Overgeneration and falsifiability in phonological theory*

Sylvia Blaho
Hungarian Academy of Sciences
blaho.sylvia@nytud.mta.hu

Curt Rice
University of Tromsø
curt.rice@uit.no

Abstract

A prevalent trend in current phonological practice, usually justified by invoking a simplistic version of Popperian *falsifiability*, is *data-fitting*: fine-tuning phonological models, often with the help of *ad hoc* restrictions, in order to exclude unattested languages from the set of possible grammars generated by the model.

We argue that this practice should be abandoned, for three main reasons. First, it is based on a dubious interpretation of Popperian falsifiability. Second, *ad hoc* restrictions do not further our understanding of language, but they decrease the coherence of a model. Third, the practice is based on data from a small subset of existing languages. We examine each of these arguments in detail.

1 Introduction

A trend often observable in current phonological analyses is what Hale & Reiss (2008) term *data-fitting*: adding *ad hoc* restrictions to models to reduce their predictive power, in order to exclude unattested languages from the set of possible grammars generated by the model. Examples of this practice are proposals of universally fixed rankings (cf. McCarthy 2002), restrictions on constraint conjuction (cf. Baković 2000) in Optimality Theory, or universal feature co-occurrence restrictions (cf. Chomsky & Halle 1968).

This practice is usually supported by invoking a simplistic version of Popper's (1959/2005, ff.) notion of *falsifiability*:

> "Every 'good' scientific theory is a prohibition: it forbids certain things to happen. The more a theory forbids, the better it is."
>
> (Popper 1963)

In this paper, we attempt to show that this kind of argumentation is not in line with the Popperian scientific method. We argue in favour of allowing phonological models to overgenerate under certain circumstances, based on theoretical and empirical arguments. In section 2, we schematically illustrate the most wide-spread method of data-fitting, (partially) fixed rankings in Optimality Theory, and discuss the sense in which we term these

*A version of this paper appeared in *La phonologie de français: normes, périphéries, modélisation*, ed. by Jacques Durand, Gjert Kristoffersen & Bernard Laks, 2014, Presses universitaires de Paris Ouest. We would like to thank Helene Andreassen, Ricardo Bermúdez-Otero, Gene Buckley, Patrick Honeybone, Pavel Iosad, Kristóf Molnár, Dave Odden, the participants of the 20th Manchester Phonology Meeting and three anonymous reviewers for SinFonIJA5 for helpful comments and discussion. Blaho's research was supported by Marie Curie grant nr. MB08-B 82438 and the Research Council of Norway's YGGDRASIL grant nr. 211286.

rankings arbitrary. In section 3, we examine Popper's and Lakatos's requirements of falsifiability, and show that neither of them favour theories with added restrictions. We also argue that arbitrary restrictions serving only to exclude certain language types decrease the coherence of a model, and are dispreferred in both Popper's and Lakatos's system. In section 4, we discuss empirical arguments against data-fitting: drawing on the distinction between attested and possible languages (Hale et al. 2007), we argue that excluding unattested languages from the set of possible grammar predicted by a model is based on an assumption rather than an empirical fact. We also present a number of case studies of recently filled gaps in typologies, that is, language types that have been unattested until recently. Finally, we return to Lakatos to expand on the difficulties of distinguishing 'facts' from 'theory'. We conclude in section 5.

2 'Accounting for gaps'

2.1 Fixed constraint rankings

To illustrate the notion of 'constraining the predictive power of a model' by imposing (partially) fixed rankings in Optimality Theory (OT, Prince & Smolensky 1993), consider the two grammars below.[1]

(1) *Grammar A*
CON = {*VELAR, ID[PLACE], *CORONAL}

(2) *Grammar B*
CON = {*VELAR, ID[PLACE], *CORONAL}
FIX = {*VELAR ≫ *CORONAL}

These two grammars are identical, except for the universally fixed ranking of the constraints *VELAR and *CORONAL in grammar B. (3) shows the violation profiles for the minimal inventory {[t], [k]}.[2]

(3)

t	*VELAR	ID.PLACE	*CORONAL
t			*
k	*	*	

k	*VELAR	ID.PLACE	*CORONAL
t		*	*
k	*		

[1] We use fixed rankings in OT as an example here, but the argument holds true of any grammar with an arbitrary set of restrictions.

[2] Deletion, 'placeless' output segments and other repair strategies are not shown for reasons of simplicity. They do not affect the argument.

