) property of these constraint families is their symmetry: MAX-IO and DEP-IO are one another's mirror images, and the formulation of IDENT-IO requires the *commensurability* of underlying form and surface form, that is, in order that the underlying and the surface element can be assessed as identical or not, they should have a common basis for comparison, which means that they should be expressed in the same kind of units.

If the underlying form and the surface form shall be expressed in comparable units, the question arises what the nature of these units is. I will share here the common view that the underlying phonological form of an utterance is basically an aggregate of lexical forms, augmented with information about morphological boundaries. These lexical forms are *economical representations*, which means that they are encoded in abstract features without much redundant information, disregarding most of what can be added by rule, such as positional variants, syllabification, or metrical structure. This *lexical economy* has been the one of the drives after proposals for contrastive underspecification of phonological features (Chomsky & Halle 1968, Steriade 1987), and in its extreme form of *lexical minimality*, it has inspired proposals for privative features (Trubetzkoy 1939), feature-tree underspecification (Jakobson, Cherry & Halle 1953), and radical underspecification (Archangeli 1984, 1988). But while there has been a heated debate about its extent, the consensus was that some sort of economy existed, implying at least, for instance, that predictable positional variants were not encoded in the lexicon. This situation appeared to change with the Optimality-Theoretic maxim of *richness of the base* (Prince & Smolensky 1993:191), which states that the input to the grammar is not restricted and that the only formal restrictions are imposed by grammars on their outputs. At first sight, this contradicts lexical economy, but in the three-grammar model defended in the current article, the lexicon is itself the output of the perception and recognition grammars, and the restrictions imposed by these grammars will be directly reflected in lexical forms (§3.6). Thus, lexical economy and richness of the base are not incompatible,

¹ I will use traditional terms like “underlying form” and “surface form” rather than OT-specific terms like “input” and “output”, because “input” means different things when applied to a speaker, to a listener, or to a learner.

and it appears to be a legitimate goal even within an OT framework to strive for efficient coding in the lexicon.

The idea of commensurability of underlying and surface forms is at odds with the early generative view (Chomsky & Halle 1968) that underlying forms were economical but surface forms were phonetically detailed. In a derivational framework, this did not matter, but in an OT framework the implications would be severe, and McCarthy & Prince (1995) chose to ignore the issue, implicitly relegating, perhaps, the handling of phonetic detail to a separate module of phonetic implementation. In OT, a detailed surface form has been proposed by some of those who want to integrate phonetic principles into phonological theory. With a ‘perceptual’ interpretation of faithfulness, namely that an underlying feature is faithfully produced if it can be heard in the output, Jun (1995) and Steriade (1995:ch.1:fn.10) realize that faithfulness constraints must basically be ranked by the ease of *perceptual recoverability* of these features. However, they had to implement this view by providing the (production) grammar with faithfulness-like constraints that favour the presence of specific acoustic cues in the output. According to this UCLA school of phonetically-driven phonology (also Flemming 1995; Hayes 1996, 1999; and Kirchner 1998), phonological surface forms must be *rich representations* that contain information about single acoustic cues. But since the grammar contains faithfulness-like constraints, its input must contain rich representations as well. This means that if the grammar is regarded as a single production grammar combining phonology and phonetic implementation, then the lexical forms, which are the input to this grammar, must be rich representations as well, strikingly violating lexical economy. The alternative, of course, is that the production grammar consists of two sequentially ordered modules and that the theory of phonetically-driven phonology only models the second of these modules, which can perhaps be identified with phonetic implementation.

The current paper seeks to bridge the gap between these two apparently conflicting viewpoints, by taking seriously the speech perception process, i.e. by making a principled distinction between *acoustic* (rich) representations and *perceptual* (economical) representations. I will model the highly language-dependent process of perceptual recovery as a full-fledged Optimality-Theoretic *perception grammar* that is aimed at constructing economical perceptual representations from rich acoustic input. When applied to the speaker’s own utterance, this grammar produces the phonological surface form, which is therefore a economical representation, so that faithfulness constraints can be evaluated on two minimal abstract representations, as desired. Several kinds of rich representations (the utterance as a set of lower-level perceptual forms) are intermediate forms in the perception process. These forms do not play any role in evaluating faithfulness, but they do play a role in learning, so that this model manages to incorporate phonetic principles in phonology, as desired. The phonologically most interesting result of the assumption of a perception grammar is that a general highly language-dependent process of *sequential abstraction*, which is performed by this perception grammar, accounts for all the phenomena traditionally ascribed to the *Obligatory Contour Principle* (OCP), i.e. the autosegmental principle that states that adjacent identical elements are forbidden, especially within morphemes (Goldsmith 1976, McCarthy 1986, 1988). As a corollary, I will show that previous attempts to incorporate the OCP instead in an Optimality-Theoretic *production* grammar lead to undesirable ambiguities.

1 Example: faithfulness as recoverability of an underlying form

As an example for a discussion about the required grammar model, I will consider the English partial utterance *it told*, which was chosen because it has two adjacent identical underlying segments.

1.1 The underlying form

The English utterance *it told* consists of a concatenation of the lexical items (word-like morphemes) |t| ‘it’ and |told| ‘told’, so its underlying phonological form is something like

- (1) *The underlying form*
 |# ɪ·t # t·o·l·d #|

where the number signs (#) delimit the boundaries between the lexical items, and the centred periods (·) explicitly indicate that the two surrounding segments belong to the same lexical item. The representation (1) is a readable shorthand for a fuller description in terms of abstract feature bundles or parallel tiers.

The general lack of direct observability of underlying forms has led to a wide diversity of proposals for what such a more atomic description should look like. If the tense-lax distinction plays any role in English phonology or lexicalization, the symbol |ɪ| will have a value of [lax] on the tenseness tier and a value of [high] on the vowel height tier, whereas if the tense-lax distinction does not play a role in English, no tenseness tier will be used, and the symbol |ɪ| will only have a value of [higher mid] on the vowel height tier. It is uncontroversial that the entities |t| and |d| have something in common but differ in another respect. The thing they have in common can be expressed as the value [alveolar] on the place tier, or as a more hierarchical combination [coronal, –distributed] if |t| and |d| happen to pattern together with [+distributed] coronals in English. The respect in which |t| and |d| contrast must be the same thing that distinguishes |p| from |b| and |k| from |g|, and we can call this voiceless-voiced, tense-lax, or fortis-lenis, depending on how well this abstract contrast generalizes to other pairs of English consonants. In general, the preferred analysis of an underlying form will depend on the behaviour of these elements in the language at hand, and on the linguist’s preferences with respect to binarity, underspecification, and feature geometry. Fortunately, I will not have to take a stance on these complicated issues in the current article, since this article will focus on sequential structure and therefore only address the other controversial issues in autosegmental phonology, namely locality, line crossing, and the OCP, of which the form |# ɪ·t # t·o·l·d #| contains several cases.

The form in (1) is economical in several respects. It abstracts away from the difference in pronunciation between word-final and word-initial |t|, it ignores the positional variation of |l| (dark or light), and it has a single symbol |o| for something that is usually pronounced as a diphthong. In short, (1) represents a concatenation of forms as they are stored economically in the lexicon. The morphological make-up is made explicit by the use of the “#” symbol, which here stands for a non-inflectional morpheme boundary, and by the use of the “·” symbol, which tells us that the constituent forms |t| and |told| are indivisible lexical items.

The question, now, is what the surface representation of this form is. In an Optimality-Theoretic framework, this question is highly relevant because we will want to evaluate faithfulness constraints as a comparison between (1) and the surface form.

1.2 First surface form: the articulatory form

In early generative phonology (Chomsky & Halle 1968), the surface form is a detailed phonetic representation, derived from the underlying form by a single stratum of serially ordered rules. We could write such a surface form as a detailed articulatory transcription, which we can simplify with traditional relatively language-independent IPA symbols:

(2) *The articulatory form*

[t̚tː^həʊːt̚d]

This representation is a linearized shorthand for a fuller multi-tiered description in terms of three tongue-tip gestures, a spreading gesture of the glottis, a lip rounding gesture, a lung compression gesture, and more. The notation (2) takes into account the fusion of heteromorphemic identical plosives into a single long closure, the post-aspiration of initial [t], the realization of [o] as a smoothly gliding diphthong, vowel lengthening before “voiced” consonant clusters, velarization of final [l], and devoicing of final lenis plosives.

The question, now, is whether (2) is a form whose similarity to the underlying form (1) can be evaluated by faithfulness constraints. The answer must be negative. The underlying form contains cognitive symbols whose task it is to represent the two morphemes in the lexicon in an economical way, whereas the articulatory form (2) represents a prescription for muscle movements. So the forms (1) and (2) simply have no common basis for comparison. Any visual similarity between parts of (1) and parts of (2) is purely coincidental: both of the mappings |t| → [t̚tː^h] and |o| → [əʊː], for instance, must be regarded as equally complicated mappings from an arbitrary cognitive symbol to an articulatory configuration.

1.3 Second surface form: the acoustic form

A better candidate for an evaluable surface form is the detailed acoustic phonetic representation, which we can again simplify by using IPA-like symbols:

(3) *The acoustic form*

[[t̚ t̚ː t̚^h əʊː t̚ d̚]]

This representation is a linearized shorthand for a fuller multi-tiered description in terms of acoustic events: [[t̚]] is a vowel (loud periodic sound) with lower-mid F1, [[t̚ː]] is a formant transition with alveolar place, [[t̚ː]] is a long silence, [[t̚]] is a fortis alveolar release burst, [[t̚^h]] is a short aspiration noise, [[əʊː]] is an overlong vowel with low-falling F1, [[t̚]] is the sound of a velarized alveolar lateral approximant, [[_]] is a short silence, and [[d̚]] is a lenis alveolar release burst (note: for reasons of space, several of these terms already refer to nonperipheral perception). Therefore, the notation (3) is visually quite similar to (2), but could hardly be more different in content. The deceptive similarity is entirely due to the pragmatic hybrid philosophy behind the International Phonetic Alphabet, in which most symbols and features are meant to have articulatory as well as acoustic correlates.

The question is whether (3) can perform in a faithfulness evaluation. Again, the answer must be negative: to assess the identity of (3) to (1), we would have to make the eleven parts of (3) correspond to the seven parts of (1) and evaluate their similarity, and this is not straightforwardly possible.

structure might be $/\# \underset{-\text{voi}}{\overset{+\text{plos}}{\text{I}}}/$, so that the lexical item |It| ‘it’ is activated, but |Iz| ‘is’ is not. For reasons of space, however, I will take a more conservative approach here and assume that the listener generally stores diacritics until she has an opportunity to resolve them, i.e. that she just has $/\# \text{I}_{-\text{leng}}/$ at this point.

- c. The sequence $[[\text{t}^{\text{h}} _ : \text{t}^{\text{h}}]]$ of transition, long silence, fortis burst, and aspiration must signal two separate underlying segments, since English has no underlying geminate segments. The first of these must be an alveolar plosive, but since the only local cue is the ambiguous unreleased transition $[[\text{t}^{\text{h}}]]$, we don’t know yet whether it is voiced or not (actually, we might conjecture voicelessness on the basis of the diacritic of the already recovered previous vowel, but I will be conservative again and take one simple step at a time), so let us denote this segment as an underspecified $/\widehat{\text{dt}}/$. The second consonant has a fortis release burst and is aspirated, so it has two cues that together reliably signal its voicelessness in English: $/\text{t}_{\text{asp}}/$. However, aspiration in English may indicate something about word boundaries, so we keep it as a diacritic for potential later use: $/\text{t}_{\text{asp}}/$. The recovered structure is now $/\# \text{I}_{-\text{leng}} \widehat{\text{dt}} \text{t}_{\text{asp}}/$. In this form, the spaces between the segments indicate that we do not yet know whether consecutive segments belong to the same word or whether there is a word boundary between them.
- d. Add a morpheme boundary between any two identical adjacent consonants, even if they differ in voicing: $/\widehat{\text{dt}} \text{t}_{\text{asp}}/ \rightarrow / \widehat{\text{dt}} \# \text{t}_{\text{asp}}/$. This step is reasonable because no English word-like morpheme contains a sequence of identical adjacent consonants or of adjacent consonants that differ only in voicing. The recovered structure is now $/\# \text{I}_{-\text{leng}} \widehat{\text{dt}} \# \text{t}_{\text{asp}}/$.
- e. Now that it has become word-final, put the consonant $/\widehat{\text{dt}}/$ into the same word as the previous vowel: $/\# \text{I}_{-\text{leng}} \widehat{\text{dt}} \# \text{t}_{\text{asp}}/$. This step is reasonable because no English word consists of a single consonant. Now that the listener has segmented out a complete word $/\# \text{I}_{-\text{leng}} \widehat{\text{dt}} \#/$, she can restrict the search space for the recognition process to just |Id| ‘id’ and |It| ‘it’.
- f. Now that we know that $/\widehat{\text{dt}}/$ is in the same morpheme as the preceding vowel $/\text{I}_{-\text{leng}}/$, the voicing of $/\widehat{\text{dt}}/$ can be resolved by the $[-\text{leng}]$ diacritic. The absence of lengthening, perhaps aided by the presence of a glottal stop, will now select the voiceless variant $/\text{t}/$, so that the recovered structure becomes $/\# \text{I}_{-\text{leng}} \cdot \text{t} \# \text{t}_{\text{asp}}/$. The listener has no choice now but to recognize the lexical item |It| ‘it’.
- g. Now that the lengthening diacritic has been useful in the previous step, it can be discarded, leaving $/\# \text{I} \cdot \text{t} \# \text{t}_{\text{asp}}/$ as the recovered structure.
- h. Map a long vowel with low-falling F1 (i.e. $[[\widehat{\text{əu}}]]$) to the segment $/\text{o}/$, marking whether it is lengthened or not. Now that we have classified $[[\widehat{\text{əu}}]]$ as $/\text{o}/$, we are capable of noting that the vowel is overlong, so we get $/\text{o}_{+\text{leng}}/$, and the recovered structure becomes $/\# \text{I} \cdot \text{t} \# \text{t}_{\text{asp}} \text{o}_{+\text{leng}}/$.
- i. Morphemically combine the aspirated plosive with the next vowel: $/\# \text{I} \cdot \text{t} \# \text{t}_{\text{asp}} \cdot \text{o}_{+\text{leng}}/$. This step is reasonable because English plosives are unaspirated across word-like morpheme boundaries.
- j. Now that the aspiration diacritic has been useful in morphemic combination, the English listener can discard it, leaving $/\# \text{I} \cdot \text{t} \# \text{t} \cdot \text{o}_{+\text{leng}}/$.

- k. Map any lateral approximant to the segment /l/. Since velarization can help in finding English morpheme boundaries, we mark the darkness by keeping it as a diacritic: /# ɪ·t # t·o_{+leng} l_{dark}/.²
- l. Morphemically combine the dark lateral with the previous vowel: /# ɪ·t # t·o_{+leng}·l_{dark}/. The step is reasonable because word-initial laterals are always light in English. At this point, the recognition system will already be trying to interpret the /tol/ part as the beginning of a lexical item. If the word |t| has already been identified as a pronoun, the syntactic part of the recognition system may already have narrowed down the lexical search to the verb form |told| ‘told’ at the expense of the phonologically better-fitting noun |tol| ‘toll’. This means that recognition may have succeeded before the end of the utterance has been processed.
- m. Map the sequence of silence and burst [[_ d̥]] to a plosive. The burst is lenis, which suggests a voiced plosive /d/, but other acoustic cues like a possible non-lengthening of a preceding tautomorphemic vowel may override this burst strength cue later, so we again underspecify the voicing of the plosive and keep the burst strength as a diacritic: /# ɪ·t # t·o_{+leng}·l_{dark} d̄_{lenis}/.
- n. Add a word boundary at the end of the utterance: /# ɪ·t # t·o_{+leng}·l_{dark} d̄_{lenis} #/. This step is reasonable, because English utterances tend to end in a word boundary.
- o. Morphemically combine the word-final consonant /dt/ with what comes before it: /# ɪ·t # t·o_{+leng}·l_{dark}·d̄_{lenis} #/. As before, this step is reasonable because no English word-like morphemes consist of a consonant only. At this point, the recognition system will be left with a single candidate for the second word, namely |told| ‘told’, even if it had not managed to do so in step (4l). Although the voicing of the final plosive has not been resolved yet, there is but one candidate left, because the English lexicon does not contain an item with the phonological form |tolt|. Again, this means that although the output of the perception system is the input to the recognition system, the recognition system may be finished with the utterance before the perception system is.
- p. Now that the lateral is followed by a tautomorphemic consonant, discard the darkness diacritic of the lateral: /# ɪ·t # t·o_{+leng}·l·d̄_{lenis} #/. This step is reasonable because the darkness diacritic of the lateral is now redundant, since all non-prevocalic laterals are dark in English.
- q. Now that we know that the preceding vowel /o_{+leng}/ is in the same morpheme, determine the voicing of /d̄_{lenis}/ on the basis of the lenis burst and the lengthening of the vowel. Both diacritics point to the voiced variant, so that we do not have to know whether vowel lengthening or burst strength is the primary acoustic cue: /# ɪ·t # t·o_{+leng}·l·d #/. We can imagine that the local cue (lenis burst) has been dissolved in this step.
- r. Now that the lengthening diacritic has been useful in determining obstruent voicing, discard it: /# ɪ·t # t·o·l·d #/.

We call the form resulting from the process in (4) the *perceptual form*, for reasons explained in §2:

² If the lateral approximant had been light, we would have marked it as /l_{light}/. Note that though the symbol /l/ looks similar to the acoustic notation of a light lateral ([[l]]), the symbol /l/ does not imply lightness.

(5) *The perceptual form*

/ʃ ɪ·t # t·o·l·d #/

This form, finally, is commensurable with the underlying form: like (1), it is a minimal representation, consisting of abstract segments and morphological boundaries.

