A stochastic OT account of paralinguistic tasks such as grammaticality and prototypicality judgments

Paul Boersma, March 18, 2004

It has been observed that grammaticality judgments do not necessarily reflect relative corpus frequencies: it is possible that structure A is judged as more grammatical than structure B, whereas at the same time structure B occurs more often in actual language data than structure A. In recent work (Boersma & Hayes 2001), we have used Stochastic Optimality Theory to model grammaticality judgments in exactly the same way as corpus frequencies, namely as the result of noisy evaluation of constraints ranked along a continuous scale. At first sight, therefore, this model seems not to be able to handle the observed facts: linguistic forms that have zero corpus frequency due to *harmonic bounding* often turn out not to be totally ungrammatical (Keller & Asudeh 2002), and 'ideal' forms found in experiments on *prototypicality judgments* often turn out to be peripheral within the corpus distribution of their grammatical category (Johnson, Flemming & Wright 1993). In this paper, I argue that the paradox is solved by assuming a listener-oriented grammar model (Boersma 1998), in phonology as well as in syntax. In that grammar model, the natural way to derive (relative) corpus frequency is to measure the *production process*, whereas grammaticality judgments naturally derive from a simpler process, namely the *inverted interpretation process.* Section 1 explains the Stochastic OT model for grammaticality judgments. Sections 2, 3, and 4 handle three respects in which grammaticality judgments deviate from corpus frequencies.

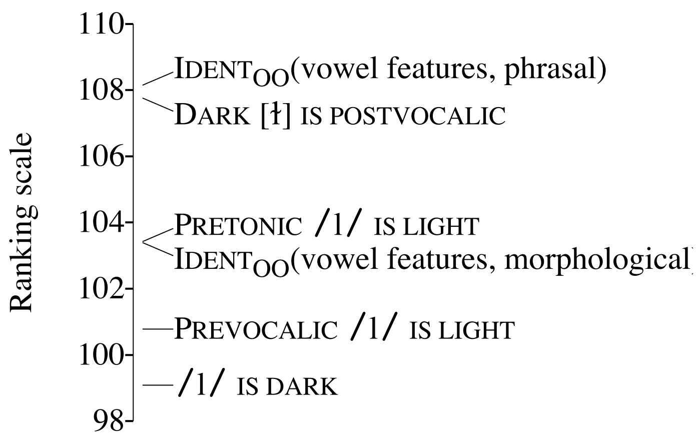
1. Simple frequency–grammaticality relations in Stochastic OT

The following example is from Hayes (2000) and Boersma & Hayes (2001). Hayes elicited grammaticality judgments (on a seven-point scale, from xx subjects) for the two major allophones of American English /l/, namely the 'light' (non-velarized) allophone [l] and the 'dark' (velarized) allophone [ɫ]. The order in which the average subject preferred the light /l/ can be summarized by the sequence of prosodic words *light*, *Louanne*, *free-ly*, *Greeley*, *feel-y*, *feel it*, and *feel*: the dark allophone was judged ungrammatical in *light* and *Louanne* and was dispreferred in *free-ly*, while the light allophone was judged ungrammatical in *feel*, strongly dispreferred in *feel it*, slightly dispreferred in *feel-y*, and slightly preferred in *Greeley.* Hayes proposed that the choice between the allophones in production is governed by six factors, all of which can be translated into an OT constraint. The constraint "PRETONIC /l/ IS LIGHT" prefers [l] to [ɫ] in *light*, while "DARK [ɫ] IS POSTVOCALIC" prefers [l] to [ɫ] in *light* and *Louanne*, and "PREVOCALIC /l/ IS LIGHT" prefers [l] to [ɫ] in *light*, *Louanne*, *free-ly*, *Greeley*, *feel-y*, and *feel it.* To counteract these three, there is the constraint "/l/ IS DARK", which prefers [ɫ] to [l] in all seven words; this ensures that at least *feel* will receive a dark [ɫ]. Finally, Hayes proposed two output-output faithfulness constraints, which favour identity of vowel allophones across morphologically related forms. For this we have to understand that /i/ is realized as [iː] before [l] and as [iə] before [ɫ]. Thus, "IDENT$_{OO}$ (vowel features /

morphological)" prefers a diphthongized dark [iə], and therefore a dark [ɫ], in *feel-y*, which is morphologically derived from *feel* [fiəɫ]. The same constraint prefers a long [iː], and therefore a light [l], in *free-ly*, which is derived from *free* [fɹiː]. Likewise, "IDENT$_{OO}$ (vowel features / phrasal)" prefers [iə] and therefore [ɫ] in *feel it.*

Boersma & Hayes (2001) modelled the English /l/ data with Stochastic OT (Boersma 1997), a form of Optimality Theory in which constraints have *ranking values* along a continuous scale rather than ordinal positions along a discrete scale, and in which *noise* is temporarily added to the ranking of each constraint at the evaluation of every tableau. This evaluation noise can lead to random variation: if constraint A has a ranking value not far above that of constraint B, it will outrank B most of the time but not always. Stochastic OT comes with a learning algorithm, the *Gradual Learning Algorithm*, which Boersma & Hayes used to compute a ranking for the six constraints, assuming a certain monotonic relationship between the subjects' grammaticality judgments and the relative corpus frequencies of the forms. The learning algorithm came up with the ranking in (1). The resulting distances along the ranking scale are appropriate for an evaluation with a root-mean-square strength ('standard deviation') of 2.0, which was also used during learning.

(1) *Ranking for determining the allophones of American English* /l/



(2) *Very strong preference for dark* [ɫ] *word-finally before unstressed vowel*

|fil#ɪt| (~ |fil| [fiəɫ])	IDENT$_{OO}$ (phrasal)	DARK [ɫ] IS POSTVOCALIC	PRETONIC /l/ IS LIGHT	IDENT$_{OO}$ (morphol)	PREVOCALIC /l/ IS LIGHT	/l/ IS DARK
[fiːlɪt]	*!					*
☞ [fiəɫɪt]					*	

Tableau (2) shows how this ranking handles the phrase *feel it.* Since "IDENT$_{OO}$ (vowel featuresphrasal) is ranked far above "PREVOCALIC /l/ IS LIGHT" (namely, 3.7 times the noise strength, as we see in (1)), the form with dark [ɫ] usually wins. Given the rankings in (1) and an evaluation noise of 2.0, /l/ ends up being dark in 99.5 percent of the cases. Boersma & Hayes propose that the 0.5 percent occurrence of light [l] corresponds to a '??' grammaticality judgment for [fiːlɪt].

(3) *Weak preference for light* [l] *word-internally before unstressed vowel*

|gɹili|	IDENT$_{OO}$ (phrasal)	DARK [ɫ] IS POSTVOCALIC	PRETONIC /l/ IS LIGHT	IDENT$_{OO}$ (morphol)	PREVOCALIC /l/ IS LIGHT	/l/ IS DARK
☞ [gɹiːliː]						*
[gɹiəɫiː]					*!	

