Phonology **26** (2009) 147–185. © Cambridge University Press 2009
doi:10.1017/S0952675709001742 Printed in the United Kingdom

*Linking speech errors and phonological grammars : insights from Harmonic Grammar networks**

Matthew Goldrick
Robert Daland
Northwestern University

Phonological grammars characterise distinctions between relatively well-formed (unmarked) and relatively ill-formed (marked) phonological structures. We review evidence that markedness influences speech-error probabilities. Specifically, although errors result in unmarked as well as marked structures, there is a markedness asymmetry: errors are more likely to produce unmarked outcomes. We show that stochastic disruption to the computational mechanisms realising a Harmonic Grammar (HG) can account for the broad empirical patterns of speech errors. We demonstrate that our proposal can account for the general markedness asymmetry. We also develop methods for linking particular HG proposals to speech-error distributions, and illustrate these methods using a simple HG and a set of initial consonant errors in English.

1 Phonological markedness and linguistic behaviour

A central concern of generative phonological theory is to characterise the relative well-formedness of phonological structures. In this paper we will use the term MARKEDNESS to refer to distinctions in well-formedness (where well-formed structures are UNMARKED and ill-formed structures are MARKED). We adopt a broad definition of markedness, encompassing binary as well as gradient well-formedness distinctions both within a language and across languages. The characterisation of markedness distinctions has played a central role in accounting for 'internal' evidence

* This research was supported by National Institutes of Health grant DC0079772. Portions of this work were presented at the Experimental Approaches to Optimality Theory workshop (Ann Arbor 2007); we thank the workshop participants for helpful feedback. For helpful feedback on the proof in the Appendix we thank Alexander Getmanenko and Marta Sales. Andries Coetzee and Joe Pater provided invaluable feedback on the manuscript. Finally, for helpful discussions on this work and related issues over many years, we thank Adam Buchwald, Brenda Rapp, Paul Smolensky and Colin Wilson.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

(Zwicky 1980) regarding the distribution of phonological structures within and across human languages. However, many theorists have claimed that markedness plays a central role in all aspects of phonological cognition – i.e. the production, perception and acquisition of phonological structures and patterns (e.g. Jakobson 1941, Smolensky *et al*. 2006).

One particular aspect of linguistic behaviour that has been claimed to be influenced by markedness is speech errors – unintentional deviations from some target linguistic form of an utterance. For example, when producing utterances quickly, speakers often switch two sounds occurring in similar environments (e.g. *she sells* may be misproduced as *see shells*). A long tradition of research claims that error probabilities are influenced by markedness. For example, Jakobson (1941) claims that universal markedness principles influence errors arising subsequent to neurological impairments (see Blumstein 1973 for the first of many studies exploring these claims more systematically). As we review in more detail in §2, a substantial body of work suggests that although errors arising within phonological processes in production result in both marked and unmarked outcomes, such errors are biased to produce unmarked structures. For example, in coda position voiced obstruents (e.g. /d/) are often assumed to be marked relative to voiceless obstruents (e.g. /t/ – this accounts for processes such as final devoicing; Lombardi 1999). Errors not only change relatively marked structures into unmarked structures (e.g. /d/ → [t]); at a lower frequency errors also change unmarked outputs into marked outputs (e.g. /t/ → [d]).

Although a great deal of empirical work has examined broad correlations between various indices of markedness (e.g. typological and within-language frequency) and speech-error probabilities (see §2 for a review), little work has addressed the question of how to link specific grammatical theories to speech-error distributions. In §§3 and 4 of the paper, we address this question in the context of Optimality Theory (OT; Prince & Smolensky 1993) and the related formalism of Harmonic Grammar (HG; Legendre *et al*. 1990, Smolensky & Legendre 2006a). In §3, we discuss how many existing grammatical accounts of probabilistic behaviour are not appropriate models for speech errors; they cannot account for the occurrence of errors in which unmarked structures are replaced by structures that are more marked along all dimensions of representation (i.e. harmonically bound outputs). In §4 we propose a novel mechanism for linking grammars to speech-error data. We show that stochastic disruption to the computational mechanisms realising an HG can account for the broad empirical patterns of speech errors. We document specific methods for linking particular HGs to speech-error distributions and illustrate their application to a set of speech-error data from English.