1.5 Satisfaction and violation of faithfulness in production

I propose, then, that the perceptual form is the true phonological surface form, and that it is the form evaluated by faithfulness constraints in an Optimality-Theoretic production grammar. To see how this works, consider the underlying form |ʃ ɪ·t # t·o·l·d #| again. One of the output candidates is the articulatory form [ɛ̞tː^həʊːt̚d̚], which, as we have seen, gives rise to the perceptual form /ʃ ɪ·t # t·o·l·d #/, which is identical to the underlying form and therefore violates no faithfulness constraints at all. But the most faithful candidate does not always win; in general, there will be competing constraints that disfavour certain aspects of the output form. In this case, the articulatory output form [ɛ̞tː^həʊːt̚d̚] contains a long tongue-tip closure gesture, which is articulatorily more difficult than a shorter closure gesture. An articulatorily easier candidate, therefore, is [ɛ̞t^həʊːt̚d̚]. According to the procedure in (4), this form would give rise to the perceptual form /ʃ ɪ t·o·l·d #/, with a space denoting ambiguity as to whether the parts /ɪ/ and /told/ are tautomorphemic or not (assuming British English, in which words may end in |ɪ|). The form /ʃ ɪ t·o·l·d #/ does violate some faithfulness constraints: one of the two underlying |t| has not been recovered, and the underlying morpheme boundary has been lost. The two candidates are evaluated in the following tableau:

(6) *The production of English*

ʃ ɪ·t # t·o·l·d #	*DELETE (t)	*DELETE (#)	*GESTURE (tongue tip: close & open / long)	*GESTURE (tongue tip: close & open / short)
 [ɛ̞tː ^h əʊːt̚d̚] ⇒ [[ɛ̞ tː _ : t ^h əʊː t̚ _ d̚]] → /ʃ ɪ·t # t·o·l·d #/			*	*
[ɛ̞t ^h əʊːt̚d̚] ⇒ [[ɛ̞ tː _ t ^h əʊː t̚ _ d̚]] → /ʃ ɪ t·o·l·d #/	*!	*		**

In this tableau, each candidate consists of three representations: the articulatory form; the acoustic form, which is derived from the articulatory form in a language-independent way by the laws of physics and peripheral auditory mechanisms, as denoted by the double arrow; and the perceptual form, which is construed from the acoustic form by the language-specific process in (4), as denoted by the single arrow. Some of the violation marks in (6) are shown at the top of their cells, some at the bottom, as an indication of the forms that they evaluate. The articulatory constraints (*GESTURE) evaluate the articulatory forms, and their violation marks therefore appear at the top of their cells. In

choosing these constraints, I follow Boersma (1998), where they are regarded as rankable by the amount of articulatory effort.³ The faithfulness constraints (*DELETE) evaluate the similarity of the perceptual forms to the underlying form, and their violation marks therefore appear at the bottom of their cells. The constraint *DELETE (x), following again Boersma (1998), is violated if an x in the underlying form does not occur in the surface form (this constitutes a violation of TRANSMIT (x)), or if it occurs in the surface form with a different value (this constitutes a violation of *REPLACE (x, y)). The constraints TRANSMIT and *REPLACE are roughly the perceptual counterparts of McCarthy & Prince's MAX-IO and IDENT-IO, although *DELETE expresses recoverability whereas MAX-IO expresses correspondence (see §6.2 for the crucial difference), and although *REPLACE, unlike IDENT-IO, has two arguments x and y , in order to take into account the rankability of this constraint by the perceptual distance between the thing replaced (x) and its substitute (y). The distinction between the roles of articulatory and perceptual representations in phonology was made explicit in Boersma (1989), who proposed a strict preference ranking of articulatory forms by amount of articulatory effort, and a strict ranking of pairs of perceptual forms by amount of perceptual confusability. This idea was translated to an Optimality-Theoretic framework by Flemming (1995), who decided, however, to ignore the perception process and to return to detailed acoustic representations like (3), i.e. the middle forms in tableau (6).

Tableau (6) expresses clearly the conflict between speaker-centred and listener-oriented constraints in speech production: the emergence of the form [ɛ̃tː^həʊːt̪] is due to the fact that its higher articulatory effort is more than outweighed by the need to enable the listener to recover the morphological and phonological make-up of the underlying form. The constraints in tableau (6) predict a typology. First, we must note that the *local ranking principle* for articulatory constraints (Boersma 1998:158) ranks the two gestural constraints in a fixed order: since they refer to comparable gestures, the amount of articulatory effort (larger for longer closures) defines a harmonic scale in the sense of Prince & Smolensky (1993:67–68). This leaves a three-way typology: if either of the faithfulness constraints outranks the anti-long-consonant constraint, the first form wins; if the highest faithfulness constraint is ranked between the two gestural constraints, the degeminated form wins, as would be usual in Dutch; if both faithfulness constraints are ranked below both gestural constraints, there is deletion of the entire alveolar gesture, and a debuccalized articulatory form like [ɛ̃^həʊːt̪].⁴

The main difference of the current proposal with almost every other theory of phonology is the existence of the process exemplified in (4), which converts a raw acoustic form into a discrete structure with language-specific elements. I will call the module that performs this process the *perception grammar*.

Conclusion of chapter 1:

The phonological surface form involved in the evaluation of faithfulness constraints in an Optimality-Theoretic production grammar is best modelled as an economical lexical-like representation constructed from the acoustic signal by a language-dependent process of perception.

³ Alternatively, Kirchner (1998) defends a single constraint LAZY, whose number of violation marks depends on articulatory effort. See next footnote for an empirical difference.

⁴ Kirchner's (1998) single LAZY constraint would not generate this three-way typology.

2 Why call this perception?

Perception is the construction of an abstract mental image from raw sensory material. This is the usual understanding of this term in the cognitive sciences (Powers 1973, Pinker 1997). In the realm of speech, I will take perception to be the construction of a discrete phonological structure from raw acoustic material. According to most psycholinguistic theories of language comprehension, this perception process feeds the process of lexical access, which we can call the *recognition* process.

2.1 What is being mapped on what?

The perception process maps raw acoustic data on more abstract representations, and it maps these more abstract representations on representations that are still more abstract. For example, the basilar spectrum, which is the inner ear's representation of the frequency spectrum of the incoming sound, is mapped on the first formant (F1), which corresponds to the lowest strong peak in the basilar spectrum, and the first formant is again mapped on the abstract feature of vowel height. The mapping from basilar spectrum to F1 is from a multidimensional continuous representation to a one-dimensional continuum, and the mapping of F1 to vowel height is from a one-dimensional continuum to a language-dependent discrete representation.

While the input to the perception process is uncontroversially acoustic, there has been considerable debate as to what the output of the perception process is. It could be phonetic features (Eimas & Corbit 1973), phonological features (Lahiri & Marslen-Wilson 1991), allophones (Wickelgren 1969), phonemes (Foss & Blanck 1980; Norris & Cutler 1988), syllables (Foss & Swinney 1973; Cole & Scott 1974; Segui, Frauenfelder & Mehler 1981; Mehler, Dommergues, Frauenfelder & Segui 1981; Dupoux & Mehler 1990), and they could contain sounds (Diehl & Kluender 1989; Jusczyk, Houston & Goodman 1998) or gestures (Lieberman & Mattingly 1985, Fowler 1986). Evidence has been adduced for all of these units, and it seems reasonable to assume that indeed all of these units are perceived at some level of abstraction (Pisoni & Luce 1987).

Features and segments may well be perceived at the same time, since the various levels of abstraction do not conflict (if we see a car, we may also notice its colour and see a couple of its wheels at the same time). For instance, the vowels /e/ and /o/ go separate ways in many languages, which is evidence for their segmenthood, and at the same time they tend to be involved together in the same process (e.g. diphthongization), which is evidence for the existence of a common feature (vowel height).

Some phonological features seem to be based on articulation, some on acoustics, and evidence for this comes from their behaviour in phonological processes. For instance, the perceptual features of *place* and *nasality* have to be derived in a non-obvious way from spectral information, and since these features define natural classes in phonological processes, it seems reasonable to assume that their definition is articulatory in origin, i.e. *place* basically refers to constriction site and *nasality* to a lowered velum (this does not mean that constriction site and velum height can always be directly perceived: if there is no airflow in my vocal tract, you cannot perceive the height of my velum, so articulatory features will always be different from perceptual features). On the other hand, many features are solely based on acoustic properties. Fricatives, for instance, tend to act as a natural class in phonological processes, so that it is desirable to give them a common feature in the lexical representation; however, the fricatives at various places of articulation are produced with widely different articulatory gestures, so their perceptual

similarity is mainly based on their acoustic similarity. Likewise, the free variation between the alveolar trill and the uvular trill in Dutch must be based on their acoustic similarity only, since their articulations differ widely.

Thus, perception is an abstraction away from acoustic features: it merges simultaneous features into more segment-like units, and merges sequences of lower-level units into larger units. This latter *sequential abstraction* is the main focus of this paper.

2.2 Sequential abstraction

If speech perception is like other kinds of perception, the mental construction of higher-order units depends on the frequency of co-occurrence of its parts. Common co-occurrence of smaller units in temporal sequence will lead to sequential abstraction. The most striking example of this is the perception of intervocalic plosives. While the centre of the sequence $[[a t^{\prime} _ t a]]$ consists of three rather different acoustic events, namely a transition, a silence, and a release burst, these three will be perceived as a single segment, so that the whole sequence becomes /ata/. Since intervocalic plosives are abundant in almost every language, the sheer necessity of a coronal release burst following a coronal transition, together sandwiching a silence, will lead to this perceptual merger in nearly all languages; evidence for this comes from the fact that /t/ tends to act as a unit in phonological processes.

Several examples of sequential abstraction can also be seen in (4). In (4c), the sequence of burst + aspiration is combined into a single segment, because this aspiration can only follow voiceless release bursts in English. Other examples in (4) involve the combination of consecutive segments into morphemic units, or indeed the insertion of morpheme boundaries, which is an anti-abstraction measure. To (4) could have been added some steps that build prosodic structure, in this case the two syllables that can be perceived in *it told*. Syllables will be organized around sonority peaks, like height differences in the landscape are perceived as a series of mountains. In the perception process, they could constitute an intermediate construct, helping in finding word boundaries, especially if the language has stress on a fixed syllable position. I will not discuss in detail the merger of segments into syllables, nor stress, since the focus of this article is on segmental phenomena.

The empirical subject of this article will be the mapping from long consonants to single or paired geminates, the mapping from sequences of nasalized vowels to polysyllabic nasality, and the mapping of sequences of high- and low-toned vowels to polysyllabic tones. In all these three cases, the input is already a discrete representation, quite well abstracted from the acoustic signal, and the output is an even more abstract discrete representation.

Conclusion of chapter 2:

The perception process maps raw sensory data onto discrete structures.

3 Why describe perception with a grammar?

The term *grammar* is used in generative linguistics to denote a description of the set of possible language forms (utterances). In phonological theory, however, the term tends to be used in the sense of a language-dependent recipe for the conversion of one form into

another, and I will understand the term in that sense here. In the rest of this chapter, I will show that the perception grammar meets both of the requirements for naming something a ‘grammar’: the formal similarity to what we know as a grammar, and the language-dependence.

3.1 The perception grammar as a formal grammar

The recipe in (4) represents a partial account of the perception process for English. Its form is reminiscent of a grammar with serially ordered rules, which we can see clearly if we regard the process as a construction of four increasingly higher levels of abstraction, with each higher level using information only from the previous level, not from any deeper levels:

(7) *Levels of abstraction*

$$\sigma_1 = [[\text{t}^{\text{h}} \text{ } _ : \text{t}^{\text{h}} \text{ } _ \text{ } \text{d}]]$$

$$\sigma_2 = / \text{I}_{\text{leng}} \widehat{\text{dt}} \text{t}_{\text{asp}} \text{O}_{\text{leng}} \text{l}_{\text{dark}} \widehat{\text{dt}}_{\text{lenis}} /$$

$$\sigma_3 = / \# \text{I}_{\text{leng}} \widehat{\text{dt}} \# \text{t}_{\text{asp}} \cdot \text{O}_{\text{leng}} \cdot \text{l}_{\text{dark}} \cdot \widehat{\text{dt}}_{\text{lenis}} \# /$$

$$\sigma_4 = / \# \text{I}_{\text{leng}} \cdot \text{t} \# \text{t}_{\text{asp}} \cdot \text{O}_{\text{leng}} \cdot \text{l}_{\text{dark}} \cdot \text{d}_{\text{lenis}} \# /$$

$$\sigma_5 = / \# \text{I} \cdot \text{t} \# \text{t} \cdot \text{O} \cdot \text{l} \cdot \text{d} \# /$$

The four steps in (7) summarize the perception process in (4) by abstracting away from the temporal order that was explicit in (4). Without the temporal factor, we can regard the process as a sequence of mappings between five consecutive levels. The first mapping (from σ_1 to σ_2) succeeds in *segmentation*, i.e. it manages to divide up the acoustic signal into the six segments that it represents. The first mapping also makes the first big advance in *categorization*, i.e. destroys a lot of acoustic information, namely all the continuous information about spectral properties and timing, and thereby has the virtue of creating a representation that is already as discrete (though not yet as economical) as the lexical representation; of the six segments, four are identified completely (i.e. successfully categorized for all features), and the remaining two are identified with respect to all of their features except voicing. Finally, the first step makes some decisions about *sequential abstraction*; for example, it maps the acoustic sequence $[[\text{t}^{\text{h}} \text{ } _ : \text{t}^{\text{h}} \text{ } _]]$ on a sequence of two separate plosives, not on a single geminate. The second mapping (from σ_2 to σ_3) succeeds (in this example) in retrieving the complete morphological make-up of the utterance on purely phonological grounds, i.e. on the basis of the utterance boundaries, the segmental make-up, and the information in the diacritics, without accessing the lexicon. This step is exclusively structure-building, therefore. The third mapping (from σ_3 to σ_4) succeeds in completing the categorization. This step is also structure-building; for example, the mapping from $/\widehat{\text{dt}}_{\text{lenis}}/$ to $/\text{d}_{\text{lenis}}/$ adds the feature value [+voi] to the representation of the plosive. The fourth step, finally, discards all the information of the diacritics because they are irrelevant for lexical access.

From the formal point of view, therefore, the name ‘perception grammar’ seems appropriate. It remains to be seen as what kind of grammar it can most conveniently be modelled: as a sequence of ungrouped rules as in (4), as a sequence of level mappings as in (7), as a structure-building tree, or as an Optimality-Theoretic grammar. Other than what the sequence of forms in (7) suggests, the three intermediate levels presented for clarity in (7) probably have no special status. For example, there seems to be no

unordered mapping between each pair of consecutive levels. The mapping from level σ_2 to σ_3 , for instance, seems to involve an iterative procedure. To show this, I will consider the English word *toast*, which will be perceived on the second level as $/t_{asp} \text{ } o_{-leng} \text{ } s \text{ } dt_{fortis}/$. The insertion of a word boundary at the end first causes the incorporation of $/s/$ into the same morph as $/dt_{fortis}/$, and then causes the incorporation of $/o_{-leng}/$ into that same morph, leading to $/\# t_{asp} \cdot o_{-leng} \cdot s \cdot dt_{fortis} \#/$ on the third level (we now see why *told* cannot be used as an example here: the morphemic connection between $/o_{+leng}/$ and $/l_{dark}/$ could also be derived from the darkness of the lateral). The mapping from σ_2 to σ_3 , then, involves recursion, i.e., the rules in (4) cannot be grouped into a small number of ordered levels of unordered rules. Probably a better way to view (4) is as a set of “as soon as” rules, i.e. each rule can apply as soon as its input conditions are met. For instance, rule (4i) can apply as soon as a sequence of an aspirated plosive and a vowel becomes available. This procedure seems to apply to all the 18 rules in (4), and it means that no extrinsic rule ordering is needed. This situation is reminiscent of the situation in some versions of the Minimalist Program in syntactic theory (Chomsky 1995 et seq.), in which Merge combines constituents into larger units as soon as they become available. As in minimalist syntax, any attempts to draw the process as the construction of a single tree will be a waste of paper, even more so in our case because the process not only introduces concatenation (the “.” symbols), but boundaries (“#”) as well, a situation that prevents intermediate forms from being representable as partial trees. It seems most appropriate, then, to use the string-based formulation of (4). In §4, however, I will abandon this formulation because it cannot handle conflict resolution, and replace it with a formulation in terms of two OT grammars, one for relating σ_1 with σ_2 , the other for relating σ_2 with σ_4 .

3.2 Language dependence of the perception process

We have seen that the perception process is amenable to modelling with grammar-like formalisms, but if the perception process is the same for all languages, the term ‘grammar’ will still be inappropriate for it. All the 18 steps in (4), however, contain the word ‘English’, suggesting that all these steps are language-specific.

Some of the steps may be nearly universal, e.g. the tendency that utterances begin and end at morpheme boundaries.

At the other extreme, the language dependence of categorization is rather uncontroversial. For instance, languages with three vowel heights map incoming higher-mid F1 as well as lower-mid F1 into the same mid-vowel category (e.g. Japanese: Akamatsu 1997), unlike languages with four vowel heights, which map them into a lower-mid and a higher-mid vowel, respectively (e.g. Italian: Leoni, Cutugno & Savy 1995). This is external, i.e. strong, evidence for language-dependent perception, because this evidence has been collected in perception experiments, independently from phonological processes identified by linguists.

The language dependence of the recovery of morphological structure is also clearly language-dependent: in many languages other than English, the darkness of a lateral, the aspiration of a plosive, the lengthening of a vowel, and the length of a silence are not conditioned by morpheme boundaries at all. This evidence is theory-internal, i.e. relatively weak, because it uses the ease of recognition as a criterion for positing elements of the perception process. This article will focus, however, on a rather controversial language dependence in perception, namely that of sequential abstraction, a phenomenon traditionally ascribed to the workings of a universal principle of phonological

representation. Leben (1973) observed in the study of African tone languages that the tone contour in the form [jévésè] must be represented as HL, not as HHL:

(8) *A constraint on representations in suprasegmental phonology*



The reason for positing this single-value representation was the simplicity of phonological analysis, resulting, among other things, from distributional considerations and from the fact that the high tone on [jévé] acts (e.g. deletes) as a unit in phonological processes. Goldsmith (1976), who did not believe in its universal truth, called the constraint against HH (or LL) sequences the Obligatory Contour Principle (OCP). McCarthy (1988) phrased it as “adjacent identical elements are forbidden”. The OCP has been extended to other domains than tone, for instance nasality (Goldsmith 1976) and gemination (McCarthy 1986), and has often be proposed as an inviolable constraint on phonological representations, at least within certain domains like the morpheme, although its universality has also been argued against (Odden 1986, 1988, 1995). In the following three sections, I will show that if the goal of sequential perceptual abstraction is the optimal recovery of underlying forms, then there is evidence for its language-specificity in the different behaviour of geminates, consecutive high-toned syllables, or consecutive nasalized syllables in various languages: some languages treat these constructs as single elements in their phonology, some as two adjacent identical elements. Since this evidence is of a phonological (i.e. theory-internal) nature, I will use the last section of this chapter to adduce some external evidence from the language-dependent lack of lexicalizability of two adjacent identical elements. In subsequent chapters, I will show that sequential abstraction in the perception grammar can handle all attested OCP effects, and that undesirable ambiguities arise if the OCP is regarded instead as a constraint in an Optimality-Theoretic production grammar.

3.3 Internal evidence for language-particular perception of geminates

The weakest (i.e. theory-internal) evidence for language-dependent perception of geminates is the maximization of perceptual recoverability. In (4c), we saw that English listeners will map [[t^ɾ _ : t^h]] to two separate segments, because that it is the best way to arrive near the lexical representation in a language that has no tautomorphic geminates; for if English listeners mapped it on a single long segment, the perceptual result would be /# ɪ:t·o:l-d #/, and the recognition process would have to split up the geminate into two segments and the word into two words. In many other languages, especially those in which geminate consonants are common, geminates do act as single segments. Again, lexical economy may be invoked, but better (cross-theoretical) evidence for this has traditionally been adduced from phonological processes and language games.

Consider the common process of palatalization of velar obstruents before high front vowels, i.e. /ki/ → /tʃi/. In most languages in which this occurs, geminates undergo this process as a whole: /kɪ/ → /tʃɪ/. This is evidence that /k:/ is a single segment in these languages, for if the geminate were represented as /kk/, we would have expected that only the second /k/ is palatalized: /kɪ/ → /ktʃɪ/ (McCarthy 1986, Hayes 1986, Keer 1999).

