Rule inversion without rules

Marc van Oostendorp, Meertens Instituut, Amsterdam

0. Introduction

Hypercorrection has often been analysed as a 'rule inversion' process in the sense of Vennemann (1972, 1974, 1978, 1986) and the formal simplicity of this approach (a grammar which has the rule α → β acquires a new rule β → α) has sometimes been used as an argument for a rule-based analysis of natural language grammar. Halle and Idsardi (1997), for instance, argue that 'rule inversion' type hypercorrection phenomena are incomprehensible in an optimality theoretic framework, based on output constraints only. Intuitively, this is correct. If we have a rule α → β, this means in OT terms that β is 'more optimal' than α; otherwise, such a change would not be warranted. Yet if at the same time we have a rule β → α, we have to conclude that α is more optimal than β. This is an undesirable effect; at least one of the two changes thus has to unnatural.

In this paper I argue that on closer inspection, the crucial examples can be reanalysed more succesfully as something similar to 'conspiracies'. What seems to be happening, at least in the examples under discussion here, is that the original rule and the hypercorrection 'conspire' to turn a phonemic difference into an allophonic one. The rules are not *exactly* the mirror image of one another. Rather, their combined effect is that we find α in one phonological context and β in another. The two elements tend to be no longer in complementary distribution; their occurrence is 'predictable'. It is hard to capture this in a rule-based account, just as it is hard to capture any kind of conspiracy in which two rules cooperate to give a certain output pattern.

In this paper, I will look into two hypercorrection / rule inversion phenomena in some detail: r insertion in non-rhotic dialects of English, and vowel lengthening in Betuwe Dutch. In section 1, I first discuss the reasons why non-rhotic dialects of English have been seen as an argument for rules, taking Halle and Idsardi (1997) as an example analysis. In section 2, I then discuss some of the recent literature which indicates that a closer analysis into the relevant representations (e.g. the structure of /r/) can shed more light on this issue, weakening the arguments in favour of a rule-based account considerably. In section 3, I then show how the new insights into the representations can actually be rendered quite succesfully into a constraint-based analysis, giving us more insight into the actual nature of the processes involved. Section 4 then shows how this is even more true for the processes of vowel shortening and vowel lengthening in Betuwe Dutch. Section 5 deals with some remaining topics: a more problematic case of hypercorrection is discussed, and we briefly go into the question how hypercorrection might develop diachronically under the assumptions made here.

1. Non-Rhotic Dialects of English as an Argument for Rules

We first study the basic facts of the non-rhotic dialects of English. Dialects of this type are spoken in many parts of the English-speaking world: in the UK (Jones 1956, Wells 1982, Giegerich 1992, 1999, Harris 1994, McMahon 2000), the United States (Kurath and McDavid 1961, Labov 1966, McCarthy 1991, 1993, 1999) and South Africa (Wells 1982, Giegerich 1992).

The core facts are that underlying /r/ is deleted before pause or a consonant (1a); and that /r/ appears in hiatus context (1b) after a [-high] vowel. It does not occur after high vowel (1c).

(1)	a.	*car* [ka]	(*the) car is*	[kar ɪs]
	b.	(*alge)bra* [bra]	(*alge)bra is*	[brar ɪs]
	c.	*see* [sij]	*see it*	[sij ɪt]

It is usually assumed that the deletion process of (1a) has been diachronically prior to the insertion of (1b). There also seem to be variants that have the former process but not the latter, but no variants in which this is the other way around. It therefore is sometimes assumed that r insertion has been the result of hypercorrection. Speakers observe that the result of there grammar is that r's are deleted in positions where this does not happen in some more prestigious variant and therefore they add a rule to their grammar which inserts r's in similar positions.[1] If this new rule is 'overgeneralized', we get results such as (1b). It is often pointed out that r insertion is a fully productive process in nonrhotic dialects. New forms, such as loan words (2a), acronyms (2b) and forms in foreign languages (2c) routinely receive an intrusive [r] (the following data are from Wells 1982:226, McMahon 2000:243):

(2)	a.	*the social milieu*[r] *of Alexander Pope*	
		the junta [r] *in Chile*	
	b.	*as far as the BUPA* [r] *is concerned*	
	c.	German:	*ich habe* [r] *einen Hund*
		Latin:	*dona*[r] *eis requiem*

In this paper we concentrate on Boston English, since this is the variant that most recent papers discuss, but some attention will be paid to some other variants as well.[2] According to McCarthy (1999), "The vowels that trigger r

[1] Synchronically, this can (no longer) be an adequate description of the relevant facts for many variants, since many prestigious forms of English (RP for instance) are themselves r dropping.

[2] One type of system which is not discussed here is the one which is "characteristic of certain conservative dialects spoken in the Upper South of the United States" according to Harris (1994:232) and in which the r also does not appear in the onsets of weak syllables: it is 'dropped' in *very, bearing, star of* and *after all* and not inserted in *law of* or *saw it*. Presumably, constraints on well-formed foot structure interact in these constraints with the ones discussed

intrusion — [ə], [ʌ], [ɑ], and [ɔ] — are just exactly the word-final monophthongs that are permitted in this dialect. All other final vowels are actually diphthongs: [ij], [ej], [uw], and [ow]." This is the reason for representing the glide in (1c).

Halle and Idsardi (1997) provide an analysis of the interaction between r drop and r insertion which is based on ordered phonological rules and the Elsewhere Condition. There analysis is based on the three rules given in (2):

(2)

a. → ə / Rhyme
Nuc
[-cons] ___ [+cons]
[+son]
[-nasal]

b. → x / Rhyme Rhyme
r Nuc Nuc
[-cons] ___ [-cons]
[-high]

c. x → / Rhyme
r Nuc
[-cons] ___

The rule in (2a) is needed to insert a (predictable) schwa between vowels and liquids in forms such as *hear it* and *feel*. Rule (2c) is the 'original' r deletion rule, deleting r in codas. (2b) is the hypercorrect rule inserting r in intervocalic position (if the preceding vowel is [-high].) The rules have to be ordered in the order in which they have been given here, as is shown in the following sample derivation:

---

below in these dialects.

In South African English, "[r] sandhi is the norm within words (*hearing* etc.) but it is uncommon between words, where instead prevocalic glottal stops (*hear* [ʔ]*it*) tend to occur" (Giegerich 1999:186). Giegerich points out that Standard German 'non-rhoticity' is subject to similar requirements.This is another type of restriction that will not be taken into account here.