2 The markedness asymmetry in speech errors

Speech errors are unintentional deviations from the target form one intends to produce. Of interest here are errors targeting phonological

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

structures; for example, in spoonerisms the initial portion of two words exchange (e.g. *dear old queen* → *queer old dean*). These errors can arise spontaneously, in experimental tasks (e.g. tongue twisters), as well as subsequent to acquired neurological impairments (e.g. aphasia). Errors have played a central role in the development of psycholinguistic theories of speech production (see Meyer 1992 for a critical review of evidence from spontaneous and experimentally induced errors, and Rapp & Goldrick 2006 for a review of evidence from aphasia).

A number of studies have shown that the probability of an error resulting in a relatively unmarked structure tends to be greater than the probability of an error resulting in a more marked structure. For example, in English, the consonant [h] is subject to a phonotactic constraint that prevents it from occurring in coda position. Consistent with the markedness of /h/ in this prosodic position in English, English speech errors resulting in [h] are much more likely to appear in onset position than coda position (e.g. errors like /mæd/ → [hæd] are much more likely than errors like /dæm/ → [dæh]). We refer to this asymmetry in error probabilities as the MARKEDNESS ASYMMETRY. The sections below review data from spontaneous speech-error corpora, experimentally induced speech errors and neurological impairment that document this asymmetry.

2.1 Categorical markedness distinctions

The English example above illustrates a case where an error outcome is *illegal* – i.e. categorically marked within a language. Analyses of a number of spontaneous speech-error corpora in a number of languages have shown that illegal error outcomes have much lower probability than legal outcomes (Arabic: Abd-El-Jawad & Abu-Salim 1987; English: Vousden *et al.* 2000; German: MacKay 1972; Mandarin: Wan & Jaeger 1998). Similar results have been reported in cases of acquired speech-production impairments. For example, the speech of individuals with 'jargon aphasia' (fluent but semantically empty speech composed of both words and nonwords) is typically reported to respect the phonotactic constraints of the individual's native language (see Marshall 2006 for a review).

However, it is important to note (*contra* many early analyses of speech errors) that although legal errors are more likely than illegal errors, both types of errors occur. Illegal errors are occasionally observed in spontaneous speech-error corpora (Stemberger 1983). Experimental studies have reported large numbers of errors that violate phonotactic constraints on segment sequences (e.g. English: Butterworth & Whittaker 1980). Similarly, in experimental studies where participants acquire novel categorical phonotactic constraints (e.g. English speakers acquire a constraint banning /f/ in coda), participants' speech errors will occasionally violate the novel phonotactic constraint (Dell *et al.* 2000, Goldrick 2004, Taylor & Houghton 2005, Warker & Dell 2006, Goldrick & Larson 2008, Warker *et al.* 2008). Markedness constraints on individual segments are also violated in speech errors. Instrumental studies have shown that in

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

experimental tasks participants' productions can be composed of illegal combinations of gestures (e.g. simultaneous tongue-tip and tongue-dorsum raising during the production of an English stop; Goldstein *et al.* 2007, Pouplier 2007, 2008); similarly, acoustic studies have shown that errors can result in segments outside the speaker's native language inventory (Laver 1980).[1] At multiple levels of phonological structure, then, speech errors can result in both legal and illegal structures; however, the former are more likely to occur.