Languages in which tautomorphemic geminates are single segments are so common that Keer (1999) restricts GEN (the Optimality-Theoretic candidate generator) to the extent that GEN will never generate paired geminates, i.e. sequences like /kk/. Such an assertion, however, is laudably open to falsification, and a single counterexample will suffice. In Polish, the adjectives /lekk-i/ ‘light’ and /mēkk-i/ ‘soft’ are usually pronounced with geminate consonants ([lek:i], [mɛŋk:i]), but their male plural forms are pronounced as [lɛktsɪ] and [mɛŋktsɪ], which are probably structured as /lɛktsɪ/ and /mɛktsɪ/. Therefore, the change /k/ → /tʂ/ does not affect the first half of the geminate, which according to McCarthy’s, Hayes’, and Keer’s criteria would be evidence that this geminate consists of two separate segments.

Thus, the acoustic form $[[k^{\cdot} _ : k]]$ will be represented as /k:/ in most languages, but as /kk/ in Polish, so the perception grammar maps $[[k^{\cdot} _ : k]]$ to the single segment /k:/ in most languages, but to the pair geminate /kk/ in Polish.

3.4 Internal evidence for language-particular perception of nasal vowels

The perception of stretches of nasality is different in different languages.

The weakest evidence for language-dependent perception is again maximal perceptual recovery. In English, vowels are nasalized before nasal consonants in the same morpheme. Thus, *pen* is pronounced [p^hɛ̃n], whereas *pet* is [p^hɛt]. Since a nasal vowel is always followed by a nasal consonant, they will be perceived as a single nasal on some level of abstraction, and [p^hɛ̃n] and [p^hɛt] will be ultimately perceived as /# p·ɛ·n #/ and /# p·ɛ·t #/, respectively. In this way, the lexicon need not maintain a representation for a nasal vowel. Unlike with following consonants, nasal vowels have no correlation with preceding consonants in English: thus, while [mɛ̃n] has a nasal vowel, it is not caused by the preceding [m], as a comparison with [mɛt] immediately shows; indeed the nasal vowel in [mɛ̃n] is conditioned by a *following* nasal, as in [p^hɛ̃n] versus [p^hɛt]. Thus, the perception of [mɛ̃n] will take the following steps:

(9) *The perception of men in English*

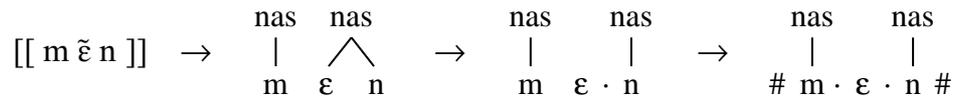
$$[[m \tilde{\epsilon} n]] \rightarrow / \# m \ \epsilon_{+nas} \ n \ # / \rightarrow / \# m \ \epsilon \cdot n \ # / \rightarrow / \# m \cdot \epsilon \cdot n \ # /$$

Phonology is little interesting without generalizations, and the generalization of the second step in (9) is that any sequence of nasalized vowel and nasal consonant becomes a tautomorphemic sequence of vowel and nasal consonant. In SPE-style (Chomsky & Halle 1968), we could formalize this as

(10) *The perception of a nasalized vowel in English*

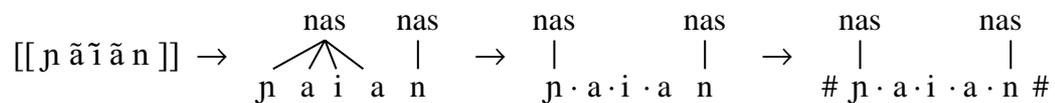
$$V_{+nas} \rightarrow V \cdot / \left[\begin{array}{c} C \\ +nas \end{array} \right]$$

or as the output filter $*/V_{+nas} \# [C, +nas]/$. We notice that the feature value [+nas] occurs twice in this formulation, and although vowel nasality is acoustically rather different from stop nasality (Delattre 1954, House 1957, Ohala 1975, Maeda 1993, Ladefoged & Maddieson 1996), we could hypothesize that they are perceived as the same thing in English on the basis of their common co-occurrence in VC sequences. If so, an autosegmental formulation would be appropriate. The second step in (9), then, involves the sequential merger of the nasal features of the vowel and the following consonant:

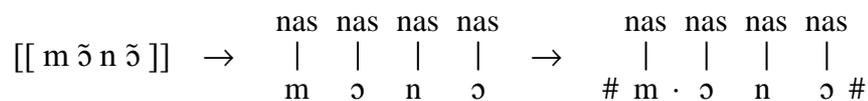
(11) *Autosegmental perception of men in English*

In the first step, we see an OCP violation: although the two [nas] values are adjacent on their tier, and are even connected to adjacent segments, they are two separate nasal feature values, on their way to build a minimal structure optimized for lexical access.

In Sundanese (Cohn 1990), the situation is the reverse from English, in that a nasal vowel is a clear sign of a *preceding* nasal consonant. The rules in the perception grammar will also be reversed. For instance, the perception of the word [ɲaiaɲ] ‘wet-ACTIVE’, which is pronounced as [ɲãĩãɲ] with rightward spreading of nasality from the first nasal, will be the mirror image of (11):

(12) *Autosegmental perception of nyaian in Sundanese*

In French, the situation is radically different from English and Sundanese, because the nasality of a vowel is not related to preceding consonants, and only hardly to following consonants. Thus, disyllabic words with non-nasal consonants can have either no nasal vowels, or a nasal vowel in the first syllable, or in the second, or in both (|bato| ‘boat’, |pɔ̃so| ‘poppy’, |lapɛ̃| ‘rabbit’, |ʃãsɔ̃| ‘song’), and the same four possibilities exist if the onset of the first syllable is nasal (|midi| ‘noon’, |mãto| ‘coat’, |matɛ̃| ‘morning’, |mãsɔ̃ʒ| ‘lie’), or even (though less stably) if the onset of the second syllable is nasal (|amur| ‘love’, |ãnuʁi| ‘boredom’, |ʃəmɛ̃| ‘road’, and perhaps |ãmãʃe| ‘put on a handle’). In the phrase [mɔ̃nɔ̃] ‘my name’, the nasality of the two vowels is not related to either nasal consonant, as we can see from the existence of e.g. [tɔ̃sɔ̃] ‘your sound’. Therefore, [mɔ̃nɔ̃] will be perceived with four separate nasal feature values, if maximal perceptual recovery is the criterion:

(13) *Autosegmental perception of mon nom in French*

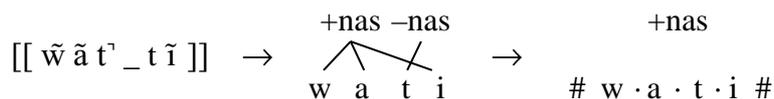
Rather than to delete the links between the nasal feature and the vowels, as the English and Sundanese perception grammars do, the French perception grammar preserves vowel nasality in order to optimize lexical access, since the lexicon contains nasalized vowels.⁵ We see that the traditional OCP is routinely violated in French *mon nom* even more than it is in English *men* or Sundanese *nyaian*.

The evidence for the language-specificity of the perception of nasality has so far come only from the theory-internal concept of the optimal preparation for lexical access. Better evidence would come from phonological processes and restrictions on lexical representations. In Southern Barasano (Smith & Smith 1971, Piggott 1992, Steriade 1993,

⁵ While the first consonant can be merged with the following vowel, the morphemic alliance of the second consonant cannot be determined in the perception grammar, since the word boundary can go before this consonant, as in *mon nom*, or after this consonant, as in *mon oncle* ‘my uncle’.

Walker 1998), lexical forms either contain non-nasal vowels only, or nasal vowels only. Thus, Southern Barasano has words like [wãtĩ] ‘dandruff?’ and [wati] ‘demon’, but no words with mixed nasal and non-nasal vowels like *[watĩ] or *[wãti]. The minimal lexical representation of these forms will have a single [+nas] or [-nas] feature value, unconnected with any specific vowel or consonant. Thus, the lexical forms will be phonologically |wati, +nas| and |wati, -nas|. The perception grammar can derive the first of these as follows:

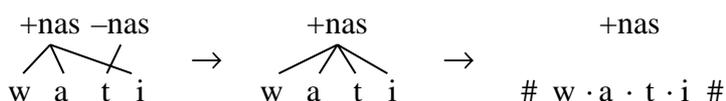
(14) *Autosegmental perception of wãtĩ in Southern Barasano*



In the first step, we see a violation of the Line Crossing Constraint (LCC), which is one of the two inviolable structure conditions of autosegmental phonology (the other is the OCP). The use of binary values for nasality in (14) makes the crossing of the two lines explicit. While the LCC is usually regarded as inviolable, we know that perception in general is discontinuous: if we see a car behind a lamppost, we see, from left to right, the front of the car, the lamppost, and the back of the car, and still we see a single car. Likewise, the nasality in (14) is perceived as a single stretch, although it is interrupted by a non-nasal oral closure.

It is not unequivocal, though, that the voiceless plosive must be marked as [-nas]. The voiceless plosives (and fricatives) of Southern Barasano have no overtly nasal counterpart, as the other consonants have ($w \sim \tilde{w}$, $^m b \sim m$, $r \sim \tilde{r}$), so they can be regarded as their own nasal counterparts, and a rather abstract view of perception as a construction process can easily posit the nasal oral plosive \tilde{t} as an intermediate form:

(15) *Abstract perception of wãtĩ in Southern Barasano*



This move simplifies the formulation of the perception grammar. In (14), the morphological concatenation of /a/ with the following vowel had to be established by an awkward rule like “merge a sequence of nasalized vowel, glottal stop, and nasalized vowel into a single word”, where the third conditioning segment is necessary because plosives are not combined with the preceding nasalized vowel if a non-nasalized vowel follows; rather, a following non-nasalized vowel would be a fairly reliable sign of a word boundary before the plosives (nasality is not spread across word boundaries in Southern Barasano). By contrast, the sequence in (15) can be expressed locally: first, mark any consonant followed by a nasalized vowel as [+nas]; second, morphemically merge every consonant with the vowel that follows it; third, morphemically merge every vowel with the following consonant if that agrees with it in nasality; fourth, set nasality afloat.⁶ So we

⁶ This may be a simplification of actual Southern Barasano. Evidence from the infrequent disharmonic words suggests that nasality can be connected to a vowel underlyingly (Piggott 1992:51), from which it spreads rightward. In an analogous analysis of the closely related language Tuyuca, Walker (1998:28) tentatively suggests that nasality could be connected to the first vowel, i.e. the underlying form of our example would be |wãti| instead of |wati, +nas|. Nothing in my story depends on this: the rightmost representations in (18) and (19) would simply receive a link from [+nas] to /a/. The disadvantage of such a move could be that lexical minimality is violated, if the failure to underspecify the default case is

see that a high degree of abstraction from reality, i.e. the filling in of something absent from the sensory data may simplify the perception process, and it will remain a question for some while how high a degree of abstraction listeners actually use. The abstractness of these nasal oral plosives is much larger than the abstractness proposed by direct realism, since the velum is not even down during the closure periods of the plosives. A perceived nasal oral plosive, then, is neither articulatorily nor acoustically present and has to be constructed completely by the perception process. It is as though the lamppost becomes the middle part of the car. We may note here that positing nasal oral plosives as intermediate forms is not an unheard-of move; an analogous construction in *production* was posited by Walker (1998), who, working within McCarthy's (1999) Sympathy Theory, considered these nasal oral plosives to be the *sympathetic forms*.

An interesting and quite common complication arises in the second step of (14), i.e. the morpheme construction step. While it is all right in this CV language to merge morphemically a sequence of consonant and vowel, i.e. put concatenation symbols in /w·a/ and /t·i/, it is a different question what we have to do with vowel-consonant sequences, i.e. whether /w·a/ and /t·i/ should be morphemically combined to give /w·a·t·i/. The answer must be: yes, if the nasality values are equal on both sides. This is because the perception process has no access to the lexicon, so that the first step in (14) has no choice but to combine the nasality of adjacent syllables into a single [+nas] feature value. For tautomorphemic $\tilde{V}C\tilde{V}$ this would be appropriate, while for heteromorphemic sequences of \tilde{V} and $C\tilde{V}$ this would be wrong. But the perception grammar cannot tell the difference between the two cases, and in order to handle the tautomorphemic case correctly, it has to morphemically merge any pair of adjacent syllables that agree in nasality. This will result in a failure to perceive a word boundary if two consecutive words agree in nasality. The recognition grammar, which maps the perceptual result to the underlying lexical form, will have to undo this inappropriate morphemic merger, and it is perfectly capable of performing this, as I will show in §5.1.

By contrast with Southern Barasano, words with consecutive nasalized vowels are represented with separate nasals in French. Phonological evidence of this separation is found in a morphological operation that changes the nasality of the second vowel but does not affect the nasality in the first vowel: /ʃɑ̃sɔ̃+je/ 'song+AGENT' → /ʃɑ̃sɔ̃nje/ 'singer', not */ʃasɔ̃nje/ (also compare the masculine-feminine pair /kəkɛ̃/-/kəkɪn/ 'rascal' with /mɔ̃dɛ̃/-/mɔ̃dɛn/ 'fashionable'). It would not be impossible to derive this from a single nasal, but the analysis would be complicated; besides, partial alterability is exactly the phenomenon that would lead proponents of the single-element view to reconsider their hypothesis (§3.3).

3.5 Internal evidence for language-particular perception of tone sequences

Starting with Leben (1973), much evidence has been adduced for the fact that in many tone languages, consecutive high-toned syllables have a single H on the tone tier, as in (8). This evidence could come, as in the case of nasal harmony discussed above, from non-underlyingly-linked elements and from spreading, but also from deletion as a unit.

Non-underlyingly linked elements occur in Mende (Leben 1973), where nouns of one syllable have H ([kɔ́] 'war'), L ([kpà] 'debt'), HL ([mbû] 'owl'), LH ([mbǎ] 'rice'), or

considered to be such a violation. It is doubtful, however, that there would be empirically differences between an account in terms of required linkage (which would almost always be to the first vowel), and an account (as here) in terms of linkage only if a non-initial vowel is the first nasal vowel of the word.

LHL ([mbã] ‘companion’), but not HLH, while nouns of two syllables have H-H ([pélé] ‘house’), L-L ([bèlè] ‘trousers’), H-L ([kéjà] ‘uncle’), L-H ([níkà] ‘cow’), or L-HL ([jàhâ] ‘woman’). By representing the five tone sequences for the bisyllabic nouns in a suprasegmental way, we can see that they are identical to those of the monosyllabic nouns:

(16) *The five bisyllabic tone sequences in Mende*



The generalization is that Mende has a morpheme structure constraint that allows only the sequences H, L, HL, LH, and LHL within morphemes. Leben’s analysis is that the segmental and tonal properties of Mende nouns are stored in the lexicon as separate sequences, e.g. |HLH-jaha| ‘woman’, and that a phonological rule maps tones to syllables in a one-to-one fashion from left to right. If there are more syllables than tones, the last tone spreads to any remaining syllables (as in |H-pɛlɛ| ‘house’), while if there are more tones than syllables, the last syllable receives any remaining tones (as in |HLH-jaha| ‘woman’).

If tones in the Mende lexicon float on their own tier, without being linked underlyingly to syllables, then the Mende perception grammar will also have to map a tone sequence on its own tier and sever the link with syllables in order to get as close as possible to the lexical representation:

(17) *Abstract Mende perception?*



The second abstraction in (17) is problematic, as were those in (14) and (15). In the perception of a complete utterance, there is no way the listener can set any tones afloat before word boundaries have been determined, which requires lexical access. This can either mean that the second form in (17) is the end result of perception (leading either to commensurability problems if the lexical form is minimal, or to lexical nonminimality if the lexical form is commensurable i.e. has tones linked to vowels), or that the comprehension process consist of cycles like those proposed for the production grammar in theories of lexical phonology (Kiparsky 1982, Mohanan 1986). In the Mende case, something can be said in favour of lexical linking, since there exist words like [mãná] ‘banana’ and [bɛsí] ‘pig’ (Leben 1973:84) and [hókpô] ‘navel’ (Odden 1995) that violate the rule of left-to-right one-to-one tone assignment.

The existence of a lexical L in Mende is crucial to expressing the contrast between bisyllabic LH and LHL: there is no way in which [jàhâ] could be represented with a H tone only. In Bantu languages, by contrast, low tones are not usually seen as playing a role in the phonology. Therefore, the lexicon will contain only H tones and these will have to be linked to vowels in the lexicon. In Zezuru Shona (Myers 1987, 1997, 1998), for instance, the forms [bázì] ‘branch’ and [bàdzá] ‘hoe’ denote two monomorphemic nouns.⁷ The first will have a H tone linked lexically to the first syllable, the second will have its H linked to the second syllable.

⁷ Segmental IPA transcriptions for Shona are due to Doke (1931).

(18) *Perception of ‘hoe’ in Shona*

$$[[\text{b} \grave{\text{a}} \text{d} \text{z} \acute{\text{a}}]] \rightarrow \begin{array}{c} \text{H} \\ | \\ \text{b} \text{a} \text{d} \text{z} \text{a} \end{array}$$

The next question is how sequences of high-toned syllables, as in the monomorphemic noun [bǎŋgá] ‘knife’ and the composite verb form [tǎŋgésérà] ‘sell to!’ should be analysed. The answer is that they are a single H tone, and the evidence comes from spreading and deletion-as-a-unit. The evidence from spreading is found in verb stems, in which an underlying H spreads maximally by two syllables rightward from the root. Thus, the root *tǎŋg* ‘buy’, augmented with the toneless default inflection terminal *-a*, becomes the imperative [tǎŋgá] ‘buy!’. This can be further extended with the toneless causative marker *-is-* to give [tǎŋgésá] ‘sell!’, and yet further with the toneless applicative marker *-ir-* to [tǎŋgésérà] ‘sell to!’. While this spreading is in itself often enough reason for linguists to posit a single H tone in [tǎŋgésérà], one could alternatively imagine that [tǎŋgésérà] is the result of copying a H tone rightward twice. Evidence against this tone-copy view comes from Meeussen’s rule, which deletes a H after an inflectional stem or clitic that ends in a single H-toned syllable. Thus, the first-singular future inflectional stem combines with the ‘sell’ verbal stem as [ndítǎtǎŋgèsà] ‘I will sell’ with deletion of the entire three-syllable H tone, and the copular clitic combines with ‘knife’ as [íbǎŋgà], with deletion of the entire two-syllable H-tone. Apparently, [bǎŋgá] is to be analysed with a single H. Whether it is also *perceived* with a single H depends on the relative problems of perceiving a single H where there are two underlying H (in case of autosegmental perception) and of perceiving multiple H tones where there is a single one underlyingly (in case of syllabic tone perception). Regarding the trouble that Shona speakers go through to prevent the surfacing of two different underlying H tones on adjacent syllables (if a word ending in two or more high-toned syllables is juxtaposed to a word starting with a high-toned syllable, the last syllable of the first word is spoken with a low tone), we must conclude that Shona listeners perceive tone autosegmentally, so that [bǎŋgá] is perceived with a single H:

(19) *Perception of [bǎŋgá] in Shona*

$$[[\text{b} \acute{\text{a}} \text{ŋ} \text{g} \acute{\text{a}}]] \rightarrow \begin{array}{c} \text{H} \\ \wedge \\ \text{b} \text{a} \text{ŋ} \text{g} \text{a} \end{array}$$

In isolation, therefore, this word is perceived exactly as it is stored in the lexicon. There are some cases, however, in which the perception does not match the lexicon. If a word ending in a single high-toned syllable, like [bǎdzá] ‘hoe’, is juxtaposed to a word starting with a high-toned syllable, like [gúrú] ‘big’, the result is a simple concatenation ([bǎdzágúrú] ‘big hoe’), and the perception grammar will reconstruct a single H:

(20) *Perception of a big hoe in Shona*

$$[[\text{b} \grave{\text{a}} \text{d} \text{z} \acute{\text{a}} \text{g} \acute{\text{u}} \text{r} \acute{\text{u}}]] \rightarrow \begin{array}{c} \text{H} \\ \wedge \\ \text{b} \text{a} \text{d} \text{z} \text{a} \text{g} \text{u} \text{r} \text{u} \end{array}$$

There is really no other way to perceive this phrase differently without lexical access, since it is tonally completely equivalent to [kùtǎŋgésá] ‘to sell’, which clearly has a

single H tone. So there is no full recovery of the two H tones in *badzá gúrú*. This just means that the final form in (20) is a surface structure with a faithfulness violation. Faithfulness violations like these can cause the emergence of other solutions than the simple concatenation of (20), which involves fusion on the tone tier.