(3) *Halle and Idsardi (1997): Some Sample Derivations*

	see it /si ɪt/	*saw it* /sa ɪt/	*hear it* /hir ɪt/	*hear* /hir/	*feel it* /fil ɪt/
(2a)	—	—	hiər ɪt	hiər	fiəl ɪt
(2b)	—	sar ɪt	—	—	—
(2c)	—	—	—	hiə	—
	[sij ɪt]	[sar ɪt]	[hiər ɪt]	[hiə]	[fiəl ɪt]

Schwa insertion (2a) applies in all those cases in which a liquid follows a vowel underlyingly. R insertion (2b) follows this rule and applies in intervocalic context. R deletion (2c) is disjunctively ordered with respect to (2b) — in a rather special manner, to be discussed in some more detail below — and applies to those cases in which r is not intervocalic.

H&I provide several arguments in favour of a rule-based analysis and against an analysis based on constraints (next to the more general argument that rule inversion phenomena seem to be captured quite easily in terms of rules).

The first of these problems was noted already by McCarthy (1993:189), who probably was the first to analyse in an optimality-theoretic framework (H&I can be read in part as a reply to McCarthy 1993): "r is demonstrably not the default consonant in English". The fact that it is epenthesized thus is remarkable at least. It would be possible to assign a default status to [r] in coda positions, as H&I point out, but this would still not be unproblematic, since [r] does not play this role in the coda of a syllable with '[-high] vowels' (as H&I call them) such as we find in *see*. In order to account for the fact that **see*[r]*ing* contrasts with *saw*[r]*ing* in non-rhotic dialects, we would need a set of NoCoda-constraints distinguishing between different positions in the syllable and between different preceding vowels.

The problem thus is that the insertion of [r] in exactly this environment seems arbitrary and phonologically unmotivated. H&I assume that this kind of arbitrariness is better captured in a theory of rules than in a theory based on surface constraints.

Another problem arises according to H&I in the opaque interaction between schwa insertion between a glide and a liquid (*feel,* [fijəl]) on the one hand and r deletion on the other: in words such as *fear,* schwa is inserted in Boston English, in spite of the fact that the r which supposedly triggered the insertion has itself been deleted on the surface ([fijə]). These examples show, according to H&I, that there still should be a difference between words ending in an underlying /r/, and those *not* ending in such a segment, since a word such as *see* does not display such a behaviour and surfaces as [si(j)] rather than *[sijə]. Furthermore, we need (extrinsic) ordering of rules: the /r/ can only have been deleted *after* insertion of schwa has applied.

Another crucial point for H&I thus is that in the synchronic grammar of Boston English, both r insertion and r deletion are present and that there is an underlying contrast between words ending in an r underlyingly and those words

'just' ending in an underlying vowel. This underlying contrast would be seen to surface in Level I suffixation forms such as those in (4):

(4) Volt[ej]ic alt[ər]ation
algebr[ej]ic Hom[er]ic

While the putative underlying /r/ surfaces in the forms in the righthand column, it does not surface in those cases in words in which the /r/ is not underlyingly present at all.

Although the rule-based analysis of H&I thus seems to work technically (but cf. McCarthy 1999 for some serious empirical problems), there are several conceptual problems with it.

In the first place, H&I's analysis is based on a number of rules which are purely arbitrary.[3] Each of the three rules involved (1a-c) is restricted in some ways that do not seem motivated and could have been different in a number of ways. It is not clear at all, for instance, why it is exactly the segment *r* that is deleted and inserted by the rules (1b) and (1c), and why in both cases a coda context (and not, for instance, the onset) is involved. Similarly, why both the schwa insertion rule and the r deletion rule are restricted to a context after a [-high] vowel (and not unrestricted, or restricted to different contexts).

H&I (p. 346) suggest that the fact that there are two rules governing *r* is the result of hypercorrection: "It is well known that the deletion of coda *r* is relatively widespread among English dialects and historically prior ro *r*-insertion. Speakers notice that coda *r*'s are missing in their utterances and attempt to correct this by *r*-insertion in some intervocalic contexts". Granting that this might be (the beginning of) a historical explanation for the way the 'rule inversion' phenomena at hand might have been initiated (cf. section 5), the question remains to what extent this can serve to explain the curious fact that a language acquiring child would set up such a complicated system of counteracting rules in a synchronic grammar.

Furthermore, H&I's explanation presupposes that the language system from which the speakers in question start out needs 'correction', at least as far as these speakers themselves are concerned. Even though this might be a plausible explanation for other 'hypercorrection' phenomenon (Van Oostendorp 1999), it is rather problematic in the case at hand. Non-rhoticity is quite widespread within the English language area. It is a feature of RP, which counts as a prestigious variant, at least in Britain. In my view, it does not seem to make sense to suppose that RP speakers observe that the output of their rule system has a result which needs correction by a new rule.

The **second problem** I want to note is that H&I have to assume that all speakers of a dialect of a given language have the same underlying form for any given word. In the case at hand, Boston English speakers distinguish between [idea]

[3] Cf. Harris (1994:238).

and [ka] as far as underlying representations are concerned (/idea/, /kar/), although these forms behave exactly the same in all relevant contexts on the surface. 'Native', 'monolingual' speakers of non-rhotic dialects do not have a lot of evidence that these forms are different. The point is that in H&I's view dialects of a language ultimately differ in their grammar, i.e. the rule system, only. This means that people have access to underlying contrasts that no longer play a role in their synchonic grammars.

Fortunately, H&I provide an empirical argument in favour of their position (following McCarthy 1991, 1993). According to H&I (p. 332) stems such as *Volta* and *Homer*, which are pronounced in the same way in all relevant respects in isolation contexts, start to behave differently once a Level I suffix is attached to them (*Voltaic, Homeric*, cf. (4)).

The question here is, of course, whether forms like 'Homeric' and the like are relevant if we are discussing situation in which apparently non-prestige variants are involved. Furthermore, it should be no coincidence that the non-r forms in the lefthand column all involve a front vowel. *Homeric* also has a front vowel plus an r, but there are no cases, as far as I now, without a back vowel but without *r* (this would be a form such as *[alt[əj]ation]). The reason for this gap is a mystery from the point of view of H&I's proposal. They need to stipulate a constraint on underlying representations to the effect that [-high] vowels need to be followed by /r/ underlyingly. Notice that this would be the third time that this particular feature would pop up in the analysis: it also appears in the unrelated rules (1a) and (1b).