2.2 Gradient markedness distinctions

Within any particular language, a large set of structures is legal. Within this set, many theories distinguish varying degrees of markedness. One criterion for distinguishing less from more marked legal structures within a language is their cross-linguistic markedness – as assessed by typological frequency in inventories, susceptibility to phonological processes, etc. For example, although English admits both [s] and [z] word-finally, [z] is cross-linguistically marked relative to [s] in this position (Lombardi 1999). Another criterion often invoked is frequency (both type and token); within a language, certain sounds or sound sequences are relatively more common (both across the lexicon and across utterances). For example, because [z] is less frequent than [s] in English, some theories assume it to be more marked. As suggested by this example, these two criteria are difficult to distinguish; in general, cross-linguistically marked structures are infrequent (Zipf 1935, Greenberg 1966, Berg 1998, Frisch 2000). For the purposes of this discussion, we will not distinguish among these various criteria. Instead, we focus more broadly on the influence of gradient markedness distinctions within a language (as determined by some criterion) on the probability of speech errors.

Studies of experimentally induced speech errors (all in English) have shown that errors on non-word targets are more likely to occur on low-frequency *vs.* high-frequency structures (Kupin 1982); furthermore, when an error occurs under these conditions, it is more likely to result in a high-frequency *vs.* low-frequency structure (Motley & Baars 1975, Levitt & Healey 1985, Goldrick 2002).[2] Goldrick & Larson (2008) show that

[1] Additionally, studies have shown that error productions can be phonetically distorted – i.e. acoustic/articulatorily distinct from non-errorful productions (Mowrey & MacKay 1990, Frisch & Wright 2002, Pouplier & Hardcastle 2005, Goldrick & Blumstein 2006, Goldstein *et al.* 2007, Pouplier 2007, 2008, McMillan *et al.* 2009).

[2] Note that many studies of spontaneous speech errors – with real word targets – have not found asymmetries in the probability of more *vs.* less marked outcomes (e.g. Shattuck-Hufnagel & Klatt 1979). Furthermore, in a series of studies with real word targets, Stemberger has shown that the general preference for less marked error outcomes is coupled with a preference for more marked structures over 'default' forms (see Stemberger 2004 for a review). Following Stemberger, we attribute this effect to properties of lexical representations (i.e. phonological retrieval processes). We believe that these effects are minimised in the studies cited above, because they rely on non-word stimuli.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

the probability of participants' error outcomes are influenced by probabilistic phonotactic constraints acquired over the course of an experiment (e.g. the probability of producing an error that results in /s/ in coda increases as the relative frequency of /s/ in coda increases). Note, however, that in all of these cases gradient markedness (as indexed by frequency) merely biases the likelihood of outcomes. For example, although an /s/ → [z] error is less likely to occur than a /z/ → [s] error, both types occur.

Similar results have been reported in cases of aphasia. Several case studies have examined individuals with deficits affecting 'postlexical' phonological/phonetic processes in speech production (i.e. processes that follow retrieval of phonological information from the lexicon, but precede motoric implementation processes; see below for further discussion). These individuals tend to produce more errors on more marked structures; the resulting error outcomes tend to be relatively unmarked (English: Buchwald *et al*. 2007, Goldrick & Rapp 2007; French: Béland & Paradis 1997; Italian: Romani & Calabrese 1998, Romani *et al*. 2002, Romani & Galuzzi 2005). For example, Romani & Calabrese (1998) document the case of an Italian individual whose errors are influenced by sonority-sequencing constraints. He produces more errors on onset segments of higher sonority (e.g. error rate on nasals: 0·1%; obstruents: 0·02%) and his errors tend to result in segments of lower sonority (across segment types, 68% of errors result in segments of lower sonority, 22% increase sonority and 10% maintain the same sonority level). Similar results have been reported in studies that analyse the aggregate performance of a number of aphasic individuals (e.g. English: Blumstein 1973; Dutch: den Ouden 2002; French: Béland *et al*. 1993).