This idea seems to be so well established that one could easily overlook the languages in which tones or tone contours are simply linked to the syllable. In Mandarin Chinese, for instance, every syllable is separately specified in the lexicon for one of the four tone complexes, and it does not do the perception process any good to combine two consecutive high-toned syllables into a single H, only to force the recognition process to separate them again. The case is analogous to the consecutive nasalized vowels in French: if, in a linguist's view, speech perception is oriented at accessing the lexicon as elegantly as possible, then Chinese listeners should keep the perceived tone sequences connected with the perceived syllables, without building pointless larger structures. Moreover, Bao (1990) mentions cases of the spreading, i.e. copying, of complete tone contours in some Chinese dialects, and argues that it cannot be analysed away as a case of reduplication.

While the Chinese case could be weakened by the assertion that every syllable is a different morpheme, such an objection is not possible for languages in which lexical tones interact with an extensive intonation system. In Venlo Limburgian (Gussenhoven & Van der Vliet 1999:129), the recoverable surface structure of the monomorphemic word |fɛilant| 'saviour' can contain as many as five H elements: two lexical tones, a focus tone, and a HH second-question intonation.

3.6 External evidence for and against sequential abstraction

Since most theories of speech production, i.e. the phonological theories of the previous century, have for their validation relied on criteria like elegance, generalization, and minimality, a grammatical account of speech perception should likewise be judged on points as elegance, generalization, and minimality, and I have indeed used those criteria to present some plausible accounts of the perception grammar. I am sure that the details of some of these accounts will have invited some readers to propose alternative analyses, and these alternative analyses were quite possibly based on the same kind of criteria.

But independent evidence is of course always a good thing to have, and a convincing case for sequential merger would be e.g. the (in)ability to lexicalize more than one feature value within a morpheme. We can expect that listeners of languages like Southern Barasano (§3.4) cannot store more than one nasal feature value for each word. Indeed, Kaye (1971) found that in loan lexicalization in Desano, Portuguese words with mixed nasal and non-nasal segments were represented either completely nasal ([jũ] for [ʒoẽũ] 'John') or completely non-nasal ([sabo] for [sɐβẽũ] 'soap').

One might argue that the lack of nasal lexicalizability in Desano is caused by a restriction in their lexicon, i.e. that Desano speakers have no provisions in their lexicon for connecting nasality to separate vowels or consonants. But this is no explanation, and it is certainly against the OT maxim of *richness of the base* (Prince & Smolensky 1993:191), according to which there are no restrictions in the lexicon and all restrictions are enforced by the grammar. A real explanation for the Desano lexicalizations is provided by the properties of the perception grammar: if the perception grammar optimizes for lexical access, and the lexicon contains single-nasality morphemes only, then the perception grammar will never construe multiple-nasality morphemes. Thus, it is the Desano perception grammar that maps the Portuguese utterance [ʒoẽũ] to the perceptual form /jũ/, which when it is heard for the first time will be stored in the

lexicon as the identical form [ɲũ]. This situation, with the perception grammar restricting possible lexical forms, is compatible with the second part of richness of the base: while there *are* restrictions in the lexicon, these are enforced by a grammar!

This may be a good place to note again that the perception grammar is not about *audibility*, i.e., the Desano perception grammar does not say that Desano listeners cannot hear the fricative in [ʒoẽũ]; it only says that Desano listeners will try to find an underlying [ɲ] in their lexicon instead of a [ʒ]. This is analogous to the situation in the production grammar, which is not about pronounceability, i.e., although the English grammar enforces a constraint against producing velar fricatives, this does not mean that these sounds cannot be produced in less-linguistic situations like imitation. The goal of the two grammars is to facilitate communication, and this has to be done in a language-specific way.

Conclusion of chapter 3:

‘Perception grammar’ is an appropriate term for the perception process, because this process submits to grammatical modelling and both of its tasks (categorization, sequential abstraction) are language-dependent.

4 How to model the perception grammar

The production grammar, i.e. the computation of a surface form from an underlying form, is what phonological theory has been modelling hitherto. We have seen here, however, that many of the concepts familiar from these models of speech production can apply to the phonology of speech perception as well: intrinsic ordering, SPE-style rules, autosegmental representations, privative features, the OCP, and the LCC. Also, the amount of unresolved issues and potential controversies seems as large as in the case of the production grammar: underspecification of boundary location (fn. 5), undoing of morphemic merger in the recognition phase (§3.4, §5.1). If the concepts and controversies are comparable, we might well choose to use a similar framework for the modelling of production and perception. In §3.1, we found that the unordered set of rules (4) was a good way to model the perception process, but if the process of speech production is described with an OT grammar, as in (6), it seems to be a logical step to formalize the perception process with an OT grammar as well. Moreover, if we can find cases of conflict resolution, then the OT formalism will prove a better way to handle speech perception than the unordered-rule formalism. In this chapter, I will show that conflict resolution plays a role in two of the main activities of the perception process, namely categorization and sequential abstraction, and I will conclude that an Optimality-Theoretic formalism is indeed appropriate for describing speech perception.

4.1 An OT account of categorization

It has been shown (Boersma 1998:ch.8) that the process of categorization along a continuous scale, as in (4b) and (4h), can be handled by an Optimality-Theoretic grammar that determines whether an acoustic value is mapped on an existing discrete phonological category (and if so, on which category), or on a new category, or on no category at all. However, it is more interesting to consider cases that have the immediate appearance of a possible conflict that has to be resolved, like the process in (4q), in which two voicing

cues may struggle for primacy: in $[[\widehat{\text{əu}}: \text{t} _ \text{d}]]$, the voicing of the final plosive is determined by the overlength of the tense vowel and by the lenis character of the release burst. The word |bolt| ‘bolt’, by contrast, would be implemented as $[[\text{b} \widehat{\text{əu}} \text{t} _ \text{t}]]$, with a simple long vowel, a short lateral, and a fortis release burst (Peterson & Lehiste 1960; Chen 1970; Raphael, Dorman, Freeman & Tobin 1975), and the perception grammar would contain a conversion of this fortis release burst into a voiceless plosive, unless an overlong vowel precedes.⁸ But what if the acoustic form is $[[\text{b} \widehat{\text{əu}}: \text{t} _ \text{d}]]$ or $[[\text{b} \widehat{\text{əu}}: \text{t} _ \text{t}]]$, i.e. if the cues of vowel length and burst strength conflict with each other? If vowel length is the stronger voicing cue (for undecisive evidence on this issue: Denes 1955; Raphael 1972; Hogan & Rozsypal 1980; Wardrip-Fruin 1982; Eilers, Oller, Urbano & Moroff 1989), $[[\text{b} \widehat{\text{əu}}: \text{t} _ \text{d}]]$ will be recognized (i.e. perceived after lexical access) as |bolt| ‘bolt’ and $[[\text{b} \widehat{\text{əu}}: \text{t} _ \text{t}]]$ as |bold| ‘bold’:

(21) *The perception of English*

$[[\text{b} \widehat{\text{əu}}: \text{t} _ \text{t}]]$	$*/V_{+\text{length}} \cdot C_0 \cdot [C, -\text{voi}]/$	$*/[C, -\text{voi}]_{\text{lenis}}/$
 /# b·o·l·d #/		*
/# b·o·l·t #/	*!	

This view of high-level cue integration, with its binary constraints working on discrete acoustic cues, is not realistic, however. The identification curves found in the literature cited above rather point to a low-level additive integration, which we could simulate with constraints that map linear combinations of *degrees* of length and burst strength to discrete contrasts. The pursuit of this is outside the scope of the current article.

4.2 An OT account of sequential abstraction

It has been shown (Boersma 1998:chs.12,18) that sequential abstraction can be described in an Optimality-Theoretic perception grammar. The relevant constraints are aptly called OCP and LCC. The constraint OCP favours the merger of two consecutive (not necessarily adjacent) elements into a higher-order unit:

(22) OCP ($f: x; cue_1 \mid m \mid cue_2$)

“A sequence of two acoustic cues cue_1 and cue_2 is perceived as a single value x on the perceptual tier f , **despite** the presence of some intervening material m .”

The constraint LCC *disfavours* the merger of two consecutive (possibly adjacent) elements into a higher-order unit:

(23) LCC ($f: x; cue_1 \mid m \mid cue_2$)

“A sequence of two acoustic cues cue_1 and cue_2 is **not** perceived as a single value x on the perceptual tier f , **because of** the intervening material m .”

Like most of the constraints introduced earlier, OCP and LCC can be locally ranked. The OCP constraint is ranked higher if the sequential combination of cue_1 and cue_2 is more

⁸ If an overlong vowel does precede, then this is a cue to a voiced plosive only if the vowel and the plosive are in the same word, since a partial phrase like *bowl to* has an overlong first vowel.

common, and it is ranked lower if there is more intervening material. The reverse correlations hold for the LCC.

The perception processes described in §3.3, §3.4, and §3.5 can now be formalized as the results of interactions of OCP and LCC constraints. The first step in (11) involves the merger of the nasalities of a vowel and the following consonant. The constraint that effects this can be written as OCP (nas: +; $\tilde{V} \mid \mid \tilde{C}$), which says “merge the nasal values of a nasalized vowel followed by a nasal consonant, if there is no intervening material”. Here, there is a condition on adjacency, as is common for the OCP. The formulation with nasalized vowel and nasal consonant assumes that some classification work must have been done before the OCP can do its work. The constraint could also have been written, then, as OCP (nas: +; $V_{+nas} \mid \mid [C,+nas]$). The merger will only be effected if this constraint outranks its LCC counterpart. Thus, in the winning form, LCC (nas: +; $\tilde{V} \mid \mid \tilde{C}$) will be violated, i.e. even if there is no intervening material, we consider the merger a line-crossing violation. This makes the name “LCC” for this constraint equally mysterious as the name “OCP” has traditionally been. The same first step in (11) also involves the non-merger of the nasalities of the first consonant and the following vowel, i.e. LCC (nas: +; $\tilde{C} \mid \mid \tilde{V}$) must outrank OCP (nas: +; $\tilde{C} \mid \mid \tilde{V}$). We can evaluate four possible candidates in a tableau:

(24) Perception of English [mɛ̃n]

[[m ɛ̃ n]]	OCP (nas: +; $\tilde{V} \mid \mid \tilde{C}$)	LCC (nas: +; $\tilde{C} \mid \mid \tilde{V}$)	LCC (nas: +; $\tilde{V} \mid \mid \tilde{C}$)	OCP (nas: +; $\tilde{C} \mid \mid \tilde{V}$)
<pre> nas nas nas m ε n </pre>	*!			*
<pre> nas nas / m ε n </pre>	*!	*		
 <pre> nas nas / m ε n </pre>			*	*
<pre> nas / \ m ε n </pre>		*!	*	

This was an example with adjacent elements on the nasal tier. In Southern Barasano, we have an example of an intervening non-nasal consonant in [wātĩ]:

(25) *Perception of Southern Barasano* [wãtĩ]

[[wã t' _ tĩ]]	OCP (nas: +; \tilde{V} C ₀ \tilde{V})	LCC (nas: +; \tilde{V} C ₀ \tilde{V})
	*!	
		*

In this case, the name of the LCC constraint is appropriate: there is a real line crossing in the winning candidate.

The morphemic merger in the second step of (11) can also be handled by an OCP constraint, namely by OCP (morpheme; \tilde{V} | | \tilde{C}).

Tone languages are handled analogously to nasal-harmony languages. The perception of the Shona form [bãdzágúru] in (20) will have a single H tone if an OCP is ranked high, and three H tones if the corresponding LCC is ranked high:

(26) *Perception of ‘big hoe’ in Shona*

[[bã dzá gú rú]]	OCP (tone: H; \tilde{V} σ \tilde{V})	LCC (tone: H; \tilde{V} σ \tilde{V})
	!	
	*!	*
		**

Note that there is no ranking of the two constraints that would make the second candidate, which would be ideal from the point of view of the lexicon, the winner. This perceptual restriction (unrecoverable material is not present in surface forms) will play a crucial role in our rejection of the OCP as a constraint in the production grammar (§5.3, §6.3).

An example of a crucial conflict between categorization and sequential abstraction will be seen in tableau (44).

4.3 An OT account of “it told”

The mapping from σ_1 to σ_2 in (7) can be handled completely by the categorization and segmental abstraction constraints of §4.1 and §4.2, e.g. with OCP (place: cor; t' | _: | t), or an analogous constraint for segmenthood instead of place. The mapping from σ_2 to σ_4 can be handled by a constraint that disfavors the absence of the morphological markers “#” and “:” in σ_4 and by negative filters like $*/V_{-leng}\cdot[C,+voi]/$, $*/[C,-voi]_{lenis}/$, $*/\#C_0\#$ (“every word has a vowel”), $*/V_{+leng}\cdot C_0\cdot[C,-voi]/$, $*/l_{dark}\cdot V/$ (“laterals are light before a vowel in the same morpheme”), and $*/C_i\cdot C_i/$ (“no sequence of identical consonants”).

within a morpheme”). All of these categorization and morphemic merger constraints work in parallel to obtain the form σ_4 as in (7).

Conclusion of chapter 4:

The perception grammar is best modelled as an Optimality-Theoretic grammar, because there are conflicts within and between its various tasks (categorization, sequential abstraction), and because the autosegmental well-formedness conditions that handle sequential abstraction (OCP and LCC) must be regarded as violable if they are defined in terms of intervening material.

5 The perception grammar in comprehension, production, and learning

The task of the perception grammar, then, is to map a raw acoustic form onto a discrete perceptual form in such a way that lexical access is maximally facilitated, which contributes to *minimization of perceptual confusion*. The perceptual form, then, has three roles: for the listener, it is the input to the recognition grammar; for the speaker, it evaluates faithfulness in the production grammar; and for the learner, it combines these two roles.

5.1 The recognition grammar

The listener will use the output of the perception grammar to access the lexicon and thereby recognize the utterance. In the case of a perceived /# ɪ·t # t·o·l·d #/, this access is particularly easy, since the lexical items |ɪt| ‘it’ and |told| ‘told’ are there for grasping. We can model this with an Optimality-Theoretic *recognition grammar*, whose input is the perceptual form and whose output is the underlying form:

(27) *The recognition of English*

/# ɪ·t # t·o·l·d #/	*INSERT (h)	*REPLACE (tense)
 ɪt ‘it’ + told ‘told’		
hi ‘he’ + told ‘told’	*!	*

The recognition grammar contains faithfulness constraints whose violation would punish candidates that are phonologically different from the perceptual form. Here again we see an example of the desirable commensurability of perceived form and underlying form: it allows us to model the recognition process with faithfulness constraints. The winning candidate in (27) simply violates no faithfulness constraints at all, since all the segments in this candidate are identical to the perceived segments. The two faithfulness constraints violated in the alternative candidate |hi| ‘he’ + |told| ‘told’ are *INSERT (h), which punishes the insertion of the segment /h/, and *REPLACE (tense), which punishes the recognition of a word containing the tense vowel /i/ while the lax vowel /ɪ/ was initially perceived.

In general, faithfulness constraints in the recognition grammar can be overruled by constraints against accessing certain lexical items. There are independent reasons for the existence of such constraints: for instance, the choice between recognizing

/# ɪ·t # t·o·l·d #/ as ‘it told’ or as ‘it tolled’ must purely be based on the semantic context, since both lexical forms have a perfect phonological match with the perceived form.⁹ And it is quite probable that this dependence on semantic context interacts with phonological faithfulness. For instance, the recognition of /# ɪ·t # t·o·l·d #/ as something meaning ‘it told’ could be overridden if the meaning ‘he told’ were much more appropriate in the context. This semantics-phonology interaction was modelled in Boersma (1999) and cannot be elaborated on here.

More interesting than the case of (27), in which the ranking of the faithfulness constraints is irrelevant, is a case in which all available candidates violate some faithfulness constraints; in such a case, we will see a conflict between the constraints, and their ranking becomes relevant. Suppose, for instance, that the speaker says [ɛ̃tʰə̃uːɪd̥] instead of [ɛ̃tʰə̃uːɪd̥], i.e. with degemination. The listener will quite probably perceive this as /# ɪ·t # t·o·l·d #/, and since there is no phonologically identical underlying form, the listener will have to choose between several phonologically imperfect matches, perhaps |ɪt| ‘it’ + |told| ‘told’, |hi| ‘hi’ + |told| ‘told’, and |ʃi| ‘she’ + |told| ‘told’. Tableau (28) shows the choice that the listener might make.

(28) *The recognition of imperfect English*

/# ɪ·t # t·o·l·d #/	*INSERT (ʃ)	*INSERT (t / _ t)	*INSERT (h)
ɪt ‘it’ + told ‘told’		*!	
 hi ‘he’ + told ‘told’			*
ʃi ‘she’ + told ‘told’	*!		

Since English speakers do not regularly delete an underlying /ʃ/ in their production, English listeners will not tend to reconstruct an underlying /ʃ/ if they do not perceive an /ʃ/. This is modelled as a high-ranking *INSERT (ʃ), which causes the candidate with the underlying /s/ to lose right away. The choice between the two remaining candidates depends on whether it is worse to recognize an unperceived geminate or to recognize an unperceived /h/. This choice is rather difficult, since on the one hand many English speakers often do not realize underlying /h/, and on the other hand many speakers will indeed degeminate here. In the example of (28), the absence of gemination is considered a more reliable cue than the absence of /h/, so the recognized form is |hi| ‘he’ + |told| ‘told’ (in practice, degemination and h deletion are probably so common that both *INSERT (t / _ t) and *INSERT (h) are ranked very low, so that the choice between ‘it told’ and ‘he told’ will be made on semantic grounds).