The case is different for *Greeley*, as (3) shows. Since the word is monomorphemic, there is a preference for a light [l]. The preference is only slight, since “PREVOCALIC /l/ IS LIGHT” is ranked less than 0.9 noise strengths above “/l/ IS DARK”. The result of this ranking is that [gɹiːliː] will occur in 73 percent of the cases and [gɹiəɫiː] in 27 percent of the cases. Boersma & Hayes propose that this 27 percent is large enough to cause a ‘√’ judgment for [gɹiːɫiː].

(4) *Medium preference for a dark* [ɫ] *stem-finally before unstressed vowel*

|fil+i| (~ |fil| [fiəɫ])	IDENT$_{OO}$ (phrasal)	DARK [ɫ] IS POSTVOCALIC	PRETONIC /l/ IS LIGHT	IDENT$_{OO}$ (morphol)	PREVOCALIC /l/ IS LIGHT	/l/ IS DARK
[fiːliː]				*!		*
☞ [fiəɫiː]					*	

Tableau (4) shows an intermediate case. For *feel-y*, the light candidate can only win if at evaluation time “PREVOCALIC /l/ IS LIGHT” is ranked both above “/l/ IS DARK” and above “IDENT$_{OO}$ (vowel features, morphological)”. This occurs in 17 percent of the cases, which is proposed to be (barely) compatible with a ‘?’ verdict.

(5) *Strong preference for light* [l] *in post-stem position before unstressed vowel*

|fri+li| (~ |fri| [friː])	IDENT$_{OO}$ (phrasal)	DARK [ɫ] IS POSTVOCALIC	PRETONIC /l/ IS LIGHT	IDENT$_{OO}$ (morphol)	PREVOCALIC /l/ IS LIGHT	/l/ IS DARK
☞ [fɹiːliː]						*
[fɹiəɫiː]				*!	*	

Tableau (5) shows the reverse case, in which the underlying |l| belongs to the suffix. It is now the candidate with dark [ɫ] that violates the output-output constraint, because of the existence of the undiphthongized form *free* [friː]. The frequency of [fɹiəɫiː] is now mainly determined by the distance between “IDENT$_{OO}$ (vowel features / morphological)” and “/l/ IS DARK”, which is 2.2 times the noise strength, as can be seen in (1). The relative frequency of [fɹiəɫiː] as predicted by the ranking is 4 percent, which is proposed to correspond to a clear “?”.

(6) *Ungrammaticality of dark* [ɫ] *utterance-initially before stressed vowel*

|lait|	IDENT$_{OO}$ (phrasal)	DARK [ɫ] IS POSTVOCALIC	PRETONIC /l/ IS LIGHT	IDENT$_{OO}$ (morphol)	PREVOCALIC /l/ IS LIGHT	/l/ IS DARK
☞ [lait]						*
[ɫait]		*!	*		*	

Tableau (6) shows an extreme case. Since the distance between "DARK [ɫ] IS POSTVOCALIC" and "/l/ IS DARK" is 4.3 times the noise strength, and "/l/ IS DARK" also has to overcome the relatively high-ranked "PRETONIC /l/ IS LIGHT", the relative frequency of [ɫait] is only 0.06 percent, which Boersma & Hayes propose is low enough to receive a "*" verdict. A perhaps even stronger ungrammaticality judgment has to befall *feel* [fiːl], whose predicted frequency is zero since none of the constraints prefers a dark [ɫ] to a light [l] in this form.

For the six minority forms discussed here, the subjects had six different ungrammaticality judgments when comparing light [l] with dark [ɫ], in the order *[fiːl] > *[ɫait] > ??[fiːlɪt] > ?[fɹiəɫiː] > ?[fiːliː] > √[gɹiəɫiː]. The four symbols "*", "??", "?", and "√" were not enough to capture the judgments, since the subjects really considered [fiːl] worse than [ɫait], and [fɹiəɫiː] worse than [fiːliː]. Rather than trying to extend the number of discrete ungrammaticality symbols, one has to conclude that grammaticality judgments are continuously gradient. The point that Boersma & Hayes made was that stochastic OT can model these continuously gradient grammaticality judgments, if these judgments can be assumed to correspond to relative frequencies in production. This assumption has been challenged directly by Keller & Asudeh (2002) on two grounds, which I address in Sections 2 and 3.

2. Frequencies are relative to the frequency of the underlying form

The first way in which grammaticality judgments deviate from *absolute* corpus frequencies is a trivial one, and would not deserve a separate section in this paper if not surprisingly many criticists of a relation between frequency and grammaticality had used it as an argument. In the example of Section 1, the form [fɹiəɫiː] is considered worse than [fiːliː], although [fɹiəɫiː] is predicted more often in a large corpus than [fiːliː]. We can see this in the following way. A search with the Internet searching facility *Google* yields 5,620,000 occurrences of the common word *freely*, and only 74,700 occurrences of the word (*touchy*-)*feely*. The ranking of Figure 1 predicts that 4 percent of the *freely* tokens, or about 220,000, will be pronounced as [fɹiəɫiː], and that 17 percent of the (*touchy*-)*feely* tokens, i.e. about 12,000, will be pronounced as [fiːliː]. Thus, what corresponds to a grammaticality judgment in not the *absolute* frequency of a form in a corpus (220,000 *vs.* 12,000), but a *relative* frequency, i.e. the *conditional* probability of a form *given the underlying form* (4 percent *vs.* 17 percent). This is a general feature of grammaticality judgments: they are always relative to the information given to the subjects. In Hayes' experiment, it was clear to the subjects what the underlying form was (the forms were not only spoken, but also presented orthographically), and it was equally clear to them that the focus was on the choice between the allophones [l] and [ɫ]. The subjects' grammaticality judgments for [fiːliː] therefore could not take into account the uncommonness of *feely*, neither could they take into account other than phonological aspects of the form.

Nevertheless, Keller & Asudeh (2002) criticize Stochastic OT precisely because of an alleged claim of an *absolute* frequency-grammaticality relationship. Their example is the pair of sentences in (7).

(7) *An alleged counterexample of the frequency-grammaticality relation*

a. *The athlete realized her goals.*
b. *The athlete realized her goals were out of reach.*

Since the second subcategorization option for *realize* (with a sentence as the object) is more common than the first option (with an NP as the object), Keller & Asudeh claim that Boersma & Hayes must predict that (7b) is more grammatical than (7a). But nothing of the sort is predicted by Boersma & Hayes' model. The two sentences mean different things, they have different underlying forms, and they are not candidates within a single tableau. Sentence (7b) competes only with forms that share its underlying form, for instance (7b′) *The athlete realized that her goals were out of reach* and perhaps (7b″) *That her goals were out of reach, the athlete realized*. It is likely that candidates (7b) and (7b′) get a "√" and candidate (7b″) gets a "?", and that these verdicts correspond to the relative corpus frequencies of these three construction types. A slightly more famous example, sometimes attributed to Chomsky, is the pair of sentences "I'm from Dayton Ohio" and "I'm from New York", whose comparable grammaticality but widely differing corpus frequency can be advanced to argue that corpus frequency and grammaticality cannot be related. Again, these forms have different meanings and do not compete. The form "I'm from Dayton Ohio" only competes with its synonyms "I'm from Dayton" and "I'm from Dayton Ohio USA", which will be judged as more grammatical than "I'm from Dayton Ohio" in the appropriate pragmatic context, for instance when speaking to people from Miamisburg Ohio and Berlin Germany, respectively.