Finally, H&I have to crucially reformulate the Elsewhere Condition. The reason for this is that the deletion of *r* is blocked both in those cases in which schwa insertion has applied and in those in which an underlying *r* could have been the result of insertion. This means that H&I's system has been set up in such a way that 'underlying' *r*'s are treated as if they were derived.[4] The distinction between underlying and derived segments thus becomes blurred; the question then arises how to account for Strict Cyclicity effects, but this is not answered by H&I. It would not have been necessary to make this complicated move, of course, if we were allowed to abandon either the assumption that *r* is ever underlying in these cases, or the idea that the grammar consists of ordered rules. H&I are not prepared to take either of these two steps. They keep to the first assumption for empirical reasons (viz. (3)), and to the second as a matter of principle. In the next section, I will show that we would be better of if we would abandon the principle.

[4] An alternative would be to order (2c) before (2b) rather than the other way around. The problem with this would be that it would violate the Elsewhere Condition, and furthermore it would involve many so-called Duke-of-York derivations in which underlying /r/'s are deleted first only to be inserted immediately afterwards. Cf. McCarthy (1999) for discussion of this point.

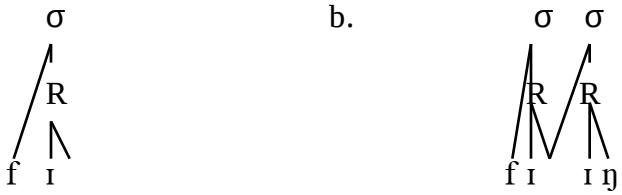
2. Non-Rhotic Dialects of English as an Argument for Representations

We thus have a constraint-based analysis and a rule-based analysis, both of which are not without problems. I think the best strategy is to take a closer look at the representations. Both in McCarthy (1991, 1993) and in H&I, the segmental representations involved are simplified: neither *r* nor schwa is decomposed into features for instance. Yet the importance of such a move for a proper understanding has been demonstrated by Ortmann (1998) and Giegerich (1999).

According to these scholars, [r] has the same phonological make-up as schwa in non-rhotic dialects. In the formalisation of Giegerich, both consist of an empty [+sonorant] root node only, that is to say, a root node which consists of the major class feature [+sonorant] with no other features attached to it.[5]

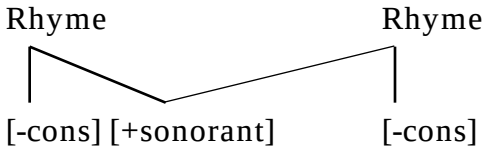
Giegerich (1999:189-190) provides us with several arguments in favour of this position. He observes, first, that [ə] and [r] are in complementary distribution: while schwa is always attached to a syllable nucleus, [r] is rather attached to marginal positions within the syllable, such as the onset and the syllable coda. Secondly, he cites Kahn (1976) and McMahon (1996) who have argued that [ə] and [r] are acoustically very similar (at least in RP). An additional argument, we might say, is that the resulting analysis is quite elegant and provides more insight into the question why it is exactly [r] that is inserted in exactly the environment after a [-high] vowel. The idea is that the representation of a word such as *fear* is (approximately) that of (5a) in an isolation context and of (5b) before a vowel:

(5) a. b.



is the symbol used by Giegerich to designate schwa and [r] alternatingly; if it occurs in the onset is interpreted as [r]; otherwise, it is interpreted as schwa. This analysis has several advantages over the one presented by H&I. The rules in (2b-c) can now be replaced by the following general resyllabification rule:

(6)



This rule is more generally applicable, as can be demonstrated on the following examples:. It is responsible not only for 'r insertion' and 'r intrusion' phenomena

[5] Ortmann assumes that there is an additional feature [-high] involved, but I do not believe this distinction is relevant for our present purposes.

— a distinction which disappears under the current theoretical assumptions — but also for gliding in cases such as those in (7):

(7) a. *I see* *see*[j]*ing*
b. *I do* *do*[w]*ing*

In order to make these analyses possible, Giegerich-Ortmann-type analyses have to assume that high vowels in English are always followed by a homorganic glide. This glide is put in the onset of the following syllable whenever this is possible (or necessary). In this view then, [r] in non-rhotic dialects is the glide of [-high] vowels in the same way as [j] is the glide for [+high, -back] vowels, and [w] for [+high, +back] vowels.

The analysis thus solves the most important problems connected to H&I's approach:

- first, at least two of the three rules have been reduced to one rule, which is much less arbitrary, but connected to general gliding processes in the phonology of English;
- secondly, we no longer have to assume that speakers all start out with the same underlying representations; specifically, the segment that surfaces as [r] has a different structure in nonrhotic dialects than in rhotic ones;
- thirdly, since the two relevant ordered rules are reduced to one, there is no need to revise the Elsewhere Condition. Indeed, this principle is not applicable to a system of one rule by definition.

In my view, then, this approach to nonrhotacism is preferable to the one presented by H&I, since it provides more insight into the nature of the process involved. Still, at least one problem connectedto H&I's approach remains. Furthermore, at least one problem H&I note in connection to constraint-based alternatives resurfaces, and furthermore, a new problem arises.

A problem that the rule-based Giegerich-Ortmann shares with H&I is that it there still are some arbitrary aspects to it. For instance, we have noted above that it has to be assumed that stem-final high vowels always develop an off-glide, and that schwa and low vowels get a final [r]. There are two ways to effectuate this. We could either assume that this is the result of a morpheme structure constraint (just stating the things we have just formulated), or that there are rules of the following type:

(7) a. Rime b. Rime

The ordering between these rules and resyllabification should be one of feeding

in order for them to work in the appropriate way. It is not clear what the function of these rules is, other than grinding out the relevant outputs.

Also the context in which these rules apply is arbitrary: at the end of the word, and not in a closed syllable. There even is a class of exceptions to this generalisation: before a liquid we do find schwa insertion, as we have seen above. Technically, this can be accomplished only by keeping the arbitrary rule (1a) in the system or, alternatively, by adding another morpheme structure constraint.