Consider now the problem signalled in §3.4 while discussing the morphemic merger of two consecutive syllables that agree in nasality. The problem was that if e.g. the two nasal words |taka, +nas| and |tuka, +nas| are concatenated, the result will be [tākātũkã], and this will be perceived as /#t·a·k·a·t·u·k·a#/, +nas/, an apparent merger that has to be undone by the recognition grammar, which will have to convert it into |taka, +nas| + |tuka, +nas|, which we can write in a commensurable representation as [# t·a·k·a, +nas # t·u·k·a, +nas #]. This just means that the winning candidate in recognition violates the faithfulness constraints *INSERT (#) and *INSERT (+nas); since this boundary insertion must be a common action of the recognition process for Southern

⁹ This ignores the syntactic environment. We can do this because there exist pairs of complete sentences like ‘we told it’ versus ‘we tolled it’, which can be distinguished neither phonologically nor syntactically.

Barasano (it will occur at half of all word boundaries), *INSERT (#) must be ranked rather low. All this does not mean that the perception grammar fails in constructing word boundaries: if two consecutive syllables do *not* agree in nasality, then this is a reliable indication of a word boundary between these syllables, and the recognition grammar should have little need to choose a candidate underlying form without a word boundary. This “little need” is implemented as a high ranking of the constraint *DELETE (#).

5.2 The production grammar

In tableau (6), we can see that the production grammar selects an articulatory output candidate on the basis of articulatory constraints and on the basis of the perceptual similarity of its automatic acoustic result to the underlying specification. The processing sequence of (6) can be depicted as follows:

(29) *The functional grammar model*

$$| \# \text{ ɪ} \cdot \text{t} \# \text{ t} \cdot \text{o} \cdot \text{l} \cdot \text{d} \# | \rightarrow [\text{ɛ} \text{t} : \text{h} \widehat{\text{ə}} \text{u} : \text{t} \text{d}] \Rightarrow [[\text{ɛ} \text{ t } _ : \text{ t } \text{ h } \widehat{\text{ə}} \text{ u} : \text{ t } _ \text{ d}]] \rightarrow / \# \text{ ɪ} \cdot \text{t} \# \text{ t} \cdot \text{o} \cdot \text{l} \cdot \text{d} \# /$$

The first arrow represents the choice of the production grammar, which combines the actions of what is often thought of the distinct modules of phonology and phonetic implementation. This arrow is to be taken with a grain of salt, since the choice is not only based on an evaluation of the second form but also on a comparison between the first and fourth forms. The second arrow represents the automatic conversion from an articulation into an acoustic event by the laws of physics. The third arrow represents the perception process, a genuine local process that involves nothing but the third and fourth forms.

It is illuminating to compare the production model in (29) with the other two major models of phonological modularization, namely the structuralist and generative models. I will argue that the model in (29) is the third logical possibility, and that it combines the virtues of the structuralist and generative models without copying the disadvantages of either.

If we are forced to label the four forms in (29) with traditional terms of phonological theory, then the first form will still be the underlying form, but the second (articulatory) and third (acoustic) forms will be called the phonetic form, and the fourth (perceptual) form will be called the phonological surface form. This constitutes a reversal if we compare it with the modern structuralist model, in which a module of phonology is followed by a module of phonetic implementation. In that model, the phonological surface form is intermediate between the underlying form and the phonetic form:

(30) *The structuralist grammar model*

$$| \# \text{ ɪ} \cdot \text{t} \# \text{ t} \cdot \text{o} \cdot \text{l} \cdot \text{d} \# | \rightarrow / \cdot \text{ɪ} \cdot \text{t} \cdot \text{o} \cdot \text{l} \cdot \text{d} \cdot / \rightarrow [\text{ɛ} \text{t} : \text{h} \widehat{\text{ə}} \text{u} : \text{t} \text{d}]$$

The phonological surface form in this example has prosodic (syllable) structure, not morphological structure, because in these models the intermediate form tends to be less abstract than σ_5 in (7). It is clear, though, that someone must have wrongly reversed an arrow, and that either the model in (29) or the model in (30) must be incorrect.

There are two reasons why the model in (30) must be regarded as incorrect. The first reason was supplied by Halle (1959), who argued that there was no basis to posit an intermediate phonological form between the underlying and phonetic forms. The main structuralist criterion for choosing between the first and second modules for any given process is that the first module handles computations with phonemes and that the second

module handles the generation of allophones. For instance, the voice assimilation of a Russian /t/ to a following /b/ would have to be in the first module, because it produces a /d/, which is a separate phoneme of Russian, whereas the voice assimilation of a Russian /tʃ/ to a following /b/ would have to be in the second module, because its result is [dʒ], which is not a phoneme of Russian so has to be an allophone of /tʃ/. This example of an obviously single process that is spread by linguists across two modules led Halle to state that the division into two modules is artificial, a point that was also pursued aggressively by the other early generativists (Chomsky 1964, Postal 1968). The grammar model defended in SPE (Chomsky & Halle 1968), therefore, had no intermediate representation:

(31) *The generative grammar model*

|# ɪ·t # t·o·l·d #| → [ɛ̣tː^həʊːt̚d]

To be sure, the arrow in (31), making no distinction between phonology and phonetic implementation, was divided into many sequentially ordered rules that created many intermediate representations, but none of these intermediate forms had any special theoretical status, and Russian voice assimilation was a single rule. The functionalist model in (29) shares this desirable property with the early generative model.

The second reason for rejecting the structuralist grammar model deserves its own section.

5.3 The recoverability assumption

Halle's account showed that the structuralist grammar model (30) was incorrect, but did not show, of course, that the direction of the arrow in the functionalist model (29) was correct, since the generative model has no second arrow at all. So I will now show the incorrectness of the direction of the arrow in the structuralist model.

After the phoneme/allophone distinction discussed above, there is another common linguistic criterion to choose between the two modules in a structuralist account of a given process, namely the idea that the second module should not neutralize. For instance, if Dutch /n/ assimilates to a following /f/, it becomes the labiodental nasal [ɱ], which can be considered an allophone of /n/, so that phonologists have safely been including this process in the phonetic implementation module (e.g. Cohen, Ebeling, Eringa, Fokkema & Van Holk 1959). However, there are speakers for whom /m/ assimilates to /f/ as well, and in this case, we say that /n/ and /m/ *neutralize* before labiodentals. This neutralization would have been sufficient reason for a structuralist to put this process in the first, more phonological module. We can write this idea in a more general form, which does not refer to any specific modules:

(32) *The recoverability assumption*

“If two forms are pronounced in the same way, their phonological surface forms must be identical. In other words, the phonological surface form must be recoverable from the phonetic form.”

For the structuralist, this is indeed an assumption, because the non-neutralizing property of the second module has to be maintained at a cost. After all, how can this module know that if it maps A to B, it is not allowed to map C to B as well? For the learner, this would require storing input-output pairs and taking action as soon as she discovers that two of these pairs have identical outputs but different inputs. This is quite a plausible situation,

since 4-to-6-year-old children tend to have fewer interword assimilation processes than adults (e.g. Hernández-Chávez, Vogel & Clumeck 1975), so that they are likely to have to move some processes from the phonetic implementation module to the phonology at some point. But to what avail? Why should the child care which module handles a certain process, as long as it is handled by *some* module? For no good reason, apparently, as long as the child's objective is to communicate, not to keep linguists happy.

The generative approach does not share this problem with the structuralists. Since there is no intermediate form, the generative phonological surface form can be identified with the phonetic form, and this makes the satisfaction of the recoverability assumption trivial in early generative phonology.

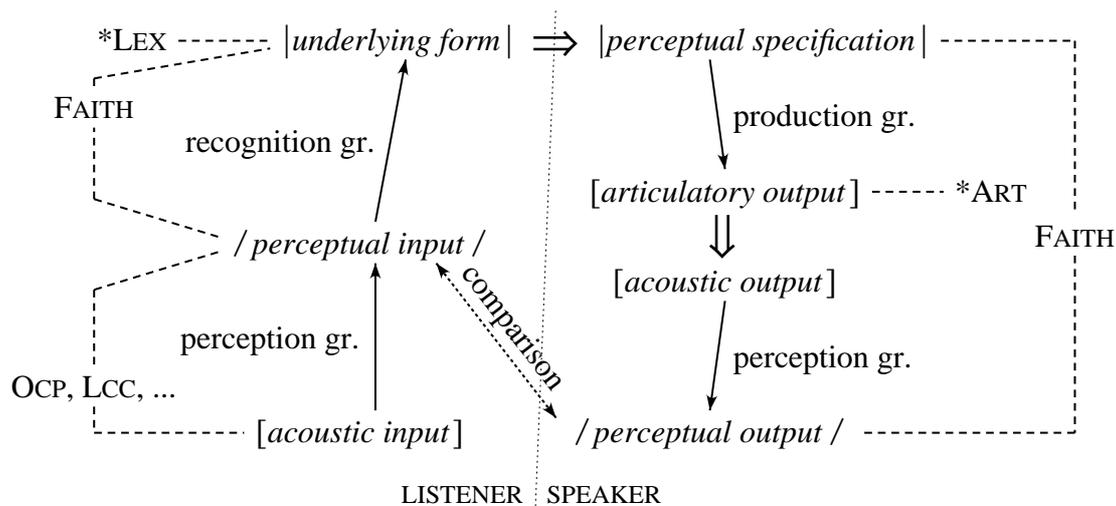
The structuralist problem with the recoverability assumption is the direction of the arrow in (30). Apparently, the relationship between phonological surface form and phonetic form is one-to-many, and to a mathematician this would mean that the phonological surface form can be a *function* of the phonetic form, and that the reverse can never hold. Therefore, the arrow has to be reversed, as in the functionalist approach of (29). Indeed, we see that if two forms are pronounced (articulatorily and, therefore, acoustically) in the same way, then the last arrow in (29) predicts that they will be perceived in the same way.

We can conclude that the recoverability assumption is shared by the structuralist model (at a cost), by the early generative model (trivially), and by the theory defended here (enforced by the direction of the arrow). Therefore, I will assume that this assumption is a desirable property for models of phonology, and I will later on come to reject accounts that violate it. Within the context of this article, the assumption must hold because it is equivalent to stating that any structural difference between two forms that are pronounced in the same way, is not perceptually recoverable without lexical access.

After identifying two problems with the structuralist approach, and showing that both the early generative approach and the functionalist approach do not copy these problems, I will have to assess the difference between the generative and functionalist approaches.

5.4 Learning

The generative move of doing away with an intermediate phonological surface form comes at a cost. Traditionally (e.g. Hockett 1965), this intermediate form has been associated with a discrete mental perception of the surface form, and has been presumed to be involved in learning based on comparing one's own surface form with those of others. The generative approach, by contrast, has no place for such a representation, and loses any advantages that it might have. In the model defended here, the perceptual surface form is used not only for evaluating faithfulness and for the preparation of recognition, but also for the acquisition of the production grammar. Figure (33) shows the complete grammar model (after Boersma 1998:143).

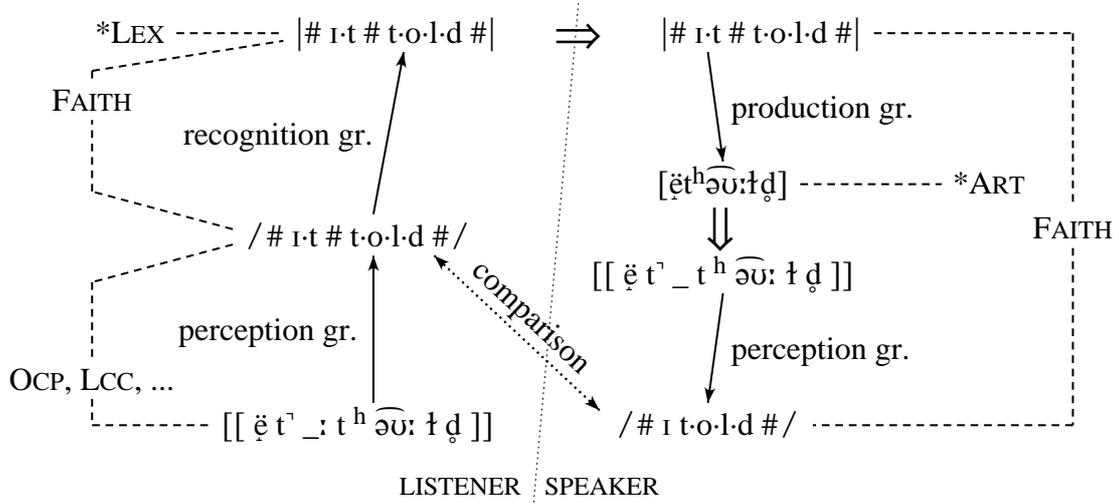
(33) *The grammar model of functional phonology*

The right-hand side of figure (33) shows the production model with the four representations known from (6) and (29), the articulatory constraints (*ART) that evaluate the articulatory output, and the faithfulness constraints (FAITH) that evaluate the perceptual similarity between the specification (underlying form) and the surface form. Both sets of constraints reside in the production grammar, which ranks them and settles their conflicts. The left-hand side of the figure shows the comprehension model as a sequence of perception and recognition. The constraint sets OCP and LCC (§4.2) handle sequential abstraction. These constraints of autosegmental perception reside in the perception grammar, together with some sets of constraints that handle categorization and are not treated in the current article (see Boersma 1998:ch.8). Note that the perception grammar occurs twice in this figure: once for providing feedback to the speaker about her own production, once for listening to others. The recognition grammar (§5.1) contains constraints against lexical access (*LEX), which are not treated in this article (see Boersma 1999), and again the faithfulness constraints (FAITH), which are the mirror images of those in the production grammar. Both sets of faithfulness constraints may well be ranked in the same order; for instance, languages that devoice final obstruents have a low-ranked *DELETE (voice) constraint in production, and listeners will have to reconstruct this voicing easily in the recognition grammar, presumably by having a low-ranked *INSERT (voice) constraint there. Likewise, the high ranking of *DELETE (#) in the recognition of Southern Barasano (end of §5.1) is probably caused by (or causes) the high ranking of *INSERT (#) in production. Finally, the double arrow between “underlying form” and “perceptual specification” simply assures the identity of the lexical form accessed in recognition with the lexical form used as the input to production, i.e. it makes explicit the reciprocity of the Saussurean sign (Saussure 1916:101).

Crucial for learning, now, is the arrow marked “comparison” in figure 31: the learner will compare her own surface forms, as she perceives them, with the adult surface forms, as the learner perceives them. An *error-driven* learner (Gibson & Wexler 1994; Tesar & Smolensky 1998; Boersma & Hayes, to appear) will change her production grammar only if these two perceptual forms are different. We can now see why the output of the listener’s perception grammar has to be an intermediate representation in comprehension: if perception and comprehension had consisted of a single module, the learner would not have had a basis for comparing her own surface forms with. As an example, I will show the case of (6), i.e. the underlying form $[\# \text{I} \cdot \text{t} \# \text{t} \cdot \text{o} \cdot \text{l} \cdot \text{d} \#]$, but for a learner whose *GESTURE constraint (against maintaining the long hold for the geminate) outranks her

*DELETE constraints, so that she will pronounce this form as the second candidate in (6), the degeminated [ɛ̃tʰəu:tɔ̃]:

(34) *Learning from a mispronunciation of <it told>*



We see that the speaker's own form will be perceived (by herself) as /# ɪ t o l d #/, while she perceives the adult form as /# ɪ t # t o l d #/ (which happens to be identical to the underlying form, but that is irrelevant; learning involves the comparison of two surface forms, not the comparison with an underlying form). These two surface forms are different, and the speaker will accordingly take action by reranking some constraints, perhaps by demoting the *GESTURE constraint below the two *DELETE constraints in one stroke (the fast learning algorithm for ordinal constraint grammars: Tesar & Smolensky 1993, 1998), or perhaps by lowering the ranking of *GESTURE and raising that of the two *DELETE constraints by a small amount along a continuous ranking scale (the gradual learning algorithm for stochastic constraint grammars: Boersma 1997, 1998, 2000; Boersma & Hayes, to appear):

(35) *Learning the production of English*

# ɪ t # t o l d # adult surface form /# ɪ t # t o l d #/	*GESTURE (tongue tip: close & open / long)	*DELETE (t)	*DELETE (#)	*GESTURE (tongue tip: close & open / short)
[ɛ̃tʰəu:tɔ̃] ⇒ [[ɛ̃ tʰ _ : tʰ əu: t _ ɔ̃]] ✓ → /# ɪ t # t o l d #/	*!→			*
☞ [ɛ̃tʰəu:tɔ̃] ⇒ [[ɛ̃ tʰ _ tʰ əu: t _ ɔ̃]] ☞ → /# ɪ t o l d #/		←*	←*	←**

In this tableau, the top left cell shows the forms assumed known to the learner, namely the underlying form and the adult form, as perceived by the learner. The learner will assume that the adult form is correct (✓), and, noticing that this form is different from her own form /# ɪ t o l d #/, she will conclude that her own form (☞) is incorrect (☞) and

rerank some constraints in the direction of the arrows. After a number of these rerankings, the *DELETE constraints will come to outrank *GESTURE, as in (6), and the learner will begin to show adult-like speech production.

5.5 Phonetic detail

The theory presented here claims to combine lexical economy with the grammaticization of phonetic principles. The perceptual processes of categorization and sequential abstraction, however, are directly aimed at producing economical representations, and lead to the impossibility for faithfulness constraints to refer to phonetic detail. It is now my responsibility, therefore, to show how phonetic detail is chosen and learned.

The perception grammar will remove much phonetic detail from the acoustic form. If an English speaker pronounces |pen| ‘pen’ as the slightly incorrect [p^hɛn], her perception will quite probably be /# p·ɛ·n #/, i.e. exactly the same as the perception of the correct [p^hɛ̃n], because the alternative, namely stranding a single consonant as in /# p·ɛ # n #/, is probably worse in English:

(36) *The perception of an English pen produced without vowel nasalization*

[[p ^h ɛ n]]	OCP (morpheme; V C ₁ #) (“merge final consonants with preceding vowel”)	LCC (morpheme; V _{-nas} [$\begin{matrix} C \\ +nas \end{matrix} \end{matrix} \end{matrix} \end{matrix})$ (“insert boundary between non-nasal vowel and nasal consonant”)
☞ /# p·ɛ·n #/		*
/# p·ɛ # n #/	*!	

We see here a crucial case of conflict resolution: an ‘as soon as’ rule formulation as in (4) cannot do the job, since this might insert the boundary between /ɛ/ and /n/ before the vowel gets a chance of being merged with the coda consonant. There also has to be a third candidate, /# p·ɛ·d #/, which satisfies both constraints in (36) but violates a categorization constraint that says that if the input contains full nasality during stop closure, then the output should contain a corresponding [+nasal] value. To rule out /# p·ɛ·d #/, the listener must rank this categorization constraint above the LCC in (36). It appears, then, that the main tasks of the perception grammar, namely categorization and sequential abstraction, can be in conflict with each other, suggesting again that it is correct to model the perception process as an OT grammar.