3 General issues in connecting speech-error patterns to grammatical mechanisms

3.1 The relevance of speech-error patterns for grammatical theories

It is well known that behavioural data of any kind reflect the interaction of a number of distinct cognitive processes. As noted by Chomsky (1980: 188) 'the system of language [e.g. components of grammar] is only one of a number of cognitive systems that interact in the most intimate way in the actual use of language'. For example, consider the speech error /kæt/ → [hæt]. This could reflect misselection of an unintended word (*hat*), correct selection of the intended word, but specification of an unintended phonological representation (e.g. [h] instead of [k]) or correct selection of the word, specification of its phonological structure, but misarticulation of the phonological targets (e.g. failure to make a sufficient velar closure).

In order to use speech-error data to draw inferences regarding cognitive structure, it is therefore critical to establish the processing locus of errors. We do so by evaluating the speech-error data above within the context of

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

current psycholinguistic theories of speech production. As reviewed by Goldrick & Rapp (2007), most of these theories assume that at least four broad stages of cognitive processes are involved in single-word speech production.[3] First is the process of LEXICAL SELECTION. A particular lexical entry is selected based on the speaker's intended meaning. For example, the concept 'furry four-legged feline' is used to retrieve the lexical entry <CAT>. Second, a PHONOLOGICAL RETRIEVAL process retrieves the phonological information associated with the selected lexical entry from long-term memory (e.g. /kæt/ is retrieved for <CAT>). This representation is analogous to the underlying representation assumed by generative phonological theories. Psycholinguistic theories often assume this representation to be relatively abstract (e.g. failing to specify many predictable aspects of phonological structure).[4] A subsequent PHONETIC ENCODING process therefore elaborates this representation, resulting in a more fully specified representation of sound structure that serves to drive the final, motoric stage of spoken production processing. This phonetic encoding process therefore computes a function analogous to that of the phonological component of generative grammars; it maps from an abstract underlying representation to a fully specified surface representation. This process is assumed to be sensitive to markedness distinctions.

The most direct evidence that the markedness asymmetry documented above arises specifically within phonetic encoding processes comes from cases of acquired neurological impairment. The case studies reviewed above report individuals who appear to have relatively selective deficits targeting phonetic encoding processes (English: Buchwald *et al.* 2007, Goldrick & Rapp 2007; French: Béland & Paradis 1997; Italian: Romani & Calabrese 1998). For example, Goldrick & Rapp (2007) report an individual (BON) who exhibited no comprehension or articulatory impairments, suggesting her deficit was localised to post-semantic, pre-articulatory processes. The absence of semantically related errors from her productions suggests lexical selection processes were intact. Furthermore, her performance was unaffected by a number of lexical properties of target items (e.g. word frequency, lexical neighbourhood density), suggesting that phonological retrieval processes were intact as well. Finally, consistent with a phonetic encoding deficit, her phonological errors were significantly influenced by various measures of markedness (e.g. phoneme frequency, cross-linguistic markedness of syllable structure).

As noted above, the speech errors of these individuals exhibit the markedness asymmetry; their errors are more likely to result in less *vs.* more marked forms, but both types of errors occur. This suggests that errors resulting in marked structures do not reflect 'contamination' of

[3] A fully specified theory will undoubtedly require additional processes, as well as further subdivision of the processes outlined above. This is intended to provide a sketch of the broad divisions between processes in production.

[4] This contrasts with assumptions of exemplar theories of speech production (Pierrehumbert 2002). See Goldrick & Rapp (2007) for evidence supporting the abstractness of representations at this level of processing.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

phonological speech errors by processes that are insensitive to markedness distinctions; such errors arise specifically within phonetic encoding processes. Furthermore, most of the speech errors produced by these individuals occurred in single-word production tasks (e.g. picture naming, single-word repetition). This suggests that the source of the errors is endogenous to phonetic encoding processes. For example, it is not the case that errors resulting in marked structures solely reflect contextual intrusions from other words in the sentence or discourse.