Grammar A allows the following language types:

(4)
ID.PLACE ≫ *VELAR,*CORONAL:	/t/→[t], /k/→[k]
*VELAR ≫ ID.PLACE,*CORONAL:	/t/→[t], /k/→[t]
*CORONAL ≫ ID.PLACE,*VELAR:	/t/→[k], /k/→[k]

Grammar B, on the other hand, prohibits all configurations where *CORONAL is not dominated by *VELAR, predicting only two language types.

(5)
ID.PLACE ≫ *VELAR ≫ *CORONAL:	/t/→[t], /k/→[k]
*VELAR ≫ ID.PLACE,*CORONAL:	/t/→[t], /k/→[t]

Proponents of data-fitting argue that Grammar B is to be preferred to Grammar A, because it predicts that languages with velar and no coronal consonants do not exist. Thus, they claim, Grammar B is more falsifiable than Grammar A.

2.2 '*Ad hoc*' vs. 'grounded' restrictions

An anonymous reviewer[3] questioned the fact that these added restrictions are arbitrary, pointing out that many proponents of universally fixed rankings (especially Steriade 2001, ff. and de Lacy 2004, ff.) are taking great care to justify these rankings by invoking extra-phonological — usually phonetic — explanations.

A strong argument against the phonetically grounded view is that it is redundant to encode functional biases in phonology, given that they can arise through diachronic change. Blevins (2004) shows that many phenomena previously thought of as phonological are emergent from the way the human perceptual and articulatory systems work. The argument is not that articulatory and perceptual factors do not play a role in shaping the phonologies of individual languages, but that their role is of a diachronic rather than of a synchronic nature. Given that there already is an extra-phonological explanation for markedness tendencies, it would be superfluous to duplicate this 'knowledge' and build it into our model of phonology.

Recent work on learnability provides strong support for this claim. Boersma et al. 2003, ff. have shown that markedness in phonology is epiphenomenal, since phonetically motivated fixed rankings can be distilled from the data during the learning process. Moreover, the learning algorithm is also capable of inductively acquiring categories based on the input data, which means that phonological features need not be innate, either.

Another argument against innate phonetically grounded constraints comes from non-spoken language. If phonological knowledge is universal, it must apply to all phonologies regardless of modality. This does not apply to phonetically based models, since acoustic perception can hardly play a role in, say, sign language. Moreover, if phonetically grounded constraints are universal, then such constraints for *all modalities* must be assumed to be innate. UG would then have to contain at least two sets of constraints: one for spoken language and one for sign language (and even more sets if the phonology of other modalities, such as tactile language, turn out to have different phonetics from spoken and signed language).

[3]We are grateful to a reviewer of SinFonIJA5 for prompting us to clarify this issue.

It is hardly necessary to point out how difficult it would be to acquire language given this scenario.

If, however, we define 'phonology' as a symbolic computational system — the view taken by proponents of substance-free phonology (including Hale & Reiss 2000a,b, ff., Hume 2003, ff., Blevins 2004, ff., Dresher et al. 1994, ff., Rice & Avery 1993, .ff., Mielke 2005, ff., Morén 2007b,a. ff., Odden 2006, ff., Blaho 2008, ff.) —, and we require that a theory of phonology should shed light on how this symbolic system works, then extra-phonological justification of these fixed rankings is meaningless: computationally, FIX = {*VELAR ≫ *CORONAL} makes just as much sense as FIX = {*CORONAL ≫ *VELAR}.

3 Falsifiability

This section evaluates the relative merits of grammars A and B within two frameworks of the philosophy of science: those of Popper and Lakatos. Although we find the latter much more explicit than the former, we aim to show that neither of them prefers grammars of the type B over those of the type A; in fact, we find that grammars of the type B fare worse on both sets of criteria.

3.1 Popper

As we illustrated on page 1 above, a superficial reading of Popper's work might easily lead one to believe that data-fitting is the most rigorous scientific method according to Popperian philosophy.

Closer examination of Popper's work, however, reveals that his notion of falsifiability refers to more fundamental properties of a model than is generally understood in phonology. This is illustrated by the following quote (discussing psycho-analysis):

> "[. . .] what kind of clinical responses would refute to the satisfaction of the analyst not merely a particular analytic diagnosis but psycho-analysis itself?" (Popper 1963)

Accordingly, we distinguish two notions of falsifiability: *surface falsifiability* and *fundamental falsifiability*. A frequently cited example for the two kinds of falsifiability relates to the theory of evolution by natural selection. Finding "fossil rabbits in the Precambrian" (attributed to J. B. S Haldane) would falsify the theory on the *surface* level, but it would not affect the validity of the basic mechanisms of natural selection. Falsifying the theory on a *fundamental* level would require proving that mutations do not occur or that they occur but cannot be inherited.