So how does the English speaker prefer to implement |pen| ‘pen’ with a nasalized vowel, i.e. as [p^hɛ̃n]? Since faithfulness constraints cannot distinguish the two candidates, the choice must lie in a conflict between articulatory constraints, for instance an organizational constraint that says that the velum must be lowered as early as possible, and its counterpart that says that the velum must be lowered as late as possible:

(37) *The production of an English pen*

# p·ε·n #	*DELETE (+nas)	VELUMEARLY	VELUMLATE
$[p^h\epsilon n]$ $\Rightarrow [[p^h \epsilon n]]$ $\rightarrow / \# p \cdot \epsilon \cdot n \# /$		*!	
 $[p^h\tilde{\epsilon}n]$ $\Rightarrow [[p^h \tilde{\epsilon} n]]$ $\rightarrow / \# p \cdot \epsilon \cdot n \# /$			*

There is probably no intrinsic fixed ranking of the two articulatory constraints, since they appeal to very different principles of effort minimization: the speaker would like an early velum lowering because the velum will be low if all the velum muscles relax, and she would like a late velum lowering because forbearance may sometimes lead to acquittance. Thus, different languages will rank these two constraints in different orders, which explains the attested typology of languages that do and languages that don't nasalize vowels before nasal consonants.

While the adult speaker may be able to get away with ranking articulatory constraints and ignoring faithfulness in her choice between $[p^h\epsilon n]$ and $[p^h\tilde{\epsilon}n]$, the English child who has to learn that the word *pen* is to be pronounced with a nasalized vowel cannot get by only by comparing the most abstract forms, since the ultimate perceptual results of $[p^h\epsilon n]$ and $[p^h\tilde{\epsilon}n]$ are identical. How then does she learn that the vowel should be nasalized, if she herself produces $[p^h\epsilon n]$, presumably because the two gestural constraints in (37) have a reversed ranking in her grammar? The answer is that she will also compare the production and comprehension of the intermediate steps in the perception process:

(38) *Learning phonetic detail*

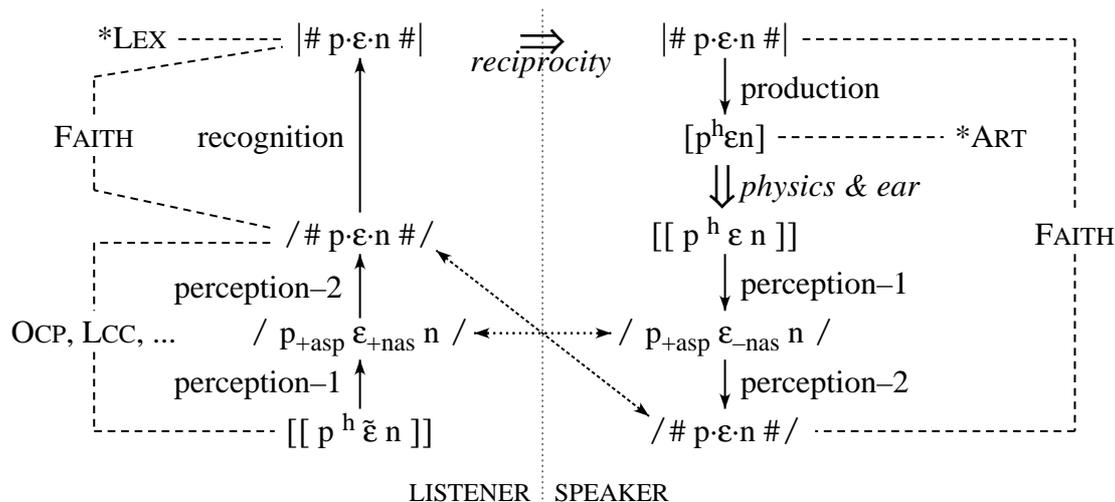


Figure (38) summarizes the complete grammar model, and specifically shows how the ‘phonetic’ nasality of the vowel in *pen* is learned. We see that while the output of the perception grammar is the same for production and comprehension (the oblique double-headed arrow), the intermediate form is different for production than for comprehension (the horizontal double-headed arrow), so that the learner will be able to change her grammar, this time by raising the rankings of the constraints violated in the candidate

[p^hɛn] - /# p·ɛ·n #/, and lowering the constraints violated in the candidate [p^hẽn] - /# p·ɛ·n #/:

(39) *The learning of an English pen*

# p·ɛ·n # adult: / p _{+asp} ɛ _{+nas} n /, /# p·ɛ·n #/	*DELETE (+nas)	VELUMLATE	VELUMEARLY
 [p ^h ɛn] ⇒ [[p ^h ɛ n]]  → / p _{+asp} ɛ _{-nas} n / ✓ → /# p·ɛ·n #/			←*
[p ^h ẽn] ⇒ [[p ^h ẽ n]] ✓ → / p _{+asp} ɛ _{+nas} n / ✓ → /# p·ɛ·n #/		*!→	

Since the two perceptual forms are identical, no faithfulness constraints can be reranked. The only constraints whose ranking can be changed are those that evaluate the articulatory forms. The existence of two opposite and rankable gestural constraints will lead to learnability in this case.¹⁰

Rich representations like / p_{+asp} ɛ_{-nas} n / are intermediate forms in perception. In Steriade's (1995) proposal, underlying forms are rich representations and faithfulness is evaluated on these rich representations (more accurately, Steriade's production grammar contains constraints that favour the presence of specified acoustic features). Since the lexicon cannot directly provide such representations, this richness restricts the scope of Steriade's analysis to the phonetic implementation module. Similar examples of rich representations in the literature are Hayes (1996), who explicitly distinguishes phonology (with markedness constraints) and phonetic implementation (with articulatory constraints); Flemming (1995); Jun (1995); and Kirchner (1998). It is interesting to see that this UCLA school, which explicitly aims at introducing phonetic principles into phonology, actually has to make an implicit structuralist distinction between a phonological and a phonetic module. By contrast, the three-grammar model defended in this article manages to combine the two modules by ensuring that faithfulness is evaluated on minimal surface representations only, as in figure (38). Rich forms like / p_{+asp} ɛ_{-nas} n / are never evaluated for faithfulness, they only play a role in learning, as figure (38) shows. In other words, the nasality of the vowel in *pen* is not caused by the presence of [+nas] in the underlying form (as in the UCLA theory), nor by the introduction of [+nas] during the derivation (as in *SPE*), but by a comparison with the surface forms of other speakers.

But the reader may feel cheated. It is possible that the choice between [p^hɛn] and [p^hẽn] is not based on gesture minimization strategies, but on perceptual considerations. The perceptual advantage of saying [p^hɛn] is that the vowel is kept more distinct from

¹⁰ Note that Kirchner's (1998) single LAZY constraint would not admit of this kind of learning.

other vowels than in the pronunciation [p^hɛ̃n], since nasalization adds to vowel confusability (ref). The perceptual advantage of saying [p^hɛ̃n], on the other hand, is that there are more acoustic cues for the correct classification of nasality. Thus, the candidate [p^hɛ̃n] will lead to more incorrect perceptions of /# p·ɪ·n #/ (and of /# p·æ·n #/, although the absence of /æ/ lengthening provides an additional cue to the identity of the vowel), whereas the candidate [p^hɛn] will lead to more incorrect perceptions of /# p·ɛ·d #/ (and of /# p·ɛ # n #/, if we take English morphological perception into account). The only way to rerank faithfulness constraints would be the occasional perception of [p^hɛn] as /# p·ɛ·d #/ or /# p·ɛ # n #/, which would be possible in a stochastic grammar. But this would lead to raising *DELETE (+nas) and *INSERT (#) in a production grammar like (39), and this would *not* lead to a higher probability of producing [p^hɛ̃n], since the ranking of these two constraints has no bearing on the choice between candidates that are perceived as identical.

As an unambiguous example of where the solution must be found, I will consider a case in which the articulatory constraints involved have a fixed mutual ranking. This is the production of an underlying vowel |a| which must be pronounced with a maximally high first formant (F_1) in order to minimize the probability that it is perceived as one of its neighbours /ɛ/ or /ɔ/ instead. But the higher F_1 must be, the more effort the speaker has to spend on opening the jaw. Accordingly, there must be some *working point*, i.e. optimal combination of F_1 and jaw height, which is 760 Hz and 2.3 cm in the following example from Boersma (1998:208):

(40) *Balancing jaw energy with vowel lowness*

a i.e. max F_1	*JAW (≥ 4 cm)	F_1 (a) ≥ 600 Hz	*JAW (≥ 3 cm)	F_1 (a) ≥ 700 Hz	*JAW (≥ 2 cm)	F_1 (a) ≥ 800 Hz	*JAW (≥ 1 cm)
[jaw 1.21 cm] ⇒ [[F_1 = 550 Hz]]		*!		*		*	*
[jaw 1.69 cm] ⇒ [[F_1 = 650 Hz]]				*!		*	*
☞ [jaw 2.25 cm] ⇒ [[F_1 = 750 Hz]]					*	*	*
[jaw 2.89 cm] ⇒ [[F_1 = 850 Hz]]			*!		*		*

From these four candidates, the winner is 750 Hz (with a finer-grained or continuous set of candidates, the winner would have been 760 Hz). The first two candidates have a too low F_1 , the fourth candidate involves a too difficult jaw gesture. The working point is where the maximum of the problems (effort and confusability) is minimal. The gestural constraints *JAW in (40) have a fixed ranking by articulatory effort. The constraints that maximize F_1 in tableau (40) are examples of *acoustical faithfulness* constraints, and they are untypical in Boersma (1998) in that they are the only faithfulness constraints that refer to phonetic detail explicitly and hence counter my claim in this article. To solve the contradiction, we must note that the faithfulness constraints in (40) are really ranked by probability of perceptual confusion. For instance, if a vowel with an F_1 of 700 Hz has only 70 percent chance of being identified as /a/, and 30 percent of being classified as /ɛ/, then the constraint “ F_1 (|a|) ≥ 700 Hz” means nothing more than “implement an

underlying |a| in such a way that it has at least 70 percent chance of being perceived as /a/, and this alternative formulation takes into account phonetic detail without making explicit reference to it. Tableau (40) would become something like

(41) *Balancing jaw energy with vowel confusion probability*

a	*JAW (≥ 4 cm)	*REPLACE (a , /ε/, ≥ 90%)	*JAW (≥ 3 cm)	*REPLACE (a , /ε/, ≥ 30%)	*JAW (≥ 2 cm)	*REPLACE (a , /ε/, ≥ 10%)	*JAW (≥ 1 cm)
[jaw 1.21 cm] ⇒ [[$F_1 = 550$ Hz]] → 1% /a/, 99% /ε/		*!		*		*	*
[jaw 1.69 cm] ⇒ [[$F_1 = 650$ Hz]] → 20% /a/, 80% /ε/				*!		*	*
☞ [jaw 2.25 cm] ⇒ [[$F_1 = 750$ Hz]] → 85% /a/, 15% /ε/					*	*	*
[jaw 2.89 cm] ⇒ [[$F_1 = 850$ Hz]] → 99% /a/, 1% /ε/			*!		*		*

The candidate cells now clearly show, unlike the candidate cells in many earlier tableaux that had to use IPA symbols for all three representations, that articulatory, acoustic, and perceptual representations are different things. The perceptual forms are expressed with percentages for the possible perceived segments, and we can assume that the speaker can compute e.g. the percentages for a jaw opening of 1.69 cm by running the acoustic input, i.e. the F_1 value of 650 Hz, through her perception grammar repeatedly and noting that 20% of these repetitions leads to a perception of /a/, 80% to a perception of /ε/. For this probabilistic interpretation to work, the perception grammar must be a *stochastic constraint grammar*, i.e. a grammar in which constraints are ranked along a continuous scale and in which some random number is added to the ranking of each constraint at evaluation time. Such a model was proposed for the production grammar by Liberman (1993:21, cited in Reynolds 1994), Zubritskaya (1997:142–4), Hayes & MacEachern (1998), Boersma (1998), and Hayes (2000), and for the perception grammar by Boersma (1997).

The constraint *REPLACE (|a|, /ε/) that appears in (41) is the perceptual counterpart of McCarthy & Prince's IDENT-IO (height), except that it has to take two arguments to accommodate the dependence of its ranking on the replacent, i.e. it is worse to implement an underlying |a| as something perceived as /e/ than as something perceived as /ε/, and we want to be able to write this universal relation as the fixed ranking *REPLACE (|a|, /e/) >> *REPLACE (|a|, /ε/). In the way described here, articulatory constraints still refer to phonetic detail explicitly, but faithfulness constraints refer only to economical representations such as lexical segments.

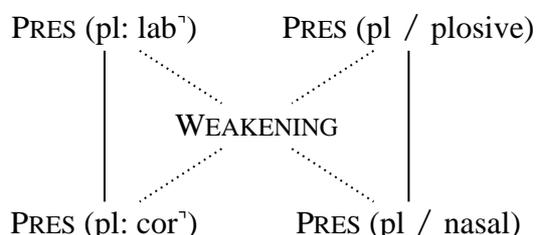
In the way described here, the incorporation of phonetic detail involves the splitting of a single constraint *REPLACE (|a|, /ε/) into a continuous family of confusion-dependent members. If this complete family outranks another constraint or constraint family, we can

see the effects of strict ranking, which is associated with “phonological” OT. Hereby, the difference between categorical effects and gradient phonetic optimization is not that these two are handled in separate modules, but that the two occupy different sizes of areas in a single ranking space.

5.6 Licensing by cues

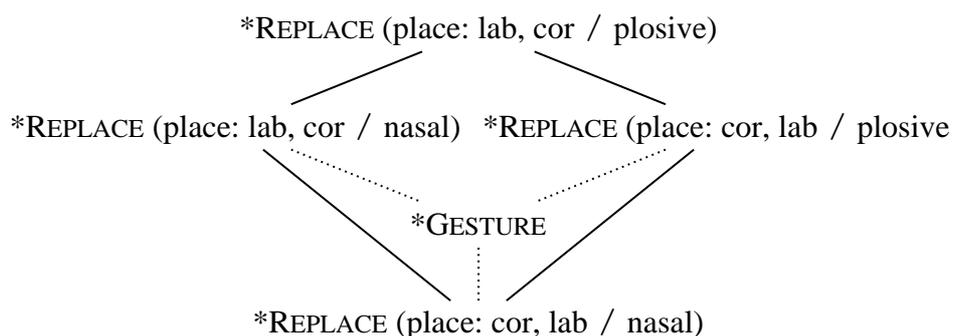
Mohanan (1993) observed that the typology of place assimilation adheres to at least two implicational universals. Restricting ourselves to labials and coronals, these universals are (1) if plosives assimilate, then so do nasals (with the same place of articulation), and (2) if labials assimilate, then so do coronals. Jun (1995) attributed these universals to his Production Hypothesis (“speakers make more effort to preserve the articulation of speech sounds with powerful acoustic cues”). Jun argued that plosives have better cues for place distinctions than nasals, and labials have better cues than coronals, and expressed this as the universally fixed faithfulness rankings $\text{PRES (pl / plosive)} \gg \text{PRES (pl / nasal)}$ and $\text{PRES (pl: lab}^\text{)} \gg \text{PRES (pl: cor}^\text{)}$, where “pl” stands for “acoustic place cues”. When combined with a freely rankable articulatory constraint (WEAKENING), this leads to a fivefold typology. If WEAKENING is ranked above all four cue preservation constraints, then all segments will assimilate (even $|\text{ap+na}| \rightarrow [\text{atna}]$), and if WEAKENING is ranked below PRES (pl / nasal) and/or below $\text{PRES (pl: cor}^\text{)}$, then no segments will assimilate (not even $|\text{an+pa}| \rightarrow [\text{ampa}]$). The Dutch and Catalan case, in which nasal coronals assimilate, but labials and plosives do not, is given by the following ranking:

(42) *Dutch and Catalan with autonomous acoustic cues (Jun 1995)*



In this figure, solid lines depict the fixed rankings, dotted lines the language-specific rankings. The remaining two types predicted by Jun’s analysis come from lowering one of the preservation constraints at the top: with $\text{PRES (pl: lab}^\text{)} \gg \text{WEAKENING} \gg \text{PRES (pl / plosive)}$ as the top ranking, we have a language in which all coronals but no labials assimilate; with $\text{PRES (pl / plosive)} \gg \text{WEAKENING} \gg \text{PRES (pl: lab}^\text{)}$, all nasals but no plosives assimilate.

But faithfulness constraints need not refer to acoustic cues. A constraint against the assimilation $|\text{at+ma}| \rightarrow [\text{apma}]$ can just as well be expressed solely with lexical features, namely as $\text{*REPLACE (place: cor, lab / plosive)}$. The previous ranking by acoustic cues now translates into a ranking by confusion probability (Boersma 1998:ch.9): since plosives have better place cues than nasals, the place value of plosives is confused less probably than the place value of nasals, and $\text{*REPLACE (place: cor, lab / plosive)}$ is universally ranked above $\text{*REPLACE (place: cor, lab / nasal)}$. We can identify four such fixed rankings, and figure (43) shows how the Dutch and Catalan case is handled.

(43) *Dutch and Catalan with faithfulness on economical representations*

A sixfold typology is generated by varying the height of the single gestural constraint, and varying the relative ranking of the two mid-level faithfulness constraints.¹¹

While the acoustical and perceptual accounts seem equally convincing, there is an interesting difference in the predicted typologies. Jun's analysis does not allow for the sixth assimilation type predicted by Mohanan's universals, namely the type in which all nasals and all coronals assimilate but |ap+na| does not assimilate to [atna]. To extend his account to accommodate this sixth type, Jun would have to add a conjoined constraint PRES (pl: lab⁷) & PRES (pl / plosive), which is violated only if |ap+na| becomes [atna], and which has to be ranked above all preservation constraints in (42). By contrast, this rather artificial extension of the constraint set is not needed if we express faithfulness constraints directly with perceptual arguments, and rank them by confusion probability, as in (43). This way of ranking corresponds well, by the way, to Steriade's (1995) version of the Production Hypothesis ("the constraint to implement a *feature value* is ranked higher in contexts with more or better acoustic cues").

As in the previous section, there seems to be no need at all to refer to acoustics or low-level perception in the formulation of faithfulness constraints, since they can be ranked by perceptual confusion, and we get a bonus by being able to accurately reflect implicational universals.

Conclusion of chapter 5:

When combined with a recognition grammar and a production grammar, the existence of a perception grammar that handles categorization and sequential abstraction ensures compatibility between lexical economy and phonetically-based ranking of articulatory and faithfulness constraints.

6 OCP effects

The perception grammar can merge two cues across a long silence into a single geminate, two consecutive high-toned syllables into a single H, and two consecutive nasalized vowels into a single [+nas]. The main evidence for this is that the result acts as a single element, i.e. it is either affected as a whole or not affected at all by other processes (such as deletion). Although the theory developed here resolves the paradox of lexical economy versus phonetic detail, it will not be an acceptable theory of phonology unless it can be shown to handle the cases that have been put forward in the phonological literature on

¹¹ In this case, a single LAZY constraint (Kirchner 1998) would do a better typological job than a freely rankable pair *GESTURE (lips) and *GESTURE (tongue) would.

phenomena ascribed here to the perception grammar, in particular OCP effects. I will discuss examples from the literature on geminates, tone, and nasal harmony, and show not only that these cases can be handled within the OCP-in-the-perception-grammar view, but also that this view can lead to simpler analyses.