A related problem may be the following: why is the [+sonorant] syllabified in syllable coda at all? From the literature (McCarthy 1991, 1993, 1999, McMahon 1996, Harris 1994, H&I, Giegerich 1999, Ortmann 1998) we know that the *r* in *Wanda returned* is different from the one in *Wanda*[r] *entered the room*. All of these authors assume that this is a difference between a monosyllabic and an ambisyllabic consonant. Yet this problem can be easily solved if we assume some form of Lexical Phonology or some other means to formalise cyclicity (or recognize that there is something special to the final position of the word, as will be argued below).

Also one of the criticisms H&I raise against constraint-based alternatives also holds against Giegerich-Ortmann. This concerns the examples in (3), which seem to show that there should be an underlying contrast between forms such as *algebra* and *Volta* on the one hand, and *alter* and *Homer* on the other. The former really do have an underlying [r], the others don't.

(8) *Volt*[a]
Volt[ej]*ic*
Volt[ar]*is*

This criticism partly depends, of course, on the assumption that the vowel shift phenomena responsible for the [a]-[e] alternation in a pair such as *Volta-voltaic* are irrelevant for the choice rule in question. If we would be able to say that the changes in vowel quality precede all other relevant rules, there would be no problem. The facts about this word would confirm our predictions rather than disconfirm them. Words such as *alter* for one reason or another do not enter into the vowel shifting pattern. With regard to glide insertion, they behave as perfectly regular:

(9) *alter - alteration*

The real problem is with words such as *Homeric*, which have a front vowel followed by [r]. This seems hard to reconcile with the Giegerich-Ortmann position indeed. As far as I can see, there is no answer to the question why this form does not surface as either (10a) or (10b);

(10) a. *Hom*[ər]*ic* b. *Hom*[ej]*ic*

In addition to this, both implementations of the general idea seem to suffer from their own problems. In the analysis of Giegerich, it is not clear why we cannot say *see*[r]*ing?* In other words, it is not clear why spreading is obligatory in these cases in his analysis. As for an analysis along the lines of Ortmann, we would have to answer the question wheter schwa is really specified as [-high] and if so, whether we can find independent evidence for this assumption (cf. Van Oostendorp 2000).

3. Non-Rhotic Dialects of English as an Argument for Constraints

In spite of these problems, I think the approach sketched by Giegerich, Ortmann and others is on the right track. Furthermore, other than the H&I approach, it can be quite easily translated into a constraint-based framework. On top of this, most of the problems mentioned earlier disappear once we have accomplished such a translation. In which the non-rhotic r-inserting dialects differ from other (older) variants by a restructuring of lexical representations rather than by rule inversion. The idea is as follows. The r is deleted if it would only occur in the coda because of the well-established Coda Constraint (McCarthy 1993) which says that consonantal material is disallowed outside the onset. It is not always deleted completely, because of interaction with Faithfulness: only the offending purely consonantal (place and manner) features are erased, leaving a trace in the form of an empty sonorant root node, which is in most cases interpreted as schwa. The schwa in *fear* [fijə] thus has a somewhat different synchronic origin than the one in *feel* [fijəl]. There is no opacity problem, and special mechanisms such as extrinsic rule ordering, Sympathy or constraint conjunction need not be invoked.

There are independent reasons to assume that the vowels classified as [-high] by H&I are different from the others. In the first place, the [-high] vowels have a tendency to be phonetically longer than [+high] vowels. In the second place, they do not have an obvious consonantal counterpart which could act as a glide (while e.g. *see* [sij] ending in a high vowel can obviously produce [sijɲ]). We argue that these two factors have caused non-rhotic dialects to apply a merger between the [-high] vowels and rhotic consonants, which is visible in the form of colouring in some English dialects. This in turn meant a minor restructuring of the constraint ranking in the grammar, to the effect that [-high] vowels can project a rhotic offglide in the same sense that [i] projects a [j] offglide in [sijɲ].

In this way, we argue, it is better understood why r insertion is restricted to [-high] contexts; this is an advantage over the H&I approach. Another advantage is that it is easier to explain why the two processes involving r start behaving as a conspiracy, aiming at a complementary distribution of [-high] vowels (before a consonant or pause) and [r] (before a vowel) in the surface representations. It is hard to capture this effect in the rule-based approach, where we basically only have two arbitrary rules.

Let us put the relevant constraints into place. In some ways the analysis has to be a little bit different from that of Giegerich-Ortmann, if only because of the

differences in background assumptions between an OT analysis and a rule-based analysis.

The basis of the analysis of course has to be that *r* should not be allowed in the syllable coda. Like most previous analysts we assume that this is the result of a constraint on coda's:

(12) **CV**
NoCoda (Itô 1986, Prince and Smolensky 1993):
Syllables should be open (no consonantal features in the coda)
Onset: (Ramm 18xx, Prince and Smolensky 1993)
Every syllable should have an onset.

The constraints NoCoda and Onset are usually distinguished, and probably for good reason. There will be no such reason here below, and therefore we will use 'CV' as a cover term for the two constraints.

Why is r the only consonant that is sensitive for this constraint? One of the reasons may be that *r* is a virtually empty consonant in the dialects in question, as it has been argued by Giegerich-Ortmann. The difference between it and e.g. [l], [n] and [m] could be the following:

(13)

[r]	[l]	[n]	[m]
[+sonorant]	[+sonorant]	[+sonorant]	[+sonorant]
	[coronal] [lateral]	[coronal] [nasal]	[labial][nasal]

The relevant Faithfulness constraints in (14) interacting with the cover well-formedness constraint CV all have the form of Identity constraints, demanding the output feature structure to be the same as the input feature structure. Also in this case it is convenient to introduce a cover constraint, which we call Ident -F, and which covers a number of Identity constraints for different features.

(14) a. Ident-F:
Neither delete nor insert feature F
b. Ident-***F*** » CV » Ident-[±sonorant]
F is the set of all (relevant) features (except [±sonorant])

The fact that /r/ (but no other segments) can be deleted, implies that CV has become sandwiched between almost-general faithfulness and faithfulness of the root node. The high ranking of other Identity constraints forbids deletion of most segments. The fact that Ident -[±sonorant] is low-ranking means that segments with only that feature, such as /r/, can be deleted:

(15)

/kal/	IDENT-*F*	CV	IDENT-[±sonorant]
☞ [kal]		[coronal], …	
[ka]	[coronal]!		*

/kar/	IDENT-*F*	CV	IDENT-[±sonorant]
[kar]		[+sonorant]!	
☞ [ka]			*

Because the segment is almost empty as far as the constraint interaction is concerned, the underlying status of [r] now actually becomes irrelevant. If /r/ disappears from the input, we still get the same outputs:

(16)

/ka/	IDENT-*F*	CV	IDENT-[±sonorant]
[kar]		[+ sonorant]!	*
☞[ka]			

b.