Analogously, Optimality Theory is fundamentally falsified by proving that candidate selection cannot be computed by humans, but not if a particular constraint ranking is shown to be at odds with empirical data. Showing that velars can exist in languages that do not have coronals (such as Hawaiian) falsifies Grammar B on the surface, that is, it falsifies that *particular instance* of a grammar with universal fixed rankings. It does not, however, fundamentally falsify the idea of OT grammars with universal fixed rankings. Thus, grammars of the *type* A are equally fundamentally falsifiable as grammars of the *type* B.

Accordingly, imposing arbitrary restrictions on constraint ranking or constraint conjunction does not make a model more fundamentally falsifiable. It does, however, lead to a less coherent and a less parsimonious model by introducing an extra mechanism (the list of arbitrary restrictions, which we termed FIX above). Moreover, since these restrictions are typically simple restatements of the observed range of data, they contradict the Popperian scientific method:

> "The introduction of an auxiliary hypothesis should always be regarded as an attempt to construct a new system; and this new system should then always be judged on the issue of whether it would, if adopted, constitute a real advance in our knowledge of the world."
>
> (Popper 1959/2005)

An anonymous reviewer[4] suggested that even though our claim that grammars of the type B are not more fundamentally falsifiable than grammars of the type A might be true, the fact that they are more *surface falsifiable* should be "worth something". We have tried to show above that, according to Popper, what we termed 'surface falsifiability' is not "worth" anything as far as evaluating the scientific or pseudoscientific nature of a theory is concerned.

To illuminate why the notion of surface falsifiability is useless from the point of view of the philosophy of science and theory construction, let us take the notion to its extreme using a *reductio ad absurdum* scenario. Our starting assumption, then, is the following:

(6) All else being equal, theory $T\prime$ is to be preferred over theory T if $T\prime$ is more surface falsifiable, that is, if it excludes more languages.

Now take a language like Hungarian (Finno-Ugric; see Siptár & Törkenczy 2000 for an overview of its phonology). This language exhibits a fairly complex morphophonology as far as its vowel harmony patterns are concerned: rounding harmony is combined with a front/back harmony. As far as the latter is concerned, front unrounded vowels show a set of unique behaviours: transparency (they are 'invisible' for harmony), anti-harmony (some stems containing only neutral vowels take front suffixes, while other stems of the same shape take back ones) and vacillation (the same stem can take front and back suffixes). A series of recent findings have sketched an even more complex picture. Beňuš & Gafos (2007) studied the phonetic properties of stems containing neutral vowels in harmonic and anti-harmonic stems and claimed that the vowels in anti-harmonic stems are pronounced with a more back articulation than the vowels in harmonic stems. Hayes et al. (2009) have shown that the number of neutral vowels at the left edge of the stem has an influence on the choice of the suffix vacillating stems take; among other factors, labial stem consonants favour back suffixes. Kálmán et al. (2010) found that vacillating stems have a preference choosing harmonic variants of suffixes so that the result is similar to existing monomorphemic words, while Törkenczy et al. (2011) have shown that there are several degrees of vacillation and transparency, and the morphological identity of both the stem and the suffix has to be taken into account.

We believe we can safely assume that no other language has exactly the same vowel system *and* exactly the same, categorical and gradient vowel harmony patterns as Hungarian.[5]

[4] Again, we are grateful to a reviewer of SinFonIJA5 for prompting us to clarify this issue.

[5] As will hopefully become clear by the end of this subsection, this assumption is not crucial to the

Now let us consider another property of Hungarian, voicing assimilation in obstruent clusters. The behaviour of the majority of obstruents can be fairly easily described: regressive assimilation across the board, with no utterance-final or syllable-final devoicing. Certain consonants, namely /j/, /v/ and /h/ deviate from this pattern in non-trivial ways. /j/ is realised as a palatal approximant in most cases. However, when preceded by a consonant and followed by a pause, it surfaces as a fricative, which is voiceless following voiceless obstruents, and voiced after voiced obstruents and sonorants. /v/ undergoes voicing assimilation if followed by an obstruent, but it does not trigger assimilation. /h/ is realised as a glottal fricative, except when it is followed by a consonant, a pause or a 'strong' morpheme boundary, when it is realised as a velar fricative. [h] triggers devoicing in preceding obstruents, but [x] fails to undergo voicing assimilation: it is voiceless even if it is followed by a voiced obstruent, creating the only type of obstruent cluster that does not agree in voicing in Hungarian.