6.1 Gemination and antigemination

A typical example of a segmental OCP effect is the allomorphy in the English past tense formation. The past tense suffix is realized as /d/ (*ban – banned*), except that it is /ɪd/ after a base that ends in /d/ or /t/ (*nod – nodded*). The reason for this apparent /ɪ/-epenthesis has traditionally been sought in the difficulty of perceptual recovery of the morpheme if epenthesis is not applied. Thus, if the past tense of [nɔd] ‘nod’ would be made by suffixation with /d/, the produced result would be something like [nɔd:], and this would have a large probability of being perceived as /nɔd/, i.e. as identical to the infinitive and several present-tense forms. In §1.4, I have proposed that long closures are perceived in English as a heteromorphemic sequence of two plosives, but those were cases in which a vowel followed. Closure length in final position, as in [[n ɔ d^ː _ : d]], is probably used in English for other purposes, like determining plosive voicing. Tableau (44) shows three obvious candidates for the perception of [[n ɔ d^ː _ : d]], if English were a language without epenthesis.

(44) *The perception of [nɔd] with a past-tense suffix without epenthesis*

[[n ɔ d ^ː _ : d]]	OCP (place; d ^ː _ : d / final)	*/[C,-voi] _{lenis} /	*/[C,+voi] _{long} /
 / n ɔ d ^ː t _{lenis, long} / → /# n·ɔ·d _{lenis, long} #/ → /# n·ɔ·d #/			*
/ n ɔ d ^ː t _{lenis, long} / → /# n·ɔ·t _{lenis, long} #/ → /# n·ɔ·t #/		*!	
/ n ɔ d ^ː d ^ː t _{lenis} / /# n·ɔ·d+d _{lenis} #/ → /# n·ɔ·d + d #/	*!		*

In final position, where aspiration is unavailable as a voicing cue, closure length is a more important voicing cue than in prevocalic position. Thus, it would be more problematic to have to distinguish *wed*, *wedd*, *wet*, and *wett*, than it is to distinguish *'e told*, *it told*, *'e doled*, and *it doled*. If the odds of perceiving [[n ɔ d^ː _ : d]], as /# n·ɔ·d #/ are large, then this means that a faithfulness constraint for the surfacing of one of the underlying |d| is violated:

(45) *The production of the past tense of 'nod' in English*

# n·ɔ·d + d #	*DELETE (+d)	*INSERT (ɪ)
[nɔdɪ] ⇒ [[n ɔ d' _ : d]] → /# n·ɔ·d #/	*!	
[nɔd] ⇒ [[n ɔ d' _ d]] → /# n·ɔ·d #/	*!	
☞ [nɔdɪd] ⇒ [[n ɔ _ d ɪ d' _ d]] → /# n·ɔ·d·ɪ·d #/		*

The winning candidate violates the faithfulness constraint *INSERT (ɪ), which militates against inserting a non-underlying vowel. Apparently, however, the cost of losing the past-tense information is greater, as symbolized by the high ranking of *DELETE (+d).

We know that historically the process was not one of epenthesis, but one of /ɪ/-deletion. The change of an earlier /bænɪd/ 'ban+ed' to a later /bænd/ may have been caused by a desire to speed up speech and was probably allowed by the cross-linguistically well-attested low ranking of faithfulness in grammatical morphemes:

(46) *Earlier English production of the past tense of 'ban'*

# b·æ·n + ɪ·d #	*DELETE (consonant)	*DELETE (vowel / base)	*TIME	*DELETE (vowel / affix)
[bæɪnɪd] → /# b·æ·n·ɪ·d #/			*****!	
☞ [bænd] → /# b·æ·n·d #/			****	*
[bɪnɪd] → /# b·n·ɪ·d #/		*!	****	
[bæɪn] → /# b·æ·n #/	*!		***	*

The constraint *TIME militates against long messages and may be speaker-oriented (shorter messages cost less effort) and/or listener-oriented (shorter messages lead to faster comprehension); in the current example, it simply counts the number of symbols in the articulatory form. The three faithfulness constraints express the idea that consonants tend to carry more information than vowels, and that information in content morphemes is more important than information in grammatical morphemes. Applied to [nɔd], the same grammar will select a different past-tense allomorph:

(47) *Earlier English production of the past tense of 'nod'*

# n·ɔ·d + ɪ·d #	*DELETE (consonant)	*DELETE (vowel / base)	*TIME	*DELETE (vowel / affix)
 [nɔdɪd] → /# n·ɔ·d·ɪ·d #/			*****	
[nɔd:] → /# n·ɔ·d #/	*!		*****	*
[nɔd] → /# n·ɔ·d #/	*!		****	*

Processes like epenthesis between similar segments, as in tableau (45), and the blocking of deletion between similar segments, as in tableau (47), have traditionally been regarded as OCP effects (McCarthy 1986, Yip 1988), the former being an example of the OCP as a *rule-triggering* device, the latter being an example of the OCP as a *rule-blocking* device. Both kinds of OCP effects turn out to be expressible in the production grammar without any reference to the OCP; only faithfulness constraints can do the job. The connection with similarity of adjacent segments is caused by the perception grammar, which tends to map sequences of similar units to a single higher-level unit.

6.2 Tonal epenthesis as an OCP effect

With the simple examples of segmental OCP effects out of the way, we can now turn to the more complicated cases of tone languages, in which the OCP often interacts with other language-specific constraints on tone sequences. A realistic example of tonal epenthesis is provided again by Shona (Myers 1997), in which a juxtaposition of two H-words surfaces with a HLH pattern, e.g. [báŋgá] + [gúru] → [báŋgàgúru]. Within an Optimality-Theoretic framework, effects like these have been ascribed to a constraint OCP in the production grammar (Myers 1997, Urbanczyk 1999). Myers' solution for [báŋgàgúru] is given in (48):

(48) *The OCP as a production constraint (Myers 1997)*

$\begin{array}{c} H_1 \quad H_2 \\ \wedge \quad \wedge \\ \text{ba} \ \eta\text{ga} \quad + \quad \text{gu} \ \text{ru} \end{array}$	OCP	UNIFORMITY	ANCHOR-L	MAX-IO (T)	MAX-IO (A)
$\begin{array}{c} H_1 \quad H_2 \\ \wedge \quad \wedge \\ \text{ba} \ \eta\text{ga} \ \text{gu} \ \text{ru} \end{array}$	*!				
$\begin{array}{c} \text{☞} \quad H_1 \quad H_2 \\ \quad \wedge \\ \text{ba} \ \eta\text{ga} \ \text{gu} \ \text{ru} \end{array}$					*
$\begin{array}{c} H_1 \quad H_2 \\ \wedge \quad \\ \text{ba} \ \eta\text{ga} \ \text{gu} \ \text{ru} \end{array}$			*!		*
$\begin{array}{c} H_1 \\ \wedge \\ \text{ba} \ \eta\text{ga} \ \text{gu} \ \text{ru} \end{array}$				*!	
$\begin{array}{c} H_1 \\ \wedge \quad \wedge \quad \wedge \\ \text{ba} \ \eta\text{ga} \ \text{gu} \ \text{ru} \end{array}$				*!	
$\begin{array}{c} H_{1,2} \\ \wedge \quad \wedge \quad \wedge \\ \text{ba} \ \eta\text{ga} \ \text{gu} \ \text{ru} \end{array}$		*!			

We have already seen (§3.5) that the words [báŋgá] and [gúrú] both have a single H tone. The tableau neatly shows how the language ranks the disadvantages of the various solutions. The first candidate violates the production-OCP constraint, which simply punishes any two identical adjacent elements. The faithfulness constraint violated in the winning candidate [báŋgàgúrú] is the deletion of an underlying association link (“A”) from the first H (H_1) to the second syllable. The third candidate, with rightward instead of leftward tone slip, violates ANCHOR-L, which says that “if an output syllable is the leftmost syllable in a tone span then its input correspondent is the leftmost syllable in a tone span”: thus, *ru* is leftmost in the output but not in the input. The fourth and fifth candidate have no correspondent of H_2 in the output, so they violate MAX-IO (T), where T stands for “tone” (the only tone in Shona is a H tone). The sixth candidate satisfies all faithfulness constraints by invoking the special trick of *multiple correspondence* (McCarthy & Prince 1995:371): the single surfacing H tone corresponds to both underlying H tones.

An account with the OCP as a constraint in the perception grammar works rather differently. It is clear, for instance, that the fifth and sixth candidates are pronounced in the same way, so their perceptual surface structures must be identical. If we talk about perceptual recovery, then multiple correspondence cannot be a solution: if the input contains two H tones, and the output only one, then one of the two has been deleted if the perception grammar cannot convert a single H into two. A candidate with four high-toned syllables, therefore, violates *DELETE (H) even if it does not violate MAX-IO (T). A more difficult case is posed by the first candidate in (48). Since Shona only has simple sequences of high- and low-toned syllables (i.e. no such complications as downstep between consecutive high-toned syllables), listeners must perceive stretches of high-toned

syllables with a single H tone, as in (26), so that the first candidate in (48) can never arise as a perceptual surface form. From the sixteen possible tone sequences for four syllables, tableau (49) evaluates the four that seem relevant:

(49) *The OCP as a perception constraint*

	$\begin{array}{c} \text{H} \quad \text{H} \\ \diagdown \quad \diagup \\ \# \text{ba} \cdot \eta \text{ga} \# \text{gu} \cdot \text{ru} \# \end{array}$	*INSERTPATH ([_H , σ])	*DELETE (H)	*DELETEPATH (H, σ)
[báŋgágúru] →	$\begin{array}{c} \text{H} \\ \diagdown \quad \diagup \quad \diagdown \quad \diagup \\ \# \text{ba} \eta \text{ga} \text{gu} \text{ru} \# \end{array}$		*!	
☞ [báŋgàgúru] →	$\begin{array}{c} \text{H} \quad \text{H} \\ \quad \diagdown \\ \# \text{ba} \eta \text{ga} \text{gu} \text{ru} \# \end{array}$			*
[báŋgágùru] →	$\begin{array}{c} \text{H} \quad \text{H} \\ \diagdown \quad \\ \# \text{ba} \eta \text{ga} \text{gu} \text{ru} \# \end{array}$	*!		*
[báŋgágùrù] →	$\begin{array}{c} \text{H} \\ \diagdown \\ \# \text{ba} \eta \text{ga} \text{gu} \text{ru} \# \end{array}$		*!	**

There are many differences between this account and Myers'. The first difference is in the representations themselves. Shona has the remarkable property that left edges of surface tone stretches (i.e. sentence-initial high-toned syllables, or high-toned syllables following a low-toned syllable) *always* signal a left edge of a word-level tone stretch (i.e., Myers' ANCHOR-L constraint is unviolated in postlexical phonology). This guarantees a reasonable recoverability of left tone edges, and the perception grammar will therefore pay special attention to left edges. In the forms in (49), I illustrate this special attention by putting the "H" of every tone stretch vertically above its leftmost syllable, thus suggesting that this leftmost syllable is the *head* of a tone stretch. Because of the commensurability between underlying and perceptual form, this representation applies to the underlying form as well as to the perceptual surface form.

Other differences are in the interpretations of the constraints. The constraint *DELETE (H) is violated even in the first candidate, simply because it contains fewer H tones (if we pay attention to the location of the heads of the two underlying tone stretches, we must conclude that it is the second H that has been deleted). The constraint *DELETEPATH (H, σ) is violated for every deleted underlying link (or *path* in the terminology by Archangeli & Pulleyblank 1994) from a H to a syllable (or to a vowel, if you like), i.e. it is violated for every underlying link that does not appear on the surface. This constraint is therefore violated twice in the fourth candidate, in contrast with Myers' MAX-IO (A), which is not violated if its tone is deleted (this alternative interpretation would evaluate the fourth candidate in (48) as better than the fifth). The two *DELETE constraints are more appropriate for evaluating faithfulness than the two MAX-IO constraints, which evaluate correspondence rather than similarity. The third constraint in (49) is equivalent to ANCHOR-L, though its formulation as *INSERTPATH ([_H, σ]) takes advantage of the perception of tone heads. Thus, the third candidate violates this constraint, because in this candidate a tone head is perceived (on *ru*) that does not occur

in the underlying form. We must note here that all three non-optimal candidates in (49) violate the similar constraint *DELETEPATH ($[\text{H}, \sigma]$), since all of them delete the underlying tone head on *gu*. Though this constraint could have done the job in (49), it cannot be a high-ranked constraint, since tone heads are never inserted in Shona, though they are often deleted, like the one on *gu* in (20). The use of *DELETEPATH ($[\text{H}, \sigma]$) instead of ANCHOR-L acknowledges the idea that many alignment constraints can be regarded as faithfulness constraints for underlying associations. This makes explicit the idea that the functional rationale behind the existence of alignment constraints is that they contribute to perceptual recoverability.

So we have seen that the assumption of a perception grammar can handle a typical tonal OCP effect rather well, reducing most relevant constraints in the production grammar to faithfulness. But there is an additional advantage, in that Myers' account violates the recoverability assumption (§5.3), as we will see in the next section.

6.3 The problem with the OCP in the production grammar

In the previous section, we saw an example of tone slip, which occurs in Shona if a word that ends in at least two high-toned syllables is juxtaposed to a word that starts with a H tone. No such tone slip occurs if the first word ends in a single high-toned syllable, as in 'big hoe', which is simply pronounced [bàdzágúru], i.e. as a concatenation of the separate words. As explained in §3.5, this form must be perceived with a single H tone on its last three syllables. However, Myers' (1997:fn.21) account is different:

(50) *An OCP violation (Myers 1997)?*

$\begin{array}{c} \text{H}_1 \\ \\ \text{6a dza} \end{array} + \begin{array}{c} \text{H}_2 \\ \wedge \\ \text{gu ru} \end{array}$	MAX-IO (T)	UNIFORMITY	ANCHOR-L	OCP	MAX-IO (A)
$\begin{array}{c} \text{H}_{1,2} \\ \wedge \\ \text{6a dza gu ru} \end{array}$		*!			
$\begin{array}{c} \text{H}_1 \quad \text{H}_2 \\ \quad \wedge \\ \text{6a dza gu ru} \end{array}$				*	
$\begin{array}{c} \text{H}_1 \\ \\ \text{6a dza gu ru} \end{array}$	*!				
$\begin{array}{c} \text{H}_2 \\ \wedge \\ \text{6a dza gu ru} \end{array}$	*!				
$\begin{array}{c} \text{H}_1 \quad \text{H}_2 \\ \quad \wedge \\ \text{6a dza gu ru} \end{array}$			*!		
$\begin{array}{c} \text{H}_1 \quad \text{H}_2 \\ \quad \\ \text{6a dza gu ru} \end{array}$			*!		*

The constraint ranking is still compatible with (48). The interpretation of the winning candidate must be that it is pronounced [bàdzágúru]. But this winner has two H tones

only because the underlying form has two H tones. Some underlying forms such as the word [kùtɛ̀ngésá] ‘to sell’ arguably have a single H tone. Therefore, the phonetically identical tone contours in [bàdzágúru] and [kùtɛ̀ngésá] have a different number of H tones on the surface, forced by the high ranking of the faithfulness constraint MAX-IO (T). This violates the recoverability assumption (§5.3), which states that phonetically identical forms have identical phonological surface structures. Thus, in one and the same language, the phonetic form LHHH can have two different phonological surface representations, depending on the underlying form. This “neutralization in phonetic implementation” violates the recoverability assumption. Therefore, the existence of a violable OCP in the production grammar is incompatible with that assumption. No such violations of this assumption are possible if the OCP is in the perception grammar (tableau (51)), because the surface form is derived from the phonetic form without any access to the underlying form, so if the phonetic forms are identical, the perceptual forms must be identical as well, regardless of any differences in the underlying forms.

(51) *No OCP violation*

	H H / # 6a·dza # gu·ru #	*INSERTPATH ([_H , σ)	*DELETE (H)	*DELETEPATH (H, σ)
☞ [bàdzágúru] →	H / # 6a dza gu ru #		*	
[bàdzágùru] →	H # 6a dza gu ru #		*	*!*
[bàdzàgúru] →	H / # 6a dza gu ru #		*	*!
[bádzàgúru] →	H H / # 6a dza gu ru #	*!		
[bádzágùru] →	H H # 6a dza gu ru #	*!		**

The same three constraints as in (49) work here. It appears that the inviolable constraint against inserting a tone head must outrank the constraint against tone deletion (and *DELETEPATH ([_H, σ) as well, cf. the discussion about this in §6.2). The winner is ultimately determined by *DELETEPATH (H, σ), which is crucially different from MAX-IO (A), which cannot distinguish between the first three candidates.

6.4 The OCP as a rule blocker

In Shona (Myers 1997), a word-final high tone spreads to the next word, if that word starts with at least two low-toned syllables, e.g. [kùtɛ̀ngésá] ‘to sell’ + [sàdzà] ‘porridge’ → [kùtɛ̀ngésásàdzà] ‘to sell porridge’.

(52) *Spreading in Shona (Myers 1997)*

	ANCHOR-L	BOUND	SPECIFY (T)	DEP-IO (A)
			***!	
			**	*
		*!	*	**
	*!		**	*

Myers' constraint BOUND says that successive syllables in a tone span must be in different domains (words), and it must be dominated by MAX-IO (T) and MAX-IO (A) in order not to break the analyses in (48) and (50). Myers' constraint SPECIFY (T) says that a syllable must be associated with a (high) tone, and it must be dominated by DEP-IO (T), a constraint against insertion of a non-underlying H, in order that underlyingly low-toned words surface as low in isolation. The constraint DEP-IO (A), finally, punishes insertion of association lines.

This time, the recoverability account of Shona spreading is analogous to Myers':

(53) *Spreading in Shona*

	*INSERTPAT H ([_H σ])	*SHIFT (H], 2)	MAXIMUM (H)	*SHIFT (H], 1)
			***!	
			**	*
		*!	*	*
	*!		**	

Of course, the surface forms cannot contain unrecoverable word boundaries as in (52). Since we only consider sentence-level phonology here (Myers' formulation of BOUND aimed at incorporating the iterative spreading in monosyllabic word-internal clitics as well), the constraint BOUND can be replaced by a faithfulness constraint against rightward

displacement of underlying right edges of tone spans by at least two syllables, i.e. *SHIFT (H), Right, 2). There are many differences (Boersma 1998:197) between such anti-displacement faithfulness constraints and McCarthy & Prince's (1993) alignment constraint ALIGN, most notably the dependence of its ranking on the amount of displacement (higher ranked if farther displaced). The constraint SPECIFY is replaced by a faithfulness constraint that aims at maximal expression of an available H tone. By ranking it with respect to a *SHIFT family, we predict a typology of spreading distance.

The OCP effect associated with tone spreading is that this spreading is blocked if the second syllable of the second word has a high tone: [kùtɛ̀ŋgésá̀bàdzá] 'to sell a hoe'. The cause, of course, is that the form with spreading ([kùtɛ̀ŋgésá̀bàdzá]) would be perceived with a single H, thus violating *DELETE (H), which has to outrank the spreading imperative MAXIMUM (H).