/ka/ + /ɪs/	IDENT-*F*	CV	IDENT-[±sonorant]
☞ [karɪs]		[+sonorant]	*
[kalɪs]	[coronal]! [lateral]	*	*
[kaɪs]		[+ sonorant], [coronal]!	

The opposition [r]-[ə] thus has become virtually non-contrastive, just as it was in Giegerich and Ortmann's analyses: because the relevant faithfulness constraint has sunk below syllable well-formedness, it no longer causes deletion. But, as a side effect and since NOCODA and ONSET can be clustered, it can now also trigger insertion. Notice however, that in the approach presented here, it is not absolutely necessary that /r/ be completely empty.

How are rhotic dialects different from the non-rhotic ones? Two possibilities are open. First, we could suppose that /r/ has more underlying material in rhotic dialects, so that some of the Identity constraints summarised here under the general heading of IDENT-***F*** now also disallow deletion of /r/. Another possibility is that /r/ has the same features in all dialects of English, but IDENT-***F*** involves more features of /r/ in rhotic than in non-rhotic dialects. Technically, these two approached amount to the same thing:

(17) Rhotic dialects:
a.

/kar/	IDENT-***F*** (inc. features r)	CV	IDENT-[±sonorant]
☞ [kar]		[+sonorant]	
[ka]	*!		*

b.

/kar/ + /ɪs/	IDENT-***F*** (inc. features r)	CV	IDENT-[±sonorant]
☞ [karɪs]		[+sonorant]	
[kalɪs]	[lateral]!	[+sonorant], [coronal],…	*
[kaɪs]	[+sonorant]!	**	

Since the relevant features are now high-ranking, both deletion *and* insertion of r is disallowed in the rhotic dialects: r is a segment like all others.

We now turn our attention back to There is still a technical problem about gliding, which can be carried over from Giegerich's approach to the one presented here: why would we prefer a glide over an 'empty' root after a high vowel? Given the constraint ranking developed until now, both options would be equally good candidates:

(18)

/si/ + /ɪŋ/	IDENT-***F***	CV	IDENT-[±sonorant]
☞ [sijɪŋ]			*
[silɪŋ]	[lateral]!		*
☞ [sirɪŋ]			*
[siɪŋ]		*!	

A possible solution can be found in the fact that syllabic positions like to be filled and that empty segments are usually avoided if they do not serve a specific function.

d. *EMPTY: (Prince and Smolensky 1993, Van Oostendorp 2000a)
Syllabic positions (i.c. onsets) should be filled with feature material.

The constraint *EMPTY as it is formulated here could of course be seen as a more specific version of the constraint ONSET. We therefore place it alongside the other CV-constraints on syllabic well-formedness in the following tableau.

(19)

/si/ + /ɪŋ/	IDENT-***F***	CV *EMPTY	IDENT-[±sonorant]
☞ [sijɪŋ]			*
[silɪŋ]	[lateral]!		*
[sirɪŋ]		*	*
[siɪŋ]		*!	

*EMPTY could be seen as a gradient constraint, in which case inserting a glide would still be better than inserting nothing an almost empty /r/. If *EMPTY is ranked below ONSET, its introduction does not affect the cases where the /r/ is preceded by a non-high vowel. In these cases, gliding is not possible. Therefore, insertion of the almost empty /r/ is still preferable to inserting nothing at all.

We also need to explain why glides can develop also at the end of the word / phrase at least in Boston English. Another well-established constraint from the literature can help us here: the constraint FINALC.

(20) FINALC: (Prince and Smolensky 1993, McCarthy 1993, Van Oostendorp 2000a, Swets in preparation)
Words should not end in a full vowel.

McCarthy (1993) has argued in favour of this constraint on the basis of facts such as the following (see also McCarthy 1999):

(21)	a.	*I'm gonna [r] eat I'm gonna eat	b.	I said I was gonna[r], and I did. *I said I was gonna, and I did.

[r] cannot be epenthesized after a function word if this function word belongs to the same syntactic constituent (*to eat*, (20a)), but it can (and in fact should) be epenthesized between a function word and a following lexical word, if the two do not belong to the same phrase (*to, and*, (20b)).

What this shows, then, according to McCarthy (1993), is that [r] surfaces not just to satisfy the ONSET constraint, but also for reasons of FINALC: only if both constraints can be satsisfied at the same time do we find [r] epenthesis.

The same constraint could of course also be put to work in order to explain gliding at the end of the word: (phonological) words prefer to end in a glide rather than in a vowel. The fact that [r] does not surface for the same reasons could be due to interaction between *EMPTY and FINALC: if the former dominates the latter, it is not allowed to insert an (almost) empty segment just to satisfy FINALC.

FINALC has incidentally been argued to have similar repercussions in other languages (like the Tilburg dialect of Dutch) as well. In this dialect, long lax vowels and diphthongs are (almost) in complementary distribution: we find the former in the middle of words and the latter at the end of them (disregarding a few complications for which see Swets in prep., Van Oostendorp 2000c)

(22)	Tilburg Dutch:	[rɛːk] 'rich', [blɛi] 'happy'
	cf. Standard Dutch:	[rɛik] 'rich', [blɛi] 'happy'

We seem to see a similar effect in Boston English: words cannot end in a (tense) vowel and they prefer to develop a glide at the end.

Summarising what we have seen so far, the difference between rhotic and non-rhotic dialects can be described as follows:

- in rhotic dialects, all relevant faithfulness constraints outrank wellformedness, so that there is a full (distinctive) contrast;
- in non-rhotic dialects, faithfulness for r has become less important (because /r/ has lost its features). The trigger for this reranking has been that /r/ could be deleted; its result is that it can also be inserted. We thus have a

'conspiracy' of deletion and insertion to the effect that schwa and [r] start to be in complementary distribution. A distinction which formerly was phonemic is now becoming allophonic. This particular case of rule inversion therefore is a phenomenon which is not more difficult but actually *easier* to describe in terms of surface constraints than in terms of rules.