Keller & Asudeh's interpretation of Boersma & Hayes' predictions is explained by their assertion (Sorace & Keller, to appear) that "subjects can judge the relative grammaticality of arbitrary sentence pairs", i.e. structures that are in different tableaus, with different underlying forms. But this does not mean that the underlying forms cannot be taken into account: indeed, it is clear that the subjects *have* to take into account hidden forms that they can reconstruct. A subject will judge the famous garden-path sentence *The horse raced past the barn fell* as ungrammatical if she interprets it as having the meaning of *The horse raced past the barn and fell* or *The horse raced past the barn, fell, and...*, probably because of high-ranked constraints against leaving out *and* and against truncation. Only when realizing or being told that *The horse raced past the barn fell* can be interpreted as *The horse that was raced past the barn fell*, will the subject change her grammaticality judgment. Even more clearly, the sentence *John shaves him every morning before breakfast* is ungrammatical in a context in which it can only mean that *John* and *him* refer to the same person, but it becomes grammatical if *him* can be made to refer to John's sick father who lives in John's house. Linguists abbreviate this situation by saying that the Logical Form *John*$_i$ *shaves him*$_j$ *every morning before breakfast* is grammatical, but *John*$_i$ *shaves him*$_i$ *every morning before breakfast* is not. In other words, grammaticality judgments are relative to known or reconstructed underlying forms, in phonology as well as in syntax.

3. Generalization to more than two candidates

As noted by Keller & Asudeh (2002), the Stochastic OT model of grammaticality judgments appears to have more trouble than in the case of Section 1 if the number of candidates is more than two.

3.1. Keller & Asudeh's apparent counterexample to Stochastic OT

Tableau (8) gives Keller & Asudeh's example of a case that they claim Stochastic OT cannot handle.[1]

(8) *The candidate with zero frequency is not the least grammatical candidate*

/S, O, V/	VERB	NOM	PRO	acceptability
☞ a. O[pro,acc] S[nom] V		*		+0.2412
b. O[acc] S[pro,nom] V		*	*	–0.0887
c. V S[pro,nom] O[acc]	*			–0.1861

This tableau is about constituent order in German subclauses. Underlyingly we have an unordered subject (S), object (O) and verb (V). The constraint VERB states that the finite verb should be the final constituent, the constraint NOM states that the nominative (nom) should precede the accusative (acc), and the constraint PRO states that a pronoun (pro) should precede a full NP. To clarify the situation with real German, (9) gives an example of each candidate (all NPs have to be masculine singular, so that there can be contrastive nominative/accusative case marking).

(9)

a. *dass ihn der Polizeibeamte erwischt*
'that him-ACC the-NOM policeman-NOM captures'
b. *dass den Dieb er erwischt*
'that the-ACC thief he-NOM captures'
c. *dass erwischt er den Dieb*
'that captures he-NOM the-ACC thief'

The problem that Keller & Asudeh note is that (8b) is judged as more grammatical than (8c), although (8b) must have zero corpus frequency since it has a proper superset of the violations of (8a); candidate (8c), by contrast, will have a non-zero frequency if VERB is ranked at a modest distance above NOM and PRO. Thus, the order of the corpus frequencies (8a > 8c > 8b) does not match the order of the grammaticality judgments (8a > 8b > 8c).

[1] They actually claim that this case is a failure of the Gradual Learning Algorithm (Boersma 1997, Boersma & Hayes 2001), not a failure of Stochastic OT. However, the Gradual Learning Algorithm is only designed to learn cases that can be described by a Stochastic OT grammar. For cases that cannot be described by a Stochastic OT grammar, no present or future learning algorithm will ever be able to come up with a working Stochastic OT grammar. So it must be Stochastic OT that apparently fails, not the Gradual Learning Algorithm.

3.2. Correction of Keller & Asudeh's example

We first note that the three candidates in (8) do not share the same underlying form, given the three constraints. That is, the fact that (8) has only three constraints means that the choice between full pronouns and full NPs is not handled by the grammar. Candidates (8b) and (8c) share the underlying form (S = 'he', O = 'the thief', V = 'capture'), whereas (8a) has a different underlying form (S = 'the policeman', O = 'he', V = 'capture'). Of course it is true that all three candidates could share the underlying form (S = 'the policeman', O = 'the thief', V = 'capture'), but in that case the grammar should handle the choice between pronouns and full NPs (i.e. between *ihn* and *den Dieb*, and between *er* and *der Polizeibeamte*), probably with the help of a constraint that bans full NPs when they are coreferential with a topic antecedent (and a constraint that forces full NPs when they are not). To enable Keller & Asudeh to maintain their point against Stochastic OT, I will assume that the three constraints in (8) are sufficient, hence that the pronouns are underlying, hence that (8) should be divided into the two tableaus (10) and (11).

(10) *Unsurprising grammaticality results*

|S = 'the policeman', O = 'the thief$_i$', V = 'capture', topic = 'the thief$_i$'|	VERB 105.0	NOM 98.0	PRO 98.0	accept-ability	corpus freq.	pairwise freq.
☞ a. *dass der Polizeibeamte ihn erwischt*			*	√	50%	83%
☞ b. *dass ihn der Polizeibeamte erwischt*		*		√	50%	83%
c. *dass erwischt der Polizeibeamte ihn*	*		*	*	0%	17%
d. *dass erwischt ihn der Polizeibeamte*	*	*		*	0%	17%

(11) *Candidate* b *has zero frequency but is not the least grammatical*

|S = 'the policeman$_i$', O = 'the thief', V = 'capture', topic = 'the policeman$_i$'|	VERB 105.0	NOM 98.0	PRO 98.0	accept-ability	corpus freq.	pairwise freq.
☞ a. *dass er den Dieb erwischt*				√	100%	100%
b. *dass den Dieb er erwischt*		*	*	??	0%	66%
c. *dass erwischt er den Dieb*	*			*	0%	34%
d. *dass erwischt den Dieb er*	*	*	*	*	0%	0%

Keller & Asudeh write that "in the [Stochastic OT] framework, differences in degree of grammaticality or frequency can only be predicted for structures that are in the same candidate set". Hence, they consider the separation into tableaus impossible. But as we have seen in Section 1, a comparison of two forms across tableaus is trivially possible if each of the two forms receives a judgment through a comparison with the other forms in its own tableau. Hence, the separation of the tableaus is possible. Moreover, I will now show that this separation is not only *possible*, but *necessary*. Consider the example by Pesetsky (1998) in (12) for the French noun phrase 'the man that I know' (the underlying form is mine, since Pesetsky does not give one).