Again, we believe we can assume that no other language has the exact same consonant inventory *and* the exact same voicing assimilation patterns as Hungarian.

Now consider two competing grammars: grammar C accounts for all and only the attested vowel harmony patterns and all and only the attested voicing assimilation patterns in the phonologies of all attested languages. Grammar D, on the other hand, accounts for all the facts that grammar C accounts for, *plus*, it also accounts for the fact that if a language has the kind of vowel harmony Hungarian has, it also has the kind of voicing assimilation Hungarian has. In other words, grammar D has some kind of prohibition against languages that have Hungarian-style voicing assimilation but not Hungarian-style vowel harmony and *vice versa*.

Now, if we assume that surface falsifiability should be used as a criterion for evaluating the merits of competing frameworks, then grammar D must be preferred over grammar C.

Hungarian also displays a fascinating (and, presumably, unique) pattern of ineffability (see Rebrus & Törkenczy 2010 for a thorough description). Let us now compare our grammar D to grammar E, one that makes all predictions that grammar D does, *plus* it states that languages that have Hungarian-style voicing assimilation and Hungarian-style vowel harmony also have to have Hungarian-style ineffability patterns (and *vice versa*). Again, grammar E is to be preferred over grammar D on the criterion of surface falsifiability.

It should now be easy to see how this line of grammars could be continued until the ultimately specific grammar that only permits attested languages. The line of grammars would not be infinite, but it would be extremely long: after all, once we arrive at a grammar that only permits the attested phonological systems, we have to go on to include syntax, semantics, the lexicon, and so on. So using surface falsifiability to evaluate our models leads us to an incredibly complex and incredibly specific grammar — but what do we end up gaining? Does 'accounting for' the fact that Hungarian vowel harmony and Hungarian voicing assimilation (not to mention Hungarian phonology and Hungarian syntax) occur in the same language advance our knowledge of the world?

We believe the answer depends on *how* our model predicts these kinds of facts. Rebrus & Törkenczy (2010) offer an analysis of the Hungarian ineffability pattern that makes a connection between ineffability and vowel-zero alternations in different stem/suffix classes, which in turn are connected to consonants cluster phonotactics.

validity of the argument we are making, it merely serves to simplify the explanation.

However, if the set of restrictions differentiating theory T and theory $T\prime$ is added by simple conjunction (such that $T\prime = T$ & FIX), then the answer is no — taking us back to Popper's notion of coherence. As we mentioned above (p. 4), Popper's discussion of Freudian and Marxist theories makes it clear that surface falsifiability is irrelevant for the purposes of theory construction.

For more elaborated and formalised notions of 'falsifiability' and 'furthering our knowledge of the world', we now turn to the work of one of Popper's greatest followers, Imre Lakatos.

3.2 Lakatos

Lakatos (1970), in an attempt to synthetise Popperian falsificationsim and Kuhn's notion of scientific paradigm in a framework he terms *sophisticated falsificationism*, defines falsifiability as a criterion to be evaluated on a series of theories rather than solitary ones.

> "For the sophisticated falsificationist a scientific theory T is *falsified* if and only if another theory $T\prime$ has been proposed withe the following characteristics:
>
> (1) $T\prime$ has excess empirical content over T: that is, it predicts *novel facts*, that is, facts improbable in the light of, or even fobidden, by T;
>
> (2) $T\prime$ explains the previous success of T, that is, all the unrefuted content of T is included (within the limits of observational error) in the content of $T\prime$; and
>
> (3) some of the excess content of $T\prime$ is corroborated."
>
> (Lakatos 1970: 116)

In the light of this, Lakatos argues, it does not make sense to evaluate the falsifiability or 'scientificness' of any theory alone; rather, one has to compare a succession of two or more theories to determine of the newer theory falsifies its predecessor or not. Lakatos further refines the notion of falsifiability in a given string of theories by introducting the concept of *progressive* vs. *degenerating problemshifts*.