Concretely, we have established the following constraint rankings for non-rhotic and rhotic dialects of English:

(23) a. *Non-rhotic dialect*: IDENT-***F*** »CV »IDENT-[±sonorant] (= only relevant feature of [r])
b. *Rhotic dialect*: IDENT-***F*** (incl. features [r]) »CV »IDENT-[±sonorant]

Since we have gone into some more detail into the Boston English variant, we have actually partly established some of the internal rankings of the constraints summarised as FinalC for this dialect (cf. McCarthy 1993):[6]

() *Boston English*
ONSET » *EMPTY » FINALC

The differences between the two types of dialect are actually very small: the only important difference is that the features of [r] are less prominent in one group than they are in the other. It therefore should come as no surprise that we can find non-rhotic dialects in many different parts of the English-speaking world.

4. Betuwe Dutch Vowel Lengthening without Rules

If non-rhotic dialects have to be described as a conspiracy, by which two segments start behaving as if they are in complementary distribution, rather than as rule inversion, we expect the same type of analysis to be possible for other kinds of 'rule inversion' hypercorrection as well. This is confirmed by Betuwe Dutch vowel lengthening (cf. Goeman and Van Reenen 1986, Goeman 2000 for an extensive description of the empirical facts and the way these were obtained). Betuwe Dutch originally had a rule of vowel shortening before a consonant cluster, e.g. in inflected verbs. Interestingly, in the same morphological environment it also developed a rule of vowel lengthening:

(24) *Betuwe Dutch* (Goeman and Van Reenen 1986, Goeman 2000)

a.	*Vowel Shortening:*	lo:pə	*hij* lopt	‘walk – he walks’
b.	*Vowel Lengthening:*	bevə	*hij* be:ft	‘tremble – he trembles’

[6] Technically, the ranking given here does not provide us with a ready-made analysis of the fact that [r] is inserted only if it also helps to satisfy FINALC. We will not go into this here.

In order to give a concrete analysis of the phenomena involved, we need to be a little bit more explicit in our assumptions about Dutch syllable structure. For the sake of concreteness, I propose the following for Standard Dutch (Van Oostendorp 2000):

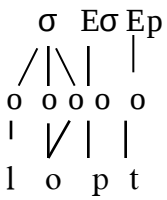
- Word-internal syllable rhymes contain maximally two positions.
- The end of the word can contain maximally one 'extrasyllabic consonant', i.e. a syllable without a (filled) nucleus.
- Voiceless coronal obstruents can be additionally licensed extraprosodically at the end of the word.

The specific way in which we want to implement notions such as 'extrasyllabicity', 'extraprosodicity', and the differences between these different forms of licensing of 'extra' material, is not relevant here.

In a Standard Dutch verbal form such as [lo:pt] 'walks' the relevant structure thus looks as in (25). The long vowel occupies (all) two positions in the rhyme; [p] is extrasyllabic and [t] is extraprosodic.

-

(25)



In a rule-based analysis, vowel shortening could look approximately as follows:

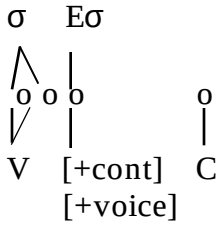
(26)

‘a long vowel shortens before a consonant cluster’

Goeman and Van Reenen (1986) analysed this as hypercorrection and more or less implicitly as a rule inversion process. What is important, however, is that the ‘hypercorrect’ lengthening is found far more often in verbs where the vowel was adjacent to an underlyingly voiced obstruent ([bevə]-[be:ft]; notice that the /v/ is devoiced to [f] in this context) or a glide ([drajə]-[dra:jt]) (cf. Goeman 2000).

In a rule-based analysis, the lengthening process would have to be arbitrarily restricted to those cases where the morpheme fits the relevant criteria:

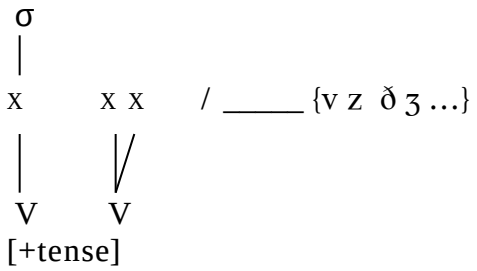
(27)



'a short vowel lengthens before a cluster of which the first segment is a voiced fricative'

We have to note various problems with this approach. In the first place, the second rule should be phonetically arbitrary (since it is just an 'inverted' rule); yet in particular the fact that it involves voiced fricatives is mirrored in other languages (such as Scottish; compare (27) to (28a), which is responsible for the pervasive length contrast illustrated in (28b)).

(28) a. Scottish Vowel Lengthening Rule (McMahon 2000)



b.

[i]	*beat*	*bean*	*greed*
[iː]	*leave*	*lease*	

Another problem is that the rules again have to be ordered by the 'special' Elsewhere Condition, or else all underlyingly long vowels before fricatives would be shortened by the original shortening rule (or we would allow Duke-of-York gambits).

Importantly, however, the original shortening rule is subject to the opposite constraints: it is often exceptionally restricted in the phonological environment before a voiced fricative. In these dialects, then, the two rules seem to conspire to lift a formerly phonemic distinction, viz. the one between long and short vowels. After both rules have applied, the former only occur before voiceless, and the latter only before voiced obstruents in 3S forms of the verb. Again, it is hard to capture this without referring to surface constraints.

The central observation, common to the analyses of both English and Dutch dialects, is that rule inversion phenomena are not arbitrary. The original rule and the inverted rule tend to create a 'conspiracy' which causes two segments to get into a state of complementary distribution. Since conspiracy effects are notoriously hard to analyse in terms of rules, rule inversion phenomena may eventually be seen as an argument *against* rules, rather than in favour of them.

A constraint-based analysis of this interaction between shortening and lengthening could be actually very similar to that of English, be it that a slightly different form of syllabic well-formedness (29a) and and a different faithfulness (29b) constraint play a role. (29a) states that extraprosodic segments are dispreferred. All segments should be in a syllable or else in an extrasyllabic position. (29b) is a constraint on identity of underlying quantity (rather than features).

(29) a. PARSE (Prosodic Licensing, Itô 1986)
Avoid extraprosodic segments.
b. IDENT-Length
Underlying length should be present on the surface.