(12) *A candidate that violates* LEFTEDGE(CP) *only is not grammatical*

|[*l'homme*$_i$ [*que*$_C$ *je connais l'homme*$_i$]$_{CP}$]$_{NP}$| 'the man$_i$ [that$_C$ I know the man$_i$]$_{CP}$]$_{NP}$'	RECOVER-ABILITY	LEFTEDGE(CP)	TELEGRAPH
a. *l'homme*$_i$ [$_{CP}$ *qui*$_i$ *que*$_C$ *je connais*		*!	*
b. *l'homme*$_i$ [$_{CP}$ *qui*$_i$ ~~*que*~~$_C$ *je connais*		*!	
☞ c. *l'homme*$_i$ [$_{CP}$ ~~*qui*~~$_i$ *que*$_C$ *je connais*			*
d. *l'homme*$_i$ [$_{CP}$ ~~*qui*~~$_i$ ~~*que*~~$_C$ *je connais*		*!	

In (12), the underlying embedded NP *l'homme*, which is coreferential with the head of the embedding NP, will be realized (if at all) by the relative pronoun *qui*. It is understood that the embedded CP is headed by the complementizer *que*. The candidate generator (GEN) allows that both *qui* and *que* optionally occur in the output. Simplifyingly, GEN assumes here that if both of these elements occur in the output, the pronoun *qui* will be adjacent to its antecedent, hence the fixed order of *qui* and *que* in the four candidates. Unpronounced constituents are striked through in the tableau. The constraint LEFTEDGE(CP) demands that every subclause should start with its head, i.e. with the complementizer *que*; this is violated if *que* is left unpronounced (candidates 12b and 12d) or if anything precedes *que* in the CP (candidate 12a). The constraint TELEGRAPH demands that function words are not pronounced; it is violated by every instance of *que* in the output. The ranking of LEFTEDGE(CP) >> TELEGRAPH now leads to the pronunciation of *que*, but not of *qui* (one can easily see the English case, in which the two constraints are ranked equally high, so that 'the man who I know', 'the man that I know', and 'the man I know' are all grammatical, but 'the man who that I know' is not). Tableau (13) shows Pesetsky's example for 'the man I have danced with'.

(13) *A candidate that violates* LEFTEDGE(CP) *only is grammatical*

|[*l'homme*$_i$ [*que*$_C$ *j'ai dansé avec l'homme*$_i$]$_{CP}$]$_{NP}$| 'the man$_i$ [that$_C$ I have danced with the man$_i$]$_{CP}$]$_{NP}$'	RECOVER-ABILITY	LEFTEDGE(CP)	TELEGRAPH
a. *l'homme*$_i$ [$_{CP}$ *avec qui*$_i$ *que*$_C$ *j'ai dansé*		*	*!
☞ b. *l'homme*$_i$ [$_{CP}$ *avec qui*$_i$ ~~*que*~~$_C$ *j'ai dansé*		*	
c. *l'homme*$_i$ [$_{CP}$ ~~*avec qui*~~$_i$ *que*$_C$ *j'ai dansé*	*!		*
d. *l'homme*$_i$ [$_{CP}$ ~~*avec qui*~~$_i$ ~~*que*~~$_C$ *j'ai dansé*	*!	*!	

In (13), the last two candidates violate the high-ranked constraint RECOVERABILITY 'a syntactic unit with semantic content must be pronounced' because the content word *avec* 'with' is left unpronounced. Now compare (12b) with (13b). In Keller's (2002) model, both candidates are equally grammatical, since they violate the same set of constraints, i.e. the constraint LEFTEDGE(CP) only. This equal grammaticality is what you must get when comparing sentences directly, without their underlying forms. We can now see where Keller's model fails: French judges regard (13b) as grammatical but (12b) as ungrammatical. When considering the rest of the two tableaus, the explanation is straightforward: (13b) is optimal for its underlying form, whereas (12b) is not. Hence it is clear that grammaticality must be taken relative to the other candidates in the same tableau.

3.3. The problem does not go away

But the problem that Keller & Asudeh noted in (8) does not go away by correctly splitting it into (10) and (11). To see this, consider how these tableaus are handled in Stochastic OT. Tableau (10) is unsurprising: since (10a) and (10b) are both generally considered grammatical German, the constraints NOM and PRO must be ranked at about the same height; no matter where VERB is ranked, candidates (10c) and (10d) will be ungrammatical as a result of superset violations. Tableau (11) is a different matter. In the column 'acceptability' we see the (discretized) judgments of Keller & Asudeh's subjects, with (11b) being more acceptable than (11c), as we remember from (8). But corpus frequencies appear to tell us a different story: candidate (11a) will win in 100 percent of the cases, independently of the ranking of the three constraints, because all other candidates have superset violations. Thus, grammaticality judgments give (11b) > (11c), whereas corpus frequencies predicted by Stochastic OT give (11b) = (11c). The discrepancy is less clear than in (8), but it is still there. Keller's (2002) 'proposal' to solve the problem in (8) is to replace the strict ranking of OT with the additive ranking of what he calls "Linear OT", which is identical to Harmonic Grammar (Legendre, Miyata & Smolensky 1990ab), the predecessor of OT. Rather than replacing OT,[2] I propose that the solution lies in seeing what is compared to what in grammatical judgment. In a pairwise test *without considering the rest of the tableau*, (11b) will be judged as better than (11c), simply because it violates lower-ranked constraints. If (11a) and (11d) had not existed, (11b) would have won in 98.7 percent of the cases, (11c) in 1.3 percent (if VERB is ranked at 105.0 and the evaluation noise is 2.0). My proposal, now, is simply: an absolute grammaticality judgment for a surface form *X* can be derived from comparing *X* with every other candidate in its tableau. Thus, (11a) beats every other candidate and gets 100%; (11d) loses to every other candidate and gets 0%; (11b) wins from (11a) in 0% of the cases, from (11c) in 98.7%, and from (11d) in 100%, for an average of 66%; candidate (11c), finally, will get an average of (100+1.3+0)/3 = 34%. We can note that the order in the column 'pairwise frequency' of tableau (11) corresponds to the acceptability order. But we can go further and do the same for tableau (10), where e.g. candidate (10d) will lose to (10a) and (10b) in 100% of the cases but to (10c) in only 50%, for an average winning percentage of (0+0+50)/3 = 17%. When comparing the results across the tableaus, we predict that the grammaticality order is (11a) > (10ab) > (11b) > (11c) > (10cd) > (11d). This order matches perfectly with the result observed in Keller & Asudeh's experiment in (8), which was (10b) > (11b) > (11c).

Does this proposal contradict Boersma & Hayes' (2001) more direct relation between corpus frequency and grammaticality judgment? Well, for the cases in §1 there were only two candidates, so there can only be one pairwise comparison, so that the current proposal predicts that for cases with only two candidates grammaticality judgments correspond directly to corpus frequencies. Thus, Boersma & Hayes' case of English /l/ was just a special case of the more general proposal forwarded here. Our pairwise within-tableau

[2] For a defence of strict ranking, and a discussion of the relation between OT and Harmony Theory, see Prince & Smolensky (1993).

comparisons may not constitute the ultimate solution to the problem, but (10) to (13) show that neither Keller & Asudeh's (2002) proposal (judgments without regard to underlying forms) nor the most straightforward generalization from Boersma & Hayes (2001) (judgments match corpus frequencies even for more than two candidates) can be correct.