> "Let us take a series of theories T_1, T_2, T_3... where each subsequent theory results from adding auxiliary clauses to (or form semantical reinterpretations of) the previous theory in order to accommodate some anomaly, each theory having at least as much content as the unrefuted content of its predecessor. Let us say that such a series of theories is *theoretically progressive* (*or 'constitutes a theoretically progressive problemshift'*) if each new theory has some excess empirical content over its predecessor, that is, if it predicts some novel, hitherto unexpected fact. Let us say that a theoretically progressive series of theories is also *empirically progressive* (*or 'constitutes an empirically progressive problemshift'*) if some of this excess empirical content is also corroborated, that is, if each new theory leads us to the actual discovery of some *new fact*. Finally, let us call a problemshift *progressive* if it is both theoretically and empirically progressive and *degenerating* if it is not. We '*accept*' problemshifts as 'scientific' only if they are at least theoretically progressive, if they are not, we '*reject*' them as 'pseudoscientific'. Progress is measured by the degree to which a problemshift is progressive, by the degree to which the series of theories leads us to the discovery of novel facts. We regard a theory in the series 'falsified' when it is superseded by a theory with higher corroborated content."

(Lakatos 1970: 118)

Given the definition above, one might judge the progression from grammars of the type A to grammars of the type B a progressive problemshift:[6] they make predictions that are unexpected according to grammars of the type B, namely that certain language types allowed by the latter do not exist. Thus, at first glance, it seems that grammars of the type B falsify grammars of the type A.

This, however, turns out to be an illusion if one takes a closer look at what 'predicting novel facts' means according to Lakatos. Similarly to Popper, Lakatos also demands that new hypotheses be more than mere restatements of some data patterns, formulating this in a more explicit way:

> "A theory without excess corroboration has no excess explanatory power; *therefore, according to Popper, it does not represent growth and therefore it is not 'scientific'; therefore, we should say, it has no explanatory power*"
> ...
> "*A given fact is explained scientifically only if a new fact is also explained with it.*"

(Lakatos 1970: 119)

As the quotes above illustrate, grammars of the type B do not constitute a progressive problemshift in Lakatos's framework: they set out to explain why certain types of languages are unattested and they propose an auxiliary hypothesis stating that these types of languages are impossible. The auxiliary hypothesis accounts for the alleged gap in the typology predicted by the model that it set out to explain, but *it makes no other novel predictions* — in other words, it *does not explain any additional facts* — , and thus does not constitute a progressive problem shift compared to Grammar B.

4 'Facts'

Having presented some theoretical arguments against data-fitting, we now turn to the question of how to use empirical data in theory construction.

4.1 Negative evidence

Both overgeneration and undergeneration are challenging to any theory of phonology, but they are of a very different nature. Since the overwhelming majority of existing languages are still undescribed, and most other languages only received impressionistic descriptions, our understanding of what is impossible is tentative at best. Thus, proposing a theoretical tool *solely* for the purpose of excluding a non-existent pattern decreases the coherence of the theory in order to accommodate a mere *assumption*, i. e., the assumption that unattested patterns are in fact universally impossible.

[6] For the moment, let us assume that all the markedness implications motivating grammars of the type B are true, keeping in mind that many of them have been shown to be incorrect: for instance, the implication that languages that have velar consonants also have coronals does not hold true of all known languages, while the statement that sibilant consonants cannot be labial or velar has, to the best of our knowledge, not been contradicted. We will return to these issues in section 4.1. In the present section, however, we would like to show that grammars of the type B do not fare better than those of the type A even on purely theoretical grounds.

Hale et al. (2007) argue convincingly that the set of attested languages is only a subset of possible human languages. Rather, the relationship is as follows:

(7) attested lgs $\subset$ attestable lgs $\subset$ humanly computable lgs $\subset$ statable lgs

Hale et al. (2007) provide the following explanation for this pattern.

> "First, the set of attested languages is a subset of the set of attestable languages (where attestable includes all linguistic systems which could develop diachronically from existing conditions — e.g., all dialects of English or Chinese or any other language in 400 years, or 4000 years, etc.). In addition, the set of attestable languages is a subset (those which can evolve from current conditions) of the set of humanly computable languages. (In our opinion, the human phonological computation system can compute a featural change operation such as /p/ $\rightarrow$ [a] /__ d but it is of vanishingly small probability that such a rule could arise from any plausible chain of diachronic changes.) Finally, the set of humanly computable languages is itself a subset of formally statable systems (which could include what we take to be humanly impossible linguistic processes such as /V/ $\rightarrow$ [V] in prime numbered syllables). The key point here is that the set of diachronically impossible human languages is not equivalent to the set of computationally impossible human languages."