6.5 Some adjacent identical tones

From the discussion of Shona, it would seem that in a language in which two consecutive high-toned syllables are perceived as a single H, it is impossible to perceive two different H tones on consecutive syllables, i.e. that surface structures such as the first candidate in (48) are universally impossible. This is not true. Two different H tones on consecutive syllables can be perceived if they are recoverable, for instance if they are not pronounced as two consecutive high-toned syllables! An example is provided by Kishambaa (Odden 1982, 1986), in which [nwáná] 'child', with a single two-syllable H tone, is juxtaposed to [dú] 'only', giving [nwáná¹dú] 'only a child', where the high tone on the third syllable is downstepped, i.e. lower than the preceding two syllables but higher than a low tone would be. This gives a third possibility: the perception grammar will map [[HHH]] to /H/, [[HHL]] to /HL/, and [[HH¹H]] to /HH/. The constraint LCC (tone: H; \acute{V} | σ] $\acute{\sigma}$ | \acute{V}) will be ranked high, i.e. the downstep is sufficient intervening material to perceive the two high-tone cues as separate H tones. This case supports the recoverability assumption: adjacent identical tone elements are possible even in a tone language, provided that they are acoustically different from a single-tone stretch, for instance if a boundary can be perceived.

Even in Shona, word boundaries can be perceived, because penultimate syllables are lengthened (Doke 1931, Myers 1987). In principle, it would have been possible to distinguish [bà:dzágú:rú] from [kùtɛ̀ŋgésá:] (§3.5) after all, and perhaps perceive two separate H tones on [bà:dzágú:rú] after inserting a word boundary on the basis of vowel lengthening. However, we know that Shona listeners do not do this, because if they did, there would be no tone retraction in [bá:ŋgàgú:rú], since the form *[bá:ŋgágú:rú] would be perceived as identical to the underlying form.

Conclusion of chapter 6:

The OCP in the perception grammar turns out to handle the linguistic data well. Faithfulness comes to replace correspondence, uniformity, anchoring, and alignment. Moreover, expressing the OCP as a violable constraint in an Optimality-Theoretic *production* grammar runs into the problem that faithfulness constraints can force different phonological surface representations for phonetically identical forms, whereas expressing the OCP as a violable constraint in an Optimality-Theoretic *perception* grammar does not run into this recoverability problem.

7 Conclusion

This article centred on the role of the perceptual recoverability of economical representations in the evaluation of faithfulness constraints in the production grammar, focussing on phenomena that have been traditionally described as OCP effects, while noting that these OCP effects are special cases of a more general process of sequential perceptual integration. In passing, we noted that a grammar model that wants to incorporate functional principles into phonology and still maintain economical representations in the lexicon, has to consist of three different grammars for production, perception, and recognition. Summarizing the conclusions of the consecutive chapters, the reasoning went as follows:

1. The phonological surface form involved in the evaluation of faithfulness constraints in an Optimality-Theoretic production grammar is best modelled as an economical lexical-like representation constructed from the acoustic signal by a language-dependent process of perception.

2. This perception process maps raw sensory data onto discrete structures.

3. ‘Perception grammar’ is an appropriate term for this perception process, because this process submits to grammatical modelling and both of its tasks (categorization, sequential abstraction) are language-dependent.

4. The perception grammar is best modelled as an Optimality-Theoretic grammar, because there are conflicts within and between its various tasks (categorization, sequential abstraction), and because the autosegmental well-formedness conditions that handle sequential abstraction (OCP and LCC) must be regarded as violable if they are defined in terms of intervening material.

5. When combined with a recognition grammar and a production grammar, the existence of a perception grammar that handles categorization and sequential abstraction ensures compatibility between lexical economy and phonetically-based ranking of articulatory and faithfulness constraints.

6. The OCP in the perception grammar turns out to handle the linguistic data well. Faithfulness comes to replace correspondence, uniformity, anchoring, and alignment. Moreover, expressing the OCP as a violable constraint in an Optimality-Theoretic *production* grammar runs into the problem that faithfulness constraints can force different phonological surface representations for phonetically identical forms, whereas expressing the OCP as a violable constraint in an Optimality-Theoretic *perception* grammar does not run into this problem.

These points are evidence for the existence of a perception grammar and suggest that the natural place for constraints that handle sequential abstraction, i.e. OCP and LCC, is in this perception grammar.

References

- Akamatsu, Tsutomu (1997). *Japanese phonetics: theory and practice*. LINCOM Europa, München.
- Archangeli, Diana (1984). *Underspecification in Yawelmani Phonology and Morphology*. PhD dissertation, MIT, Cambridge. [New York: Garland Press, 1988]
- Archangeli, Diana (1988). Aspects of underspecification theory. *Phonology* 5. 183–207.
- Archangeli, Diana & Douglas Pulleyblank (1994). *Grounded Phonology*. Cambridge, Mass.: MIT Press.
- Bao, Zhi-ming (1990). *On the nature of tone*. PhD dissertation, MIT, Cambridge.

- Boersma, Paul (1989). Modelling the distribution of consonant inventories by taking a functionalist approach to sound change. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **13**. 107–123.
- Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam* **21**. 43–58.
- Boersma, Paul (1998). *Functional phonology: Formalizing the interactions between articulatory and perceptual drives*. PhD dissertation, Univ. of Amsterdam. The Hague: Holland Academic Graphics.
- Boersma, Paul (1999). Phonology-semantics interaction in OT, and its acquisition. Rutgers Optimality Archive **369**, <http://ruccs.rutgers.edu/roa.html>. To appear in Robert Kirchner, Wolf Wikeley & Joe Pater (eds.) *Papers in Experimental and Theoretical Linguistics*. Vol. **6**. Edmonton: University of Alberta.
- Boersma, Paul (2000). Learning a grammar in Functional Phonology. In Joost Dekkers, Frank van der Leeuw & Jeroen van de Weijer (eds.) *Optimality Theory: Phonology, syntax, and acquisition*. Oxford University Press. 465–523. [ch. 14 of Boersma 1998]
- Boersma, Paul & Bruce Hayes (to appear). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*.
- Chen, Matthew (1970). Vowel length variation as a function of the voicing of the consonant environment. *Phonetica* **22**. 129–159.
- Chomsky, Noam (1964). *Current issues in linguistic theory*. The Hague: Mouton.
- Chomsky, Noam (1995). *The minimalist program*. Cambridge: MIT Press.
- Chomsky, Noam & Morris Halle (1968). *The sound pattern of English*. New York: Harper and Row.
- Cohen, A., C.L. Ebeling, P. Eringa, K. Fokkema & A.G.F. van Holk (1959). *Fonologie van het Nederlands en het Fries*. The Hague: Martinus Nijhoff.
- Cohn, Abigail (1990). Phonetic and phonological rules of nasalization. PhD dissertation, UCLA. *UCLA Working Papers in Phonetics* **76**. i–224.
- Cole, Ronald A. & Brian Scott (1974). Toward a theory of speech perception. *Psychological Review* **81**. 348–374.
- Delattre, Paul (1954). Les attributs acoustiques de la nasalité vocalique et consonantique. *Studia Linguistica* **8**. 103–109.
- Denes, Paul (1955). Effect of duration on the perception of voicing. *Journal of the Acoustical Society of America* **27**. 761–764.
- Doke, Clement (1931). *A comparative study in Shona phonetics*. Johannesburg: University of the Witwatersrand Press.
- Dupoux, Emmanuel & Jacques Mehler (1990). Monitoring the lexicon with normal and compressed speech: frequency effects and the prelexical code. *Journal of Memory and Language* **29**. 316–335.
- Eilers, Rebecca, D.K. Oller, Richard Urbano & Debra Moroff (1989). Conflicting and cooperating cues: perception of cues to final consonant voicing by infants and adults. *Journal of Speech and Hearing Research* **32**: 307–316.
- Eimas, Peter D. & John D. Corbit (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology* **4**. 99–109.
- Flemming, Edward (1995). *Auditory representations in phonology*. PhD dissertation, UCLA.
- Fortune, George (1955). *An analytical grammar of Shona*. London: Longmans, Green and Co.
- Foss, Donald J. & Michelle A. Blanck (1980). Identifying the speech codes. *Cognitive Psychology* **12**. 1–31.
- Foss, Donald J. & David A. Swinney (1973). On the psychological reality of the phoneme: perception, identification and consciousness. *Journal of Verbal Learning and Verbal Behavior* **12**. 246–257.
- Fowler, Carol (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics* **14**. 3–28.
- Fowler, Carol (1992). Vowel duration and closure duration in voiced and unvoiced stops: there are no contrast effects here. *Journal of Phonetics* **20**. 143–165.
- Gibson, Edward & Kenneth Wexler (1994). Triggers. *Linguistic Inquiry* **25**. 407–454.
- Goldsmith, John (1976). *Autosegmental Phonology*. PhD thesis, MIT, Cambridge. [New York: Garland, 1979]
- Gussenhoven, Carlos & Peter van der Vliet (1999). The phonology of tone and intonation in the Dutch dialect of Venlo. *Journal of Linguistics* **35**. 99–135.
- Halle, Morris (1959). *The sound pattern of Russian*. The Hague: Mouton.
- Hayes, Bruce (1986). Inalterability in CV phonology. *Language* **62**. 321–351.
- Hayes, Bruce (1996). Phonetically driven optimality-theoretic phonology. Handout of a LOT course, Utrecht.
- Hayes, Bruce (1999). Phonetically-driven phonology: the role of Optimality Theory and Inductive Grounding. In Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer & Kathleen Wheatley (eds.) *Functionalism and Formalism in Linguistics*, Vol. I: *General Papers*. Amsterdam: John Benjamins. 243–285. [ROA **158**, 1996]
- Hayes, Bruce (2000). Gradient well-formedness in Optimality Theory. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.) *Optimality Theory: Phonology, syntax, and acquisition*. Oxford: Oxford University Press. 88–120.
- Hayes, Bruce & Margaret MacEachern (1998). Quatrain form in English folk verse. *Language* **74**. 473–507.

- Hernández-Chávez, Eduardo, Irene Vogel & Harold Clumeck (1975). Rules, constraints and the simplicity criterion: An analysis based on the acquisition of nasals in Chicano Spanish. In Charles A. Ferguson, Larry M. Hyman & John J. Ohala (eds.) *Nasálfest*. Stanford University. 231–248.
- Hockett, Charles. F. (1965). Sound change. *Language* **41**. 185–205.
- Hogan, John T. & Anton J. Rozsypal (1980). Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant. *Journal of the Acoustical Society of America* **67**: 1764–1771.
- House, A.S. (1957). Analog studies of nasal consonants. *Journal of Speech and Hearing Disorders* **22**:190–204.
- Jakobson, Roman, Colin Cherry & Morris Halle (1953). Toward the logical description of languages in their phonemic aspect. *Language* **29**. 34–46.
- Jun, Jongho (1995). Place assimilation as the result of conflicting perceptual and articulatory constraints. *West Coast Conference of Formal Linguistics* **14**. 221–237.
- Jusczyk, Peter W., Derek Houston & Mara Goodman (1998). ‘Speech perception during the first year.’ In Alan Slater (ed.): *Perceptual development: visual, auditory, and speech perception in infancy*. Hove: Psychology Press. 357–388.
- Kaye, Jonathan (1971). Nasal harmony in Desano. *Linguistic Inquiry* **2**. 37–56.
- Keer, Edward (1999). *Geminates, the OCP, and the nature of CON*. PhD dissertation, Rutgers University, New Brunswick, New Jersey.
- Kiparsky, Paul (1982). From cyclic phonology to lexical phonology. In Harry v.d. Hulst & Norval Smith (eds.) *The structure of phonological representations*. Volume I. Dordrecht: Foris. 131–175.
- Kirchner, Robert (1998). *Lenition in phonetically-based Optimality Theory*. PhD dissertation, UCLA.
- Ladefoged, Peter & Ian Maddieson (1996). *The sounds of the world's languages*. Oxford: Blackwell.
- Lahiri, Aditi & William Marslen-Wilson (1991). The mental representation of lexical form: a phonological approach to the recognition lexicon. *Cognition* **38**. 245–294.
- Leben, William (1973). *Suprasegmental phonology*. PhD dissertation, MIT, Cambridge Mass. [New York: Garland Press, 1980]
- Leoni, F.A., F. Cutugno & R. Savy (1995). The vowel system of Italian connected speech. *Proceedings of the XIIIth International Congress of Phonetic Sciences* **4**: 396–399.
- Liberman, Mark (1993). Optimality and optionality. Ms. Univ. of Pennsylvania, Philadelphia. [not seen]
- Liberman, Alvin M. & Ignatius G. Mattingly (1985). The motor theory of speech perception revised. *Cognition* **21**. 1–36.
- Maeda, Shinji (1993). Acoustics of vowel nasalization and articulatory shifts in French nasal vowels. In Marie K. Huffman & Rena A. Krakow (eds.) *Phonetics and phonology, Volume 5: Nasals, nasalization, and the velum*. San Diego, Calif.: Academic Press. 147–167.
- McCarthy, John (1986). OCP effects: gemination and antigemination. *Linguistic Inquiry* **17**. 207–263.
- McCarthy, John (1988). Feature geometry and dependency: a review. *Phonetica* **45**. 84–108.
- McCarthy, John (1999). Sympathy and phonological opacity. *Phonology* **16**. 331–399.
- McCarthy, John & Alan Prince (1993). Generalized alignment. In Geert Booij & Jaap van Marle (eds.) *Yearbook of Morphology 1993*. Dordrecht: Kluwer. 79–153.
- McCarthy, John & Alan Prince (1995). Faithfulness and reduplicative identity. In Jill Beckman, Laura Walsh Dickey & Suzanne Urbanczyk (eds.) *Papers in Optimality Theory*. University of Massachusetts Occasional Papers **18**. Amherst, Mass.: Graduate Linguistic Student Association. pp. 249–384. [Rutgers Optimality Archive **60**, <http://rucss.rutgers.edu/roa.html>]
- McQueen, James M. & Anne Cutler (1997). Cognitive processes in speech perception. In William J. Hardcastle & John Laver (eds.) *The handbook of phonetic sciences*. Oxford: Blackwell. 566–585.
- Mehler, Jacques, Jean Yves Dommergues, Uli Frauenfelder & Juan Segui (1981). The syllable’s role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior* **20**. 298–305.
- Mohanan, K.P. (1986). *The theory of Lexical Phonology*. Reidel, Dordrecht.
- Mohanan, K.P. (1993). Fields of attraction in phonology. In John Goldsmith (ed.) *The last phonological rule: Reflections on constraints and derivations*. University of Chicago Press. 61–116.
- Myers, Scott (1987). *Tone and the structure of words in Shona*. PhD dissertation, University of Massachusetts, Amherst. [New York: Garland Press, 1991]
- Myers, Scott (1997). OCP effects in Optimality Theory. *Natural Language and Linguistic Theory* **15**. 847–892.
- Myers, Scott (1998). AUX in Bantu morphology and phonology. In Larry M. Hyman & Charles W. Kisseberth (eds.) *Theoretical aspects of Bantu tone*. Stanford, Calif.: CSLI. 231–264.
- Norris, Dennis & Anne Cutler (1988). The relative accessibility of phonemes and syllables. *Perception & Psychophysics* **43**. 541–550.
- Odden, David (1981). Problems in tone assignment in Shona. PhD dissertation, University of Illinois at Champaign-Urbana.
- Odden, David (1982). Tonal phenomena in Kishambaa. *Studies in African Linguistics* **13**. 177–208.
- Odden, David (1986). On the role of the Obligatory Contour Principle in phonological theory. *Language* **62**. 353–383.
- Odden, David (1988). Anti antigemination and the OCP. *Linguistic Inquiry* **19**. 451–475.

- Odden, David (1995). Tone: African languages. In John Goldsmith (ed.) *The handbook of phonological theory*. Oxford: Blackwell. 444–475.
- Ohala, John J. (1975). Phonetic explanations for nasal sound patterns. In Charles A. Ferguson, Larry M. Hyman & John J. Ohala (eds.) *Nasálfest: Papers from a symposium on nasals and nasalization*. Linguistics Department, Stanford University, California. 289–316.
- Peterson, Gordon E. & Ilse Lehiste (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America* **32**. 693–703.
- Piggott, Glyne (1992). Variability in feature dependency: The case of nasality. *Natural Language and Linguistic Theory* **10**. 33–78.
- Pinker, Steven (1997). *How the mind works*. New York & London: Norton.
- Pisoni, David B. & Paul A. Luce (1987). Acoustic-phonetic representations in word recognition. *Cognition* **25**. 21–52.
- Postal, Paul M. (1968). *Aspects of phonological theory*. New York: Harper and Row.
- Powers, William T. (1973). *Behavior: The control of perception*. Chicago: Aldine.
- Prince, Alan & Paul Smolensky (1993). *Optimality Theory: Constraint interaction in generative grammar*. Rutgers University Center for Cognitive Science Technical Report **2**.
- Raphael, Lawrence J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *Journal of the Acoustical Society of America* **51**: 1296–1303.
- Raphael, L.J., M.F. Dorman, F. Freeman, & C. Tobin (1975). Vowel and nasal duration as cues to voicing in word-final stop consonants: spectrographic and perceptual studies. *Journal of Speech and Hearing Research* **18**. 389–400.
- Reynolds, William (1994). Variation and phonological theory. PhD dissertation, University of Pennsylvania, Philadelphia.
- Saussure, Ferdinand de (1916). *Cours de linguistique générale*. Edited by Charles Bally & Albert Sechehaye in collaboration with Albert Riedlinger. Paris: Payot & C^{ie}. [2nd edition, 1922]
- Segui, Juan, Uli Frauenfelder & Jacques Mehler (1981). Phoneme monitoring, syllable monitoring and lexical access. *British Journal of Psychology* **72**. 471–477.
- Smith, Richard & Connie Smith (1971). Southern Barasano phonemics. *Linguistics* **75**. 80–85.
- Steriade, Donca (1987). Redundant values. In A. Bosch, B. Need & E. Schiller (eds.) *CLS 23: Papers from the Parasession on Autosegmental and Metrical Phonology*. Chicago Linguistic Society. 339–362.
- Steriade, Donca (1993). Closure, release, and nasal contours. In Marie Huffman & Rena Krakow (eds.) *Nasals, nasalization, and the velum*. San Diego, Calif.: Academic Press. 401–470.
- Steriade, Donca (1995). Positional neutralization. Unfinished manuscript. UCLA.
- Tesar, Bruce & Paul Smolensky (1993). *The learnability of Optimality Theory: an algorithm and some basic complexity results*. Ms. Department of Computer Science & Institute of Cognitive Science, University of Colorado at Boulder. [ROA **2**]
- Tesar, Bruce & Paul Smolensky (1998). Learnability in Optimality Theory. *Linguistic Inquiry* **29**. 229–268.
- Trubetzkoy, Nikolaj (1939). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck & Ruprecht.
- Urbanczyk, Suzanne (1999). Double reduplications in parallel. In René Kager, Harry van der Hulst & Wim Zonneveld (eds.) *The prosody-morphology interface*. Cambridge University Press. 390–428. [Rutgers Optimality Archive **73**, <http://rucss.rutgers.edu/roa.html>, 1995]
- Walker, Rachel (1998). *Nasalization, neutral segments, and opacity effects*. PhD dissertation, University of California, Santa Cruz.
- Wardrip-Fruin, Carolyn (1982). On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue to voicing in final stop consonants. *Journal of the Acoustical Society of America* **71**: 187–195.
- Wickelgren, Wayne A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review* **76**. 1–15.
- Yip, Moira (1988). The obligatory contour principle and phonological rules: a loss of identity. *Linguistic Inquiry* **19**. 65–100.
- Zubritskaya, Katya (1997). Mechanism of sound change in Optimality Theory. *Language Variation and Change* **9**. 121–148.