4. Remaining problem: prototypes are peripheral

Having established that the most accurate proposal to date is that there is a direct relation between grammaticality judgments and within-tableau pairwise evaluations with Stochastic OT, we are ready to attack a notorious example of the difference between judgments and frequencies, namely the "/i/ prototype effect" in phonology: if the experimenter asks a subject to choose the most /i/-like vowel from among a set of tokens that vary in their spectral properties, the subject will choose a very peripheral token, i.e. one with a very low first formant and a very high second formant (Johnson, Flemming & Wright 1993; Frieda, Walley, Flege & Sloane 2000). Such extreme formant values rarely occur in actual speech. Apparently, then, the token that the subject prefers is much more /i/-like than the average realization of the vowel /i/ is.

4.1. Why the /i/ prototypicality effect is a problem for linguistic models

The prototypicality judgment task involves a mapping from a discrete abstract surface form (SF), namely the segment /i/, to a continuous overt auditory form (OF), namely a value of the first formant (F1). The most common phonological grammar model contains just such a surface-to-overt mapping, and it is called *phonetic implementation*. This grammar model, probably believed in by a majority of phonologists, consists of a sequence of two mappings. The first of these maps an abstract underlying form (UF) to the surface form, and is called *phonology*. For instance, Hayes (1999) remarks that "Following Pierrehumbert (1980) and Keating (1985), I assume that there is also a phonetic component in the grammar, which computes physical outcomes from surface phonological representations. It, too, I think, is Optimality-theoretic [...]". This grammar model can be abbreviated as UF→SF→OF, i.e. a serial sequence of phonology and phonetic implementation. In this serial model, the /i/ prototypicality effect is a problem, since if the SF→OF mapping is used for both phonetic implementation and the prototypicality task, corpus frequencies (which result from phonetic implementation) should be the same as grammaticality judgments (whose best result is the prototype). For this reason, Johnson, Flemming & Wright (1993) proposed an intermediate representation, the *phonetic target*, which is 'hyperarticulated': UF→SF→HyperOF→OF. The prototypicality task, then, is proposed to tap HyperOF, whereas corpus frequencies obviously reflect OF. I will show, however, that in a listener-oriented grammar model this extra representation and extra processing stratum are superfluous, and the /i/ prototypicality effect arises automatically without invoking any additional machinery.

4.2. A grammar model for which the /i/ prototypicality effect is not a problem

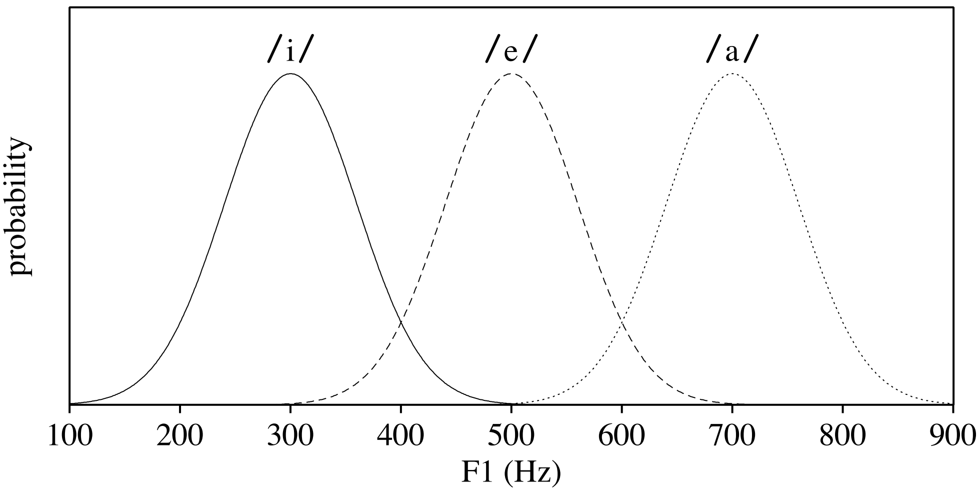
Functional Phonology (Boersma 1998) has no mapping from SF to OF. It does have the reverse mapping, OF→SF, which is called *perception* (I will defer discussion of production to §4.5). In general, perception is the process that maps continuous sensory information onto a more abstract mental representation. In phonology, perception is the process that maps continous auditory (and sometimes visual) information onto a discrete phonological representation. This OF→SF mapping is language-specific, and several aspects of it have been modelled in OT: categorizing auditory features to phonemes (Boersma 1997; Escudero & Boersma 2003, to appear), building metrical foot structure (Tesar 1997, Tesar & Smolensky 2000), and assigning autosegmental elements for tone (Boersma 2000) and nasality (Boersma 2003). An analogous mapping exists in syntax, namely the mapping from Phonetic Form (PF) to Logical Form (LF), which can be called *syntactic interpretation*. This is uncontroversially language-specific as well: although two overt strings of different languages can be word-by-word translations of each other, their corresponding logical forms may differ with respect to tree structure, movements, scope, and binding.

4.3. Formalizing perception in Stochastic OT

Since phonological perception and syntactic interpretation are language-dependent, it is useful to model them by linguistic means, and I will assume that the OF→SF and PF→LF mappings can be handled with Optimality-Theoretic grammars that may contain abstraction constraints, which evaluate the relation between OF and SF or between PF and LF, and structural constraints, which evaluate the output representation SF or LF.

In our case, perception maps F1 and F2 values to vowel segments such as /i/. For simplicity, I will discuss the example of a language with three vowels, /a/, /e/ and /i/, in which the only auditory distinction between these vowels is their F1 values. Suppose that the speakers realize these three vowels most often with the F1 values of 700 Hz, 500 Hz, and 300 Hz, respectively, but that they also vary in their realizations. If this variation can be modelled with Gaussian curves with standard deviations of 60 Hz, the distributions of the speakers' productions will look as in figure (14).

(14) *Production distributions of three vowels*



Now how do listeners classify incoming F1 values, i.e. to which of the three categories /a/, /e/ and /i/ do they map a certain incoming F1 value *x*? Following Escudero & Boersma (2003, to appear), I assume that this mapping is handled by an Optimality-Theoretic constraint family that can be abbreviated as "an F1 of *x* Hz is not the vowel *y*", for all values of *x* between 100 and 900 Hz and all three vowels *y*. Examples are in (15).

(15) *Categorization constraints*

"an F1 of 340 Hz is not /a/"
"an F1 of 340 Hz is not /e/"
"an F1 of 340 Hz is not /i/"
"an F1 of 539 Hz is not /i/"

The ranking of these constraints has to result from perceptual learning. Let us intercept the acquisition process at a point where the listener has already learned that this language has three vowel categories and already has correct lexical representations, so that if she misperceives the speaker's intended /pit/ as /pet/, her lexicon, which contains the underlying form |pit|, will tell her that she should have perceived /pit/. When detecting an error in this way, the learner will take action by changing the ranking of some constraints. Suppose that at some point during acquisition some of the constraints are ranked as in (16). The learner will then perceive an incoming F1 of 380 Hz as the vowel /e/, which is indicated by the pointing finger in (16). We can also read from (16) that 320 Hz will be perceived as /i/, and 460 Hz as /e/.