The view that follows from this approach is that our model of phonology should predict the set of humanly computable languages, not the set of attested languages.

In fact, many language types thought to be impossible have recently been shown to exist. We will review a few of these examples below.

One of the best known examples is Lesgian (Blevins 2004). In this language, only voiced obstruents can occur word-finally, which contradicts the markedness implication that voiceless obstruents are preferred over voiced ones in this position. This pattern is considered phonetically 'unnatural', since the cues for obstruent voicing can be perceived poorly in this position Steriade (2001). However, Blevins (2004) describes a scenario for how this pattern could evolve: intervocalic voicing being (diachronically) followed by loss of word-final vowels.

An example recently discussed by Davis et al. (2006) concerns initial consonant clusters. Contrary to the observation that #TR clusters are less marked than #TT clusters both from an acoustic and a perceptual point of view, in Hocank the former are broken up by a schwa, but the latter are retained. While the phenomenon can be given a diachronic explanation based on the perceptual similarity of #TR and #TəR, a model incorporating constraints propagating the ease of articulation or perception can hardly account for this pattern.

Examining the markedness of place of articulation in consonants, K. Rice (2004) shows that, although coronal is generally considered to be the unmarked place for stops, there are languages where the only stops are labial (e.g. Nimburan) and velar (e.g. Fuzhou). Moreover, any two of these three places of articulation can be found in languages to the exclusion of the third place: both labial and velar, but not coronal stops are found in dialects of Vietnamese, coronals and labials, but not velars in Kiowa, and coronals and velars, but not labials in some Chinese dialects.

C. Rice (2011) discusses a similar state of affairs with respect to languages with ternary rhythm. He notes that Hayes (1980) argues against Halle's inventory of metrical feet based on the (incorrect) assumption that ternary feet are impossible:

> "I know of no languages whose stress patterns could simply be described using feet of the [ternary] form."
>
> (Hayes 1980: p. 115)

Since then, however, convincing evidence for ternary feet has been found in a number of languages, some of the clearest cases being Cayuvava, Tripura Bangla and Chugach Alutiiq (see Rice 2011 for an overview of the data and the relevant literature).

After it had been established that ternary rhythm does exist, the typological work continued to determine what types of ternary feet were possible.

Within the OT paradigm, Buckley (2009) shows that a ternary stress pattern excluded by the typology of Kager (2001) is well attested in Kashaya, and examples of other patterns that are harmonically bounded in Kager's system can be found in Indonesian and Spanish.

In a review of McCarthy (2002), (Odden 2003: 164) discusses an interesting 'gap' in the typology that had been 'filled'. He notes that although it had long been assumed that there are no languages with both clicks and pharyngeal approximants, Dahalo has both of these groups of sounds.

4.2 'Theory' *vs.* 'facts'

Finally, we return to Lakatos's seminal paper to discuss a crucial issue that had been ignored throughout this paper: the nature of 'facts'. Some proponents of data-fitting have argued that the lack of certain language types is not the only indication that such language types are impossible, and proposed experimental evidence to support the claim that certain ('unnatural') patterns are in fact unlearnable, or at least they are less readily learned than other ('natural') patterns (see Hayes & White 2012, to appear for an overview of these studies).

However, Lakatos argued convincingly that making a principled distinction can be made between 'theory' and 'facts' is far from trivial.

In the scenario propagate by what Lakatos terms *naive falsificationsim*, testing a theory empirically is straightforward: if a fact of nature is at odds with a prediction of the theory, then the theory is falsified; if no such empirical fact is found, the theory stands.