In sum, there is evidence that speech errors can arise within a component of the spoken production system that corresponds closely to that of the phonological component of the grammar. Errors arising within this processing component result in forms that are less as well as more marked than the target. However, there is a markedness asymmetry; errors tend to result in less marked forms. Having established a plausible link between (certain) speech errors and the phonological component of the grammar, we may now ask how to connect an OT or HG grammatical theory – which makes specific claims about markedness – to speech-error patterns.

3.2 Issues in using purely grammatical mechanisms to capture speech-error patterns

3.2.1 *Speech errors and harmonic ascent.* A natural place to begin is to consider adapting existing grammatical mechanisms to account for speech-error data. Note that speech errors are inherently probabilistic; an error is a probabilistically occurring 'variant' of the target form (e.g. an error such as /kæt/ → [kæd] exists in variation with correct productions /kæt/ → [kæt]). OT proposals for incorporating variation into the grammar might therefore offer a promising avenue for purely grammatical mechanisms to capture speech errors.

The issue is that many of these mechanisms are too restrictive in the types of variation they allow. As repeatedly noted above, error outcomes can result in more marked structures than their targets. In OT terms, this means that an unmarked input is mapped onto a more marked, unfaithful output. The critical cases concern forms where the unfaithful output is universally more marked than the input. For example, /kæt/ → [kæd] violates not only faithfulness constraints (e.g. IDENT[voice]): 'Corresponding segments in the input and output must have identical voicing specifications'), but also markedness constraints (e.g. *LAR: 'Laryngeal features should not appear in the output' – violated by voiced obstruents such as [d]). Critically, in this particular error, there is no markedness constraint that disprefers the target form relative to the error form (i.e. there is no constraint that disprefers final voiceless stops relative to voiced stops). The violations of the error form are therefore a superset of the violations of the fully faithful candidate. If this condition holds, there is no OT grammar that can specify this mapping; this candidate is harmonically bound by the fully faithful candidate (see Moreton 2004 for formal

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

analyses of this restriction). Prince (2007) refers to this principle as 'harmonic ascent'; in OT, input–output mappings must 'ascend' in harmony.

To illustrate this point more concretely, consider the tableau in (1).

(1) *OT tableau illustrating harmonic ascent*

/kæt/	IDENT[voi]	*LAR
a. kæt		
b. kæd	*	*

Given these two constraints, there is no ranking under which [kæd] will be the output of the grammar for input /kæt/. With respect to all constraints in the grammar, /kæt/ → [kæd] is dispreferred relative to /kæt/ → [kæt]; it violates harmonic ascent. Of course, this situation may be quite uncommon; for most pairs of candidates there is at least one constraint that prefers one member of the pair relative to the other. For example, consider two possible outputs for *lackey*: [læki] *vs.* [lægi]. With respect to a constraint like *LAR, the candidate with [k] is preferred; however, with respect to a constraint against intervocalic voiceless stops (e.g. *VTV), the candidate with [g] is preferred. The possibility of the existence of such constraints makes identifying mappings that violate harmonic ascent difficult; however, it is clear that in principle they can exist.

Several studies have reported significant numbers of errors which arguably violate harmonic ascent. For example, in tongue-twister tasks, a number of studies have reported significant numbers errors where initial stops are replaced by fricatives (e.g. although /s/ → [t] is more frequent than /t/ → [s], both error types are observed; Levitt & Healey 1985, Goldrick 2002). Fricatives are not only typologically dispreferred relative to stops (Maddieson 1984); in initial position they are more cross-linguistically marked with respect to sonority-sequencing constraints (Clements 1990). It is therefore unlikely that in absolute initial, prevocalic position of a word there is a markedness constraint preferring fricatives to stops. Errors violating harmonic ascent have also been documented in single-word production tasks in individuals with deficits affecting phonetic encoding (e.g. clusters substituting for singleton consonants (Goldrick & Rapp 2007); more sonorous segments replacing less sonorous segments in onset position (Romani & Calabrese 1998)).