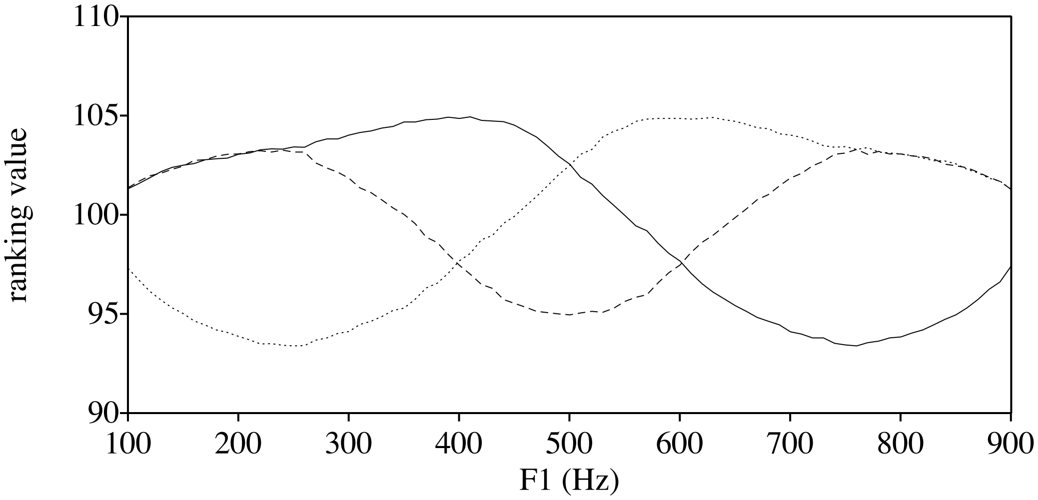
(16) *Learning to perceive vowel height*

[380 Hz] (UF = |i|)	320 Hz not /a/	380 Hz not /a/	460 Hz not /i/	320 Hz not /e/	460 Hz not /a/	380 Hz not /i/	380 Hz not /e/	320 Hz not /i/	460 Hz not /e/
/a/		*!							
☞ /e/							←*		
√ /i/						*!→			

If the lexicon now tells the learner that she should have perceived /i/ instead of /e/, she will regard this as the correct adult SF, as indicated by the check mark in (16). According to the Gradual Learning Algorithm for Stochastic OT (Boersma 1997, Boersma & Hayes 2001), the learner will take action by raising the ranking value of all the constraints that prefer the adult form /i/ to her own form /e/ (here only "380 Hz is not /e/") and by lowering the ranking value of all the constraints that prefer /e/ to /i/ (here only "380 Hz is not /i/").

To see what final perception grammar this procedure leads to, I ran a computer simulation analogous to the one by Boersma (1997). A virtual learner has 243 constraints (F1 values from 100 to 900 Hz in steps of 10 Hz, for all three vowel categories), and starts out with all of them at the same ranking value of 100.0. The learner then hears 10 million F1 values randomly drawn from the distributions in (14), which an equal probability of 1/3 for each vowel. She is subjected to the learning procedure exemplified in (16), with full knowledge of the lexical form, with an evaluation noise of 2.0, and with a *plasticity* (the amount by which ranking values rise or fall when a learning step is taken) of 0.01. The result is shown in (17).

(17) *The final ranking of* "an F1 of *x* Hz is not /vowel/", *for the vowels* /i/ *(solid curve)*, /e/ *(dashed curve), and* /a/ *(dotted curve)*



The figure is to be read as follows. F1 values under 400 Hz will mostly be perceived as /i/, since there the constraint "an F1 of *x* Hz is not /i/" (the solid curve) is ranked lower than the constraints "an F1 of *x* Hz is not /e/" (the dashed curve) and "an F1 of *x* Hz is not /a/" (the dotted curve). Likewise, F1 values above 600 Hz will mostly be perceived as /a/, and values between 400 and 600 Hz mostly as /e/. For every F1 value the figure shows us not only the most often perceived category but also the degree of variation. Around 600 Hz, /e/ and /a/ perceptions are equally likely. Above 600 Hz it becomes more likely that the listener will perceive /a/, and increasingly so when the distance between the curves for /e/ and /a/ increases. This distance is largest for F1 values around 750 Hz, where there are 99.8% /a/ perceptions and only 0.1% perceptions of /i/ and /e/ each. Above 750 Hz, the curves approach each other again, leading to more variation in categorization.

The curves can be explained as follows. Between 250 and 750 Hz, the curves are approximately probability-matching: the probability that a listener perceives vowel *y* is close to the probability that the vowel was intended as *y* by the speaker. Outside this region we see the effects of low corpus frequencies: around 100 and 900 Hz the curves are very close because the listener has heard very few utterances with extreme F1 values.

4.4. Formalizing the prototypicality task in Stochastic OT

While phonological perception (OF→SF) and syntactic interpretation (PF→LF) are *linguistic* mappings, i.e. they have been optimized during human evolution and come with their own learning algorithms, the reverse mappings (SF→OF and LF→PF) are *paralinguistic*, i.e. they are not used in normal language use and can be elicited only in experimental tasks such as the prototypicality judgment task: "of all the vowels that you are going to hear, choose the best /i/." How would a subject in such an experiment proceed? She has nothing more than the constraints that handle the OF→SF and PF→LF mappings, and their ranking values. The simplest strategy available is to invert these mappings by using SF or LF as the input and OF or PF as output candidates, and borrowing the existing constraint ranking as the decision mechanism. Tableau (18) shows how this works for /i/.

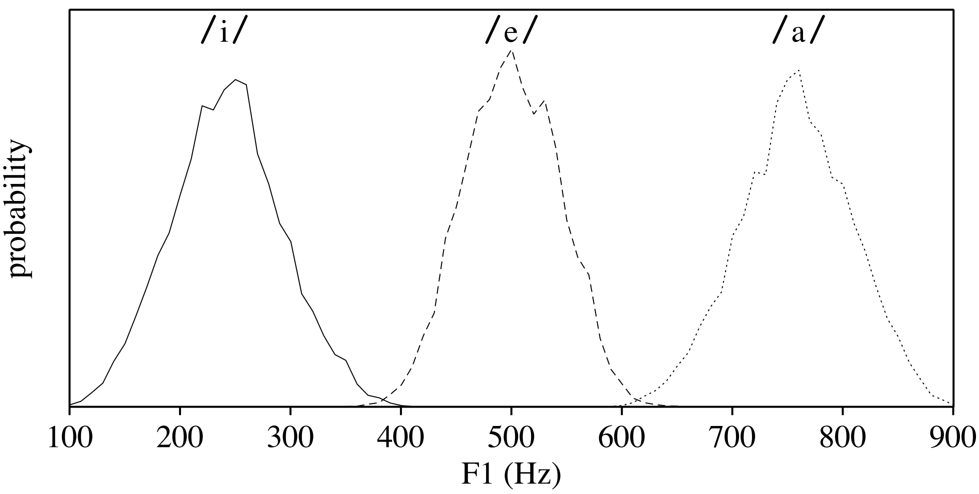
(18) *The auditory F1 value that gives the best* /i/

/i/	320 Hz not /i/	310 Hz not /i/	170 Hz not /i/	180 Hz not /i/	300 Hz not /i/	190 Hz not /i/	290 Hz not /i/	200 Hz not /i/	280 Hz not /i/	210 Hz not /i/	270 Hz not /i/	230 Hz not /i/	220 Hz not /i/	240 Hz not /i/	260 Hz not /i/	250 Hz not /i/
[170 Hz]			*!													
[180 Hz]				*!												
[190 Hz]						*!										
[200 Hz]								*!								
[210 Hz]										*!						
[220 Hz]													*!			
[230 Hz]												*!				
[240 Hz]														*!		
☞ [250 Hz]																*
[260 Hz]															*!	
[270 Hz]											*!					
[280 Hz]									*!							
[290 Hz]							*!									
[300 Hz]					*!											
[310 Hz]		*!														
[320 Hz]	*!															