Again, the relevant well-formedness constraint has become sandwiched between an almost general type of faithfulness, and one very specific type of faithfulness (the one that would block shortening). The reason here is that shortening has to be allowed, presumably for reasons of syllable well-formedness. A word-final consonant cluster is better after a short vowel than after a long one:
(30)

/lo:p/ + /t/	IDENT-***F***	PARSE	IDENT-Length (+, maybe, tense)
[lo:pt]		*!	
[lo:p]	[coronal],…!		*
☞ [lopt]			*

Other dialects of Dutch would differ from Betuwe in the same way that rhotic dialects of English differ from rhotic dialects. The relevant wellformedness constraint is no longer really 'sandwiched'. Again, this can be implemented in one of two ways: either the representation of the vowels is slightly different in these other dialects — we could argue for instance, that in other dialects, tenseness counts rather than length; cf. Van Oostendorp (2000) — so that the relative ranking of IDENT-Length simply is irrelevant:[7]

[7] The account as to *why* we would get laxing in these cases of course has to be a little bit different than why we get shortening in Betuwe Dutch. Cf. Van Oostendorp (2000) for discussion of this issue.

(31) Standard Dutch:

/lo:p/ + /t/	IDENT-***F*** (incl. tense)	PARSE	(IDENT-Length)
☞ [lo:pt]		*!	
[lo:p]	[coronal]!		*
[lɔpt]	[lax]!		*!

An alternative possibility is to assume that IDENT-Length has got a higher ranking in these other variants of Dutch, now making part of the IDENT-***F*** cluster. This of course gives us the same results. In any case, the Betuwe Dutch system should have developed by a lower ranking of the faithfulness constraints

But if PARSE can be sandwiched in this way, other well-formedness constraints may become similarly sandwiched, because faithfulness to length also starts ranking below them. In this case, length becomes completely undistinctive and also subject to the following universal tendency:

(32) VFG
(Voiced Fricative Generalisation, cf. Van Oostendorp 2000b)
A vowel to the left of a voiced fricative should be long.

This is of course the constraint version of the rule in (28a), and a reflex of some (possibly phonetic) universal tendency for vowels to be long in front of voiced fricatives. The fact that the original vowel process was subject to exceptions in contexts before fricatives, is an indication that at least some speakers in at least some cases entertained the ranking VFG » PARSE. (Violations of syllable well-formedness were allowed in exactly those cases where this helped a long vowel to surface before a voiced fricative.) But since » is a transitive relation, a reranking of Identity such that PARSE » IDENT-Length, automatically also implied VFG » IDENT-Length, so that lengthening was also allowed in those cases in which this helped satisfy VFG.

(33)

/bev/ + /t/	IDENT-***F***	VFG	PARSE	IDENT-Length
☞ [be:ft]			*	*
[be:f]	*!			*
[beft]		*!		

A technical problem still remains, viz. how the [f] (devoiced as a result of anb obligatory process of Final Devoicing which Dutch shares with German and other languages) can still trigger VFG if we are only dealing with surface constraints. On first sight, this looks like a classical opacity (counterbleeding) relation between process, just like the interaction between schwa insertion and r deletion in nonrhotic variants of English.

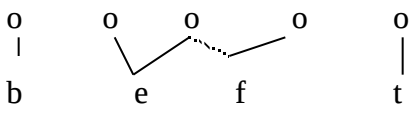
The solution should be found, just like in the case of English, in a closer look at the representations. A first (functional) approximation might be that the vowel lengthening serves to express the underlying vocing contrast on the fricative which itself disappears.

More formally we might look for a solution along the following lines. In Van Oostendorp (2000b) it is argued on independent grounds that the relevant distinction of fricatives is one of length rather than a feature tense in most, if not all, variants of West Germanic: superficially 'voiced' fricatives are phonologically short, and superficially 'voiceless' fricatives are phonologically long:

(34)

It thus looks as if the vowel in this case starts filling the position of the consonant after 'devoicing' (i.e. degemination):

(35)



Rather than filling the extra position (responsible for devoicing) by the fricative alone — thereby lifting the contrast with voiceless fricatives completely, the vowel is filled.

Concludingly, Betuwe Dutch Vowel Lengthening and Shortening behave as a 'conspiracy': together they lift the phonemic distinction between long and short vowels and turn it into an allophonic one. The net result is that vowel length is completely predictable from the phonological environment. It is hard to see how such a language change can be insightfully described in terms of rules. They can be more succesfully described in terms of constraints on representations than in terms of rules.

5. Discussion: A (possibly problematic) case

It is of course possible that the cases discussed until now are atypical. Many phenomena have been studied in the literature under the general rubric of hypercorrection and rule inversion. One of the more problematic phenomena is named '*Rückumlaut'* by Vennemann (1986). At a certain stage in the history of German, there was a productive process of umlauting: fronting a vowel, for instance turning /a/ to [e] before an [i]. This applied in some morphological circumstances (viz. in those in which an [i] was indeed added to the stem), but not in others.

In a following stage in the language history, the [i] disappeared in at least some cases. This left the language systems with alternation such as those in (36a) with a present tense form with an 'umlauted' [e] and other tenses with non-umlauted

[a]. At this point at least some of the language users started restructuring their language system in such a way that they took *brennen* to be the basic form and started to derive *brannte* etc. from this by a kind of 'backward umlaut'. That this is true can be observed from the fact that

(36) *Rückumlaut*
  a. brennen - brannte- gebrannt 'burn: present, past tense, perf. part.'
  b. lernen - larnte 'learn: present -past tense'

The change of [e]-[a] does not follow from any phonological redistribution at all: it is not clear how we could say that 'rückumlauting' could have ever led to a new, allophonic redistribution of the segments [e] and [a]. Rather, the rule of rückumlauting here seems to have been really arbitrary. This seems a case of morphological restructuring, in need of separate treatment.

In my view, the questions concerning this process cannot really be answered if we do not really have a clear idea about the way in which these conspiracy systems start out? In the discussion above we have treated the systems with hypercorrections as if they are uniform synchronic systems. This is of course an oversimplification. The issue of hypercorrection proper, the fact that people sometimes really try to 'reverse engineer' a linguistic process, should also be adressed. I have tried to develop a view on this in Van Oostendorp (1999). The idea there is based on the assumption that speakers have more than one language system as part of there mental make-up; a language system in this conception is a grammar (i.e. an OT ranking of universal constraints) plus a lexicon.