Speech-error patterns therefore stand in marked contrast to grammatical patterns. Categorical phonological patterns observed in human languages do not violate harmonic ascent (Moreton 2004). It is likely that this generalisation holds for grammatical variation as well. Variationist research has repeatedly emphasised the parallels between variable and non-variable phonological patterns (e.g. Guy & Boberg 1997). Building on this assumption, many cases of empirically documented variation have been successfully analysed within variationist frameworks that respect harmonic ascent. In the next section, we illustrate how several current OT

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

and HG approaches to variation adopt this approach – modelling variation while respecting harmonic ascent. (We return to models which do not respect harmonic ascent below.)

3.2.2 *Harmonic ascent in variationist OT models.* In many variationist OT models, the assumption that variation respects harmonic ascent is most strongly embedded within frameworks that assume variation reflects the use of multiple strict rankings by an individual speaker. If these different rankings produce different outputs, variation will result. The most general version of this approach is adopted by Jarosz (2006), who assumes that the learner simply assigns a probability to each possible total ordering of the constraint set. Other proposals have adopted more restrictive mechanisms. One approach allows some constraints to take on multiple rankings relative to other constraints (e.g. by stating a partial order over the constraint set (Anttila 1997) or by allowing certain constraints to 'float' in their placement in a constraint hierarchy (Reynolds 1994)). An alternative mechanism associates constraints with a probability distribution along an (arbitrary) continuous scale; rank order on this scale determines total ordering (Boersma 1997; see also Boersma & Hayes 2001).

Critically, in all of these accounts variation reflects the presence of a set of OT grammars (i.e. strict rankings). By virtue of being an OT grammar, each member of this set necessarily respects harmonic ascent. Thus, these approaches predict that variation will necessarily respect harmonic ascent. Consider the constraints in (1). Following Anttila (1997), we could produce variation by specifying a partial order; the constraints *LAR and IDENT[voice] could be unranked with respect to one another. For an underlying form like /kæd/ this will result in variation. Half of the time the mapping /kæd/ → [kæd] will be optimal (IDENT[voice] ≫ *LAR); the rest of the time /kæd/ → [kæt] will be optimal (*LAR ≫ IDENT[voice]). But certain types of variation will never be produced. For an underlying form like /kæt/, no matter how the constraints are ranked in (1) the optimal output is [kæt]. These grammatical mechanisms can only produce outputs that are optimal under some constraint ranking. Since forms violating harmonic ascent are optimal under no ranking, these mechanisms cannot account for the full range of speech errors.

3.2.3 *Harmonic ascent in Harmonic Grammars.* Harmonic Grammar (Legendre *et al.* 1990, Smolensky & Legendre 2006a, Boersma & Pater 2008, Potts *et al.* 2008, Pater, to appear) is the connectionist antecedent of OT. Like OT, it specifies phonological grammars through a set of ranked, violable constraints. Unlike OT, these constraints are not strictly ranked; they are numerically weighted.

One method for specifying the output of an HG assumes that each constraint ranking (or weighting) specifies a deterministic mapping from underlying to surface forms. For each candidate, the sum of the weighted violations of each constraint determine its harmony. The output is the

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

candidate with the highest harmony. Note that like standard OT, under certain conditions this system is bound to respect harmonic ascent. Consider the HG tableau in (2). Assuming x and y are restricted to positive values, there is no weighting of the two constraints that will assign higher harmony to candidate (b). ($\mathcal{H}$ denotes the harmony of the candidate; each constraint column gives the constraint's total contribution to harmony.)