With the ranking in (17) and zero evaluation noise, the listener will choose an F1 of 250 Hz as the optimal value for /i/. This is more peripheral (more towards the edge of the F1 continuum) than the most often produced /i/, which is 300 Hz. The size of the effect (50 Hz) is comparable to the effect found by Johnson, Flemming & Wright (1993) and Frieda, Walley, Flege & Sloane (2000). If the evaluation noise is 2.0, the outcome will vary as in figure (19), which was computed by running each of the three vowels 100,000 times through the inverted perception grammar.

The production/prototypicality differences seen when comparing figures (14) and (19) are very similar to the production/perception differences in the experiments by Johnson, Flemming & Wright (1993) and Frieda, Walley, Flege & Sloane (2000). To sum up: if the prototype task borrows its constraint ranking from perception, auditorily peripheral segments will be judged best if their auditory values are extreme, because the perceptual constraints have automatically been ranked in such a way that extreme auditory values are least likely to be perceived as anything else.

(19) *Prototypicality distributions for the three vowels*



4.5. Formalizing production

Now that we have seen how inverted perception accounts for the /i/ peripherality effect, we still have to see how it is possible that these same peripheral values are not used in *production*. The answer is that in production more constraints play a role. Articulatory constraints evaluate the articulatory ease of overt candidate forms (OF). Faithfulness constraints evaluate the extent to which the surface form (SF) resembles the underlying form (UF). In the listener-oriented model of Boersma (1998), the three representations are connected as UF→(OF→SF), not as UF→SF→OF as in the sequential production model. The abbreviation UF→(OF→SF) means that the perceived surface form (SF) is a function from the overt form (OF), i.e. in the production tableau the speaker chooses an overt form (e.g. an F1 value), but evaluates this overt form partly on the basis of the extent to which she thinks the listener will be able to reconstruct the abstract surface form (e.g. /i/). Boersma (2003) argues, for instance, that if faithfulness constraints are to capture all cases of neutralization (i.e. different underlying forms with the same overt form), the production model has to be UF→(OF→SF), not UF→SF→OF. Tableau (20) shows how an underlying |i| could be produced. Tableau (20) shows how a conflict between articulatory constraints and faithfulness constraints is resolved. In each candidate cell we see three representations: articulatory, auditory, and surface. The overt form is divided into an articulatory part and an auditory part. The articulatory part shows the gestures needed for articulating [i]-like sounds. For simplicity I assume that the main issue is the precision with which the tongue has to be bulged towards the palate, and that more precision yields lower F1 values, e.g. a precision of '7' yields an F1 of 240 Hz whereas a precision of '1' yields an F1 of 330 Hz. These precison values are evaluated by articulatory constraints that are ranked by the amount of effort involved, i.e. "the precision should not be greater than '6'" has to outrank "the precision should not be greater than '4'". The arrows in (20) indicate perception: each of the four auditory F1 values is turned into a vowel category by the grammar in (17). The probability that this vowel will be /i/ can be computed from (17) as well and is recorded in (20). The probability that the perceived vowel is identical to the underlying intended vowel |i| is evaluated by the usual underlying-surface identity constraint IDENT_{US} (McCarthy & Prince 1995), which is ranked here directly by confusion probability. Thus, "surface should be identical to underlying in more than 80% of the cases" has to outrank "surface should be identical to underlying in more than 90% of the cases". The result of all these constraints, with the ranking in (20), is that the overt form [F1 = 300 Hz] wins. Forms with lower F1 are too effortful, forms with higher F1 too confusable.

Tableau (20) may not be the final word on how production proceeds, but it is included here to illustrate the principles that have to be weighed in production. The result is very different from the prototypicality task. One can easily see that if the faithfulness constraints had been the only constraints in (20), the same candidate would have won as in (18). Thus, we must conclude that the difference between the two tasks can be reduced to the presence of the articulatory constraints in production and their absence in perception. The acquisition of production will lead to the child learn to match the corpus probabilities of her environment (Boersma & Hayes 2001), whereas the acquisition of perception will lead the child to choose more peripheral tokens in the prototypicality task.

(20) *Producing vowel height*

|i|	*[prec] > 6	IDENT_{US} (> 80%)	*[prec] > 4	IDENT_{US} (> 90%)	IDENT_{US} (> 98%)
$[\text{prec} = 9]_{Art}$ $[\text{F1} = 225\text{ Hz}]_{Aud}$ → 97% /i/	*!		*		*
$[\text{prec} = 7]_{Art}$ $[\text{F1} = 250\text{ Hz}]_{Aud}$ → 99% /i/	*!		*		
$[\text{prec} = 5]_{Art}$ $[\text{F1} = 275\text{ Hz}]_{Aud}$ → 95% /i/			*!		*
$[\text{prec} = 3]_{Art}$ $[\text{F1} = 300\text{ Hz}]_{Aud}$ ☞ → 86% /i/				*	*
$[\text{prec} = 1]_{Art}$ $[\text{F1} = 325\text{ Hz}]_{Aud}$ → 70% /i/		*!		*	*

4.6. Comparison with earlier explanations

The resulting tableaus automatically predict that, if the child is given enough time to learn even the rare overt forms, the best OF is the one that is least likely to be perceived as anything else than the given SF. This is the natural explanation in a grammar model without direct SF→OF mappings, such as Boersma's (1998) listener-oriented grammar model, where comprehension is OF→SF→UF and production is UF→(OF→SF). Such a grammar model can only be implemented in the framework of Optimality Theory, not in a serial rule-based framework, since the reversed 'phonetic implementation' in production, i.e. the '(OF→SF)' between parentheses, can easily be incorporated into the candidate cells of the production tableau. No such option was available to Johnson, Flemming & Wright (1993), who had to posit a serial production model with an extra representation, the 'hyperarticulated phonetic target'. Frieda, Walley, Flege & Sloane (2000) invoke a similar extra representation, the 'prototype'. With Occam's razor, the listener-oriented explanation has to be preferred, since it does not have to invoke the help of this otherwise unsupported representation.[3]

4.7. Predictions for grammaticality judgments

We have seen that an F1 of 240 Hz turns out to be a more grammatical instance of /i/ than an F1 of 300 Hz, although it has lower corpus frequency.