If the systems are very similar — if they are seen as 'variants' of the same language — and if furthermore the language user does not have full command of both systems, he might apply some principles of economy, for instance in the form of correspondence rules between the two language systems.

Something like this might be going on in some of the non-rhotic dialects of English, right now, since rhotic 'standard American counts as a prestige variant at least in some circles. This causes [r] to appear in at least some circumstances in which it is no longer 'linking': "Bostonians who say *Chinar and Japan* are employing an intrusive /r/ which is part of their native accent; while if they say *Japan and Chinar*, they are indulging in hypercorrection." (Trudgill 1986:74).

If a speaker observes, e.g. that many forms ending in [a] in one variant, end in [ar] , he may formulate some correspondence rule on this. The linguistic faculty of the person in question would then look roughly as follows (the relevant wellformedness constraint could now be FINALC, operating on its own in those variants where people say *Japan and Chinar*):

(37)

Non-Rhotic Variant		**Rhotic ('prestigious') Variant**
ca	a ⇔ ar	*car*
spa		*spar*
China		*??*
IDENT »WFC	reranking	WFC » IDENT

We thus have now introduced a kind of 'rule', albeit one which is bidirectional (a ⇔ ar, cf. McCarthy 1991) and which is not internal to a language system, but rather serves to map one language system onto another for reasons of representational economy. At the point where the language learner would come into play, the two 'systems' as depicted in (37) would get integrated. Because at this point, the constraint ranking of the two grammars could also merged; the principles described in this article would start to apply and we would get the kind of 'sandwiching' of constraints. It is not yet clear, of course, whether the non-rhotic dialects would start behaving in this way, since other than language internal factors (such as the ever-changing pattern of social prestige of certain variants) are relevant for this as well.

References

Chomsky, Noam, and Morris Halle. 1968. *The Sound Pattern of English.* Cambridge, Mass.: Cambridge University Press.

Giegerich, Heinz. 1992. *English Phonology. An Introduction*. Cambridge: Cambridge University Press.

—. 1999. *Lexical Strata in English. Morphological Causes, Phonological Effects.* Cambridge: Cambridge University Press.

Goeman, Ton. 2000. Subjectief ervaren afstand tussen dialecten en objectief taalgedrag. *Taal en Tongval* 52.1.

Goeman, Ton, and Piet van Reenen. 1986. Hypercorrecte klinkerverlenging. Een voorbeeld van dialectonderzoek. *Onze Taal* 55.1:8-10.

Halle, Morris, and William Idsardi. 1997. *r,* Hypercorrection, and the Elsewhere Condition. In Iggy Roca (ed.) *Derivations and Constraints in Phonology*. Oxford: Clarendon, 331-348.

Harris, John. 1994. *English Sound Structure*. London: Blackwell.

Jones, Daniel. 1956. *An Outline of English Phonetics.* Cambridge: Heffer.

Kahn, Daniel. 1976. *Syllable-Based Generalizations in English Phonology*. Bloomington: Indiana University Linguistics Club.

Kurath, Hans and Raven I. McDavid. 1961. *The Pronunciation of English in the Atlantic States.* Ann Arbor: University of Michigan Press.

Labov, W. 1966. *The Social Stratification of English in New York City*. Washington DC: Center for Applied Linguistics.

McCarthy, John. 1991. Synchronic Rule Inversion. *Proceedings of the Annual Meeting of the Berkeley Linguistics Society* 17: 192-207.

—. 1993. A Case of Surface Constraint Violation. *Canadian Journal of Linguistics* 38:169-195.

—. 1999. A Note on Boston r and the Elsewhere Condition. Manuscript, University of Massachusetts at Amherst. Available at http://www-unix.oit.umass.edu/~jjmccart/appendix.pdf.

McMahon, April M.S. 1996. On the Use of the Past to Explain the Present: The History of /r/ in English and Scots. In: Derek A. Britton (ed.) *English Historical Linguistics 1994. Papers from the 8th International Conference on English Historical Linguistics.* Amsterdam: John Benjamins.

—. 2000. *Lexical Phonology and the History of English.* Cambridge: Ca,mbrdige University Press.

Oostendorp, Marc van. 1999. Hypercorrectie als hyperoptimaliteit. *Proceedings van de Derde Sociolinguïstische Conferentie.* Delft: Eburon.

—. 2000a. *Phonological Projection*. Berlin: Mouton De Gruyter

—. 2000b. Ambisyllabicity and Fricative Voicing in West Germanic dialects. To appear in: Caroline Féry and Ruben van de Vijver (eds.) *The Syllable in Optimality Theory*. Cambridge: CUP.

—. 2000c. The Phonology of Postvocalic R in Brabant Dutch and Limburg Dutch. Manuscript, Meertens Institute. To appear in: R. van Hout and H. van der Velde (eds.) *Proceedings of the 'r-atics Symposium*.

Ortmann, Albert. 1998. Consonant Epenthesis: Its Distribution and Phonological

Specification. In: Wolfgang Kehrein and Richard Wiese (eds.) *Phonology and Morphology of the Germanic Languages.* Tübingen: Max Niemeyer, 51-76.
Swets, Francine. In preparation. The Word in Tilburg Dutch.
Trudgill, Peter. 1986. *Dialects in Contact*. Oxford: Basil Blackwell.
Van de Velde, Hans. 1996. *Variatie en verandeirn in het gesproken Standaard-Nederlands.* PhD Dissertation, Nijmegen University.
Venneman, Theo. 1972. Rule Inversion. *Lingua* 29: 209-242.
—. 1974. Restructuring. *Lingua* 33: 137-156.
—. 1978. Rule Inversion and Lexical Storage. The Case of Sanskrit Visarga. In: Jacek Fisiak (ed.) *Recent Developments in Historical Phonology*. 's-Gravenhage: Mouton, 391-408.
—. 1986. Rucksümläut. In Dieter Kastowsky and Aleksander Swadesh (eds.) *Linguistics across Historical and Geographical Boundaries. In Honour of Jacek Fisiak on the Occasion of his Fiftieth Birthday*. Berlin: Mouton de Gruyter, 701-723.
Wells, John. 1982. *Accents of English.* Cambridge: Cambridge University Press. 3 Vols.