As Lakatos points out, we do not have direct access to 'pure' facts: all of our observations and experimental results are dependent on the methodology and subject to different interpretations. Moreover, an implicit part of all theories is a *ceteris paribus* clause: our model predicts these results, *if no other factors interfere*. Thus, if some 'fact' is in contradiction with a model, any of these three components can be at fault:

1. the theory: this is the trivial case, the one recognised by naive falsificationism.
2. 'the facts', that is, the experimental setup, the interpretive theory, the properties of the instruments used, etc.
3. the *ceteris paribus* clause

He also points out that it is rather arbitrary which components of a model belong to which group:

> "Whether a proposition is a '*fact*' or a '*theory*' in the context of a test-situation depends on our methodological decision."
>
> (Lakatos 1970: 129)

Therefore, if sophisticated falsificationism, there is no empirical battle between theory and facts, and a theory is not immediately falsified if an 'observation' is in conflict with it. Rather, what we can detect are *inconsistencies* between a theoretical model and the interpretive theory (formerly know as 'the facts'). In Lakatos's words:

> "The problem is then *shifted* from the old problem of replacing a theory refuted by the 'facts' to the new problem of how to resolve inconsistencies between closely related theories. Which of the mutually inconsistent theories should be eliminated? the sophisticated falsificationist can answer that question easily: one had to try to replace the first one, then the other, then possibly both, and opt for that new set-up which provides the biggest increase in corroborated content, which provides the most progressive problemshift."
>
> (Lakatos 1970: 130)

As for the *ceteris paribus* clause, Lakatos has no better advice to offer than to try to control for every conceivable factor that could have an effect on our experimental results. But in a discipline as young as linguistics, where empirical statements that have been considered rock-hard facts for decades have been shown to be completely false, one should be especially careful about interpreting 'empirical facts'.

In fact, Hayes & White, while arguing for the idea that naturalness should somehow be a part of phonological knowledge, are well aware of the pitfalls of interpreting the results of experiments too readily, and conclude that the evidence currently available on this matter is far from conclusive.

5 Summary

We argue that, all else being equal, a theory with *ad hoc* restrictions of the type "Constraint1 must universally dominate Constraint2" is not preferable to a theory without such restrictions: it is not more falsifiable under a careful interpretation of Popperian falsifiability, it is less coherent than a theory without arbitrary restrictions, and it is constructed based on an *assumption* that unattested patterns are in fact impossible. Thus, theoretical innovations should always be based on existing patterns, not motivated by trying to exclude unattested phenomena.

An important part of this enterprise will include the pursuit of large-scale, careful, theoretically informed projects documenting languages and language variation in ways that will give a sounder empirical basis for the theoretical work of modelling the range of variation to be allowed by universal grammar.

References

Baković, Eric (2000). 'Harmony, dominance and control'. Ph.D. thesis, Rutgers University. `http://roa.rutgers.edu/files/360-1199/roa-360-bakovic-1.pdf`.

Beňuš, Štefan & Adamantios I. Gafos (2007). 'Articulatory characteristics of Hungarian 'transparent' vowels'. *Journal of Phonetics* **35**, pp. 271–300.

Blaho, Sylvia (2008). 'The syntax of phonology. A radically substance-free approach'. Ph.D. thesis, University of Tromsø. `http://ling.auf.net/lingBuzz/000672`.

Blevins, Juliette (2004). *Evolutionary phonology. The emergence of sound patterns*. Cambridge University Press, Cambridge.

Boersma, Paul, Paola Escudero & Rachel Hayes (2003). 'Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories'. In: 'Proceedings of the 15$^{\text{th}}$ International Congress of Phonetic Sciences', pp. 1013–1016, pp. 1013–1016.

Buckley, Eugene (2009). 'Locality in metrical typology'. *Phonology* **26**, pp. 389–435.

Chomsky, Noam & Morris Halle (1968). *The Sound Pattern of English*. Harper and Row, New York.

Davis, Stuart, Karen Baertsch & William Anderson (2006). 'Explanation in phonetics and phonology: Understanding Dorsey's Law in Hocank (Winnebago)'. Paper presented at 14$^{\text{th}}$ Machester Phonology Meeting.

Dresher, B. Elan, Glyne Piggott & Keren Rice (1994). 'Contrast in Phonology: Overview'. *Toronto Working Papers in Linguistics* **14**, pp. iii–xvii.

Hale, Mark & Charles Reiss (2000a). 'Phonology as cognition'. In: Noel Burton-Roberts, Philip Carr & Gerard Docherty (eds.), 'Phonological Knowledge. Conceptual and Empirical Issues', Oxford University Press, Oxford, pp. 161–184.

Hale, Mark & Charles Reiss (2000b). '"Substance abuse' and "dysfunctionalism": current trends in phonology'. *Linguistic Inquiry* **31**, pp. 157–169.

Hale, Mark & Charles Reiss (2008). *The Phonological Enterprise*. Oxford University Press, Oxford.