Analogous examples can be predicted in syntax. The speaker, for a given target logical form (TF), chooses a pronunciation (phonetic form, PF) that strikes a balance between speaker-based requirements (at PF) and listener-based requirements, which are nothing else than faithfulness, i.e. the extent to which the logical form (LF) that is constructed by the listener resembles the target form. The optimizable linguistic mappings are again production, i.e. TF→(PF→LF), and interpretation, i.e. PF→LF. I propose that the grammaticality judgment task is a simple LF→PF mapping ("given a meaning, choose the overt form that comes closest to implementing it"), and that this mapping uses the same constraint ranking as interpretation. We will thus find grammaticality versus corpus frequency mismatches in those cases where speakers use a non-low-ranked constraint at PF. If, for instance, Heavy NP Shift is speaker-based, corpus frequencies of this shift will be larger than grammaticality judgments would predict; if the shift is listener-based, frequencies and judgments would match. It has been reported (John Hawkins, p.c.) that, yes, corpus frequencies of Heavy NP Shift are larger than judges would like, so we can tentatively conclude that a speaker-based constraint is involved there.

Conclusion

We have seen that in a listener-oriented model of grammar, the bidirectional use of a single OT perception grammar may cause a mismatch between corpus frequencies and acceptability judgments if constraints at OF or PF play a role in production. The simplest recipe for arriving at a relative judgment when starting from *production* tableaus now seems to be the following: (1) delete the columns with constraints that evaluate OF or PF, from every tableau; (2) compare every candidate with all other candidates in its own tableau with for its satisfaction of the remaining constraints (faithfulness constraints and constraints that evaluate SF or LF); (3) turn the average percentages of these comparisons into a judgment; (4) if needed, compare these judgments across tableaus.

The key point is that the language user only optimizes her grammar for the core linguistic tasks, namely production and comprehension. She does not optimize her grammar for handling paralinguistic tasks such as grammaticality and prototypicality judgments, and that is why we find mismatches with corpus frequencies in precisely those tasks.

[3] Kuhl (1991) gave another reason for positing the 'prototype' level, namely the 'perceptual magnet' effect. However, Boersma, Escudero & Hayes (2003) show that in an OT model of distributional learning, the perceptual magnet effect emerges as an epiphenomenon.

References

Boersma, Paul (1997). How we learn variation, optionality, and probability. *Proceedings of the Institute of Phonetic Sciences* **21**: 43–58. University of Amsterdam.

Boersma, Paul (1998). *Functional phonology: formalizing the interactions between articulatory and perceptual drives*. Doctoral dissertation, University of Amsterdam.

Boersma, Paul (2000). *The OCP in the perception grammar*. Ms., University of Amsterdam. [Rutgers Optimality Archive **435**, http://roa.rutgers.edu]

Boersma, Paul (2003). Overt forms and the control of comprehension. In Jennifer Spenader, Anders Eriksson, and Östen Dahl (eds.), *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Department of Linguistics, Stockholm University. 47–56.

Boersma, Paul, Paola Escudero, and Rachel Hayes (2003). Learning abstract phonological from auditory phonetic categories: An integrated model for the acquisition of language-specific sound categories. *Proceedings of the 15th International Congress of Phonetic Sciences*. 1013–1016.

Boersma, Paul, and Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry* **32**: 45–86.

Escudero, Paola, and Paul Boersma (2003). Modelling the perceptual development of phonological contrasts with Optimality Theory and the Gradual Learning Algorithm. In Sudha Arunachalam, Elsi Kaiser, and Alexander Williams (eds.), *Proceedings of the 25th Annual Penn Linguistics Colloquium. Penn Working Papers in Linguistics* **8.1**: 71–85.

Escudero, Paola, and Paul Boersma (to appear). Bridging the gap between L2 speech perception research and phonological theory. To appear in *Studies in Second Language Acquisition*.

Frieda, Elaina M., Amanda C. Walley, James E. Flege, and Michael E. Sloane (2000). Adults' perception and production of the English vowel /i/. *Journal of Speech, Language, and Hearing Research* **43**: 129–143.

Hayes, Bruce (1999). Phonetically-driven phonology: the role of Optimality Theory and Inductive Grounding. In Michael Darnell, Edith Moravcsik, Michael Noonan, Frederick Newmeyer, and Kathleen Wheatley (eds.) *Functionalism and Formalism in Linguistics*, Vol. I: *General Papers*. Amsterdam: John Benjamins. 243–285.

Hayes, Bruce (2000). Gradient well-formedness in Optimality Theory. In Joost Dekkers, Frank van der Leeuw, and Jeroen van de Weijer (eds.), *Optimality Theory: phonology, syntax, and acquisition*. Oxford: Oxford University Press. 88–120.

Johnson, Keith, Edward Flemming, and Richard Wright (1993). The hyperspace effect: phonetic targets are hyperarticulated. *Language* **69**: 505–528.

Keller, Frank (2002). *Gradience in grammar: experimental and computational aspects of degrees of grammaticality*. Doctoral thesis, University of Edinburgh.

Keller, Frank, and Ash Asudeh (2002). Probabilistic learning algorithms and Optimality Theory. *Linguistic Inquiry* **33**: 225–244.

Kuhl, Patricia K. (1991). Human adults and human infants show a "perceptual magnetic effect" for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* **50**: 93–107.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky (1990a). Harmonic Grammar — A formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. In *Proceedings of the Twelfth Annual Conference of the Cognitive Sciences*. Cambridge, Mass.: Lawrence Erlbaum. 388–395.

Legendre, Géraldine, Yoshiro Miyata, and Paul Smolensky (1990b). Harmonic Grammar — A formal multi-level connectionist theory of linguistic well-formedness: an application. In *Proceedings of the Twelfth Annual Conference of the Cognitive Sciences*. Cambridge, Mass.: Lawrence Erlbaum. 884–891.

McCarthy, John, and Alan Prince (1995). Faithfulness and reduplicative identity. In Jill Beckman, Laura Walsh Dickey & Suzanne Urbanczyk (eds.) *Papers in Optimality Theory*. University of Massachusetts Occasional Papers **18**. Amherst, Mass.: Graduate Linguistic Student Association. 249–384.

Pesetsky, David (1998). Some optimality principles of sentence pronunciation. In Pilar Barbosa, Danny Fox, Paul Hagstrom, Martha McGinnis, and David Pesetsky (eds.), *Is the best good enough?: optimality and competition in syntax*. Cambridge, Mass.: MIT Press. 337–383.

Prince, Alan & Paul Smolensky (1993). *Optimality Theory: Constraint Interaction in Generative Grammar*. Technical Report TR-**2**, Rutgers University Center for Cognitive Science.

Sorace, Antonella, and Frank Keller (to appear). Gradience in linguistic data. *Lingua*.

Tesar, Bruce (1997). An iterative strategy for learning metrical stress in Optimality Theory. In Elizabeth Hughes, Mary Hughes, and Annabel Greenhill (eds.), *Proceedings of the 21st annual Boston University Conference on Language Development*, 615–626. Somerville, Mass.: Cascadilla.

Tesar, Bruce, and Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, Mass.: MIT Press.