Hale, Mark, Charles Reiss & Madelyn J. Kissock (2007). 'Microvariation, variation, and the features of Universal Grammar'. *Lingua* **117**, pp. 645–665.

Hayes, Bruce (1980). 'A metrical theory of stress rules'. Ph.D. thesis, MIT.

Hayes, Bruce & James White (2012, to appear). 'Phonological naturalness and phonotactic learning'. *Linguistic Inquiry* .

Hayes, Bruce, Kie Zuraw, Péter Siptár & Zsuzsa Londe (2009). 'Natural and unnatural constraints in Hungarian vowel harmony'. *Language* **85**, pp. 822–863.

Hume, Elizabeth (2003). 'Language specific markedness: The case of place of articulation'. *Studies in Phonetics, Phonology and Morphology* **9**, pp. 295–310.

Kager, René (2001). 'Rhythmic directionality by positional licensing'. Handout of paper presented at the 5$^{\text{th}}$ Holland Institute of Linguistics Phonology Conference, Potsdam. Available as ROA-514 from the Rutgers Optimality Archive.

Kálmán, László, Péter Rebrus & Miklós Törkenczy (2010). 'Lehet-e az analógiás nyelvtan szinkrón?' Paper presented at A magyar nyelvészeti kutatások újabb eredményei II. conference.

de Lacy, Paul (2004). 'Markedness conflation in optimality theory'. *Phonology* , pp. 1–55.

Lakatos, Imre (1970). 'Falsification and the methodology of scientific research programmes'. In: Imre Lakatos & Alan Musgrave (eds.), 'Criticism and the growth of knowledge. Proceedings of the International Colloquium in the Philosophy of Science, London, 1964', Cambridge University Press, Cambridge, vol. 4, pp. 91–196.

McCarthy, John J. (2002). *A Thematic Guide to Optimality Theory*. Cambridge University Press, Cambridge.

Mielke, Jeff (2005). 'Modeling distinctive feature emergence'. In: 'Proceedings of the West Coast Conference on Formal Linguistics 24', pp. 281–289, pp. 281–289.

Morén, Bruce (2007a). 'Minimalist/substance-free feature theory: Case studies and implications'. Course held at the 14th EGG Summer School, Brno, the Czech Republic.

Morén, Bruce (2007b). 'Minimalist/substance-free feature theory: Why and how'. Course held at the 14th EGG Summer School, Brno, the Czech Republic.

Odden, David (2003). 'John J. McCarthy (2002). a thematic guide to Optimality Theory. (Research Surveys in Linguistics.) Cambridge: Cambridge University Press. pp. xiii+317'. *Phonology* **20**, pp. 163–167.

Odden, David (2006). 'Phonology ex nihilo'. Paper presented at the Tromsø Phonology Project Group Meeting.

Popper, Karl (1959/2005). *The logic of scientific discovery*. Routledge, London and New York.

Popper, Karl (1963). *Conjectures and Refutations: The Growth of Scientific Knowledge*. Routledge, London and New York.

Prince, Alan & Paul Smolensky (1993). 'Optimality Theory: Constraint interaction in generative grammar'. Ms., Rutgers University & University of Colorado at Boulder.

Rebrus, Péter & Miklós Törkenczy (2010). 'Assamese nasals blocking vowel harmony'. In: Curt Rice & Sylvia Blaho (eds.), 'Modeling ungrammaticality in Optimality Theory', Equinox, London, pp. 195–235.

Rice, Curt (2011). 'Ternary rhythm'. In: Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume & Keren Rice (eds.), 'The Blackwell Companion to Phonology', Blackwell, Oxford, pp. 1228–1244.

Rice, Keren (2004). 'Neutralization and epenthesis: Is there markedness in the absence of contrast?' Paper presented at GLOW 27, Thessaloniki, Greece.

Rice, Keren & Peter Avery (1993). 'Segmental complexity and the structure of inventories'. *Toronto Working Papers in Linguistics* **12**, pp. 131–154.

Siptár, Péter & Miklós Törkenczy (2000). *The phonology of Hungarian*. Oxford University Press, Oxford.

Steriade, Donca (2001). 'The phonology of perceptibility effects: the P-map and its consequences for constraint organization'. Ms., , MIT.

Törkenczy, Miklós, László Kálmán, Péter Rebrus & Péter Szigetvári (2011). 'Harmony that cannot be represented'. Paper presented at SinFonIJA 4, Budapest.