(2) *HG tableau illustrating harmonic ascent*

weight	x	y	$\mathcal{H}$
/kæt/	IDENT[voi]	*LAR	
a. kæt			0
b. kæd	$-x$	$-y$	$-x-y$

To generate variation within this framework, Boersma & Pater (2008) follow the method of Boersma (1997). For any particular instance of generating a surface form based on an underlying representation, the weights of the constraints are perturbed by some random amount. If two constraints have similar weight values, this noise will cause their relative strengths to switch, altering the relative harmony levels of relevant candidates. These shifts in the harmony of competing candidates can result in variation. However, assuming constraint weights are restricted to positive values, this approach makes similar predictions to that of Boersma (1997) – variation is restricted to the range of input–output mappings that can be specified by a grammar. Since all HGs with positive constraint weights respect harmonic ascent, this means that this HG model will be unable to account for the full range of speech-error patterns.

3.2.4 *Summary : purely grammatical approaches to speech errors.* Many of the existing OT mechanisms that can accommodate probabilistic phenomena tend to place strong restrictions on the range of variable patterns they can produce. Specifically, many variationist OT and HG grammars categorically ban underlying–surface mappings that violate the principle of harmonic ascent. This is a positive feature of such grammars, as both categorical and variable grammatical patterns within natural languages appear to respect this principle. In contrast, speech errors frequently violate this principle; these purely grammatical mechanisms are therefore inappropriate models of this form of linguistic behaviour.

It should be noted that some OT and HG mechanisms do allow for violations of harmonic ascent. Boersma & Pater (2008) examine HGs in which constraint weightings are not restricted to positive values. These can produce violations of harmonic ascent. For example, if $y < -x$ in (2), candidate (b) will be optimal. Other methods allow for violations of harmonic ascent by assuming the grammar specifies a probability distribution over candidates. Coetzee (2006) proposes that the rank order (from most to least harmonic) of the candidate set specifies the relative probability of

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

different outputs. This mechanism could, in principle, allow for violations of harmonic ascent. Consider the tableau in (1) above. The harmonic order of the candidates is [kæt] ≻ [kæd]. Therefore, according to this model, [kæd] could be produced, although with a relatively small probability. Similarly, many authors have situated HGs within the more general context of Maximum Entropy statistical models, which have been widely utilised in a number of computational applications (for specific applications to OT-type grammatical models, see Johnson 2002, Goldwater & Johnson 2003, Wilson 2006, Jäger 2007, Hayes & Wilson 2008). These models allow an HG to specify a probability distribution over candidates (where probability is a function of relative harmony). As in Coetzee's (2006) proposal, this allows for violations of harmonic ascent (Jäger & Rosenbach 2006).

4 Proposal: disruption to grammatical processing mechanisms

Rather than rely on grammatical mechanisms to model speech errors, we have chosen to pursue an alternative strategy that builds on psycholinguistic models (e.g. Dell 1986). These view speech errors as reflecting stochastic, temporary disruptions to speech-production mechanisms. We propose that speech errors arising during phonetic encoding processes reflect probabilistic disruptions to grammatical processing mechanisms. These disruptions alter the relative harmony of various output forms; this sometimes results in a competitor form being more harmonic than the correct target, resulting in an error. For example, when determining the optimal output for /kæd/, a probabilistic disruption might increase the harmony of [kæt], resulting in a /d/ → [t] error. After outlining our specific assumptions regarding grammar processing and disruptions to processing, we show how this accounts for the markedness asymmetry in speech errors. We then discuss several methods for directly relating specific grammatical theories to predicted speech-error probability distributions.

4.1 Connectionist mechanisms supporting the processing of HGs

Processing of OT and HG grammars can be realised by a number of different computational architectures. We have chosen to focus on a connectionist processing architecture in the interest of connecting this model of speech errors to other psycholinguistic approaches (which are overwhelmingly connectionist) and, more broadly, with theories that aim for an integrated approach to computation in the mind/brain (e.g. Smolensky 2006b).

We focus on one specific type of connectionist network – a Hopfield network (Hopfield 1982) – which is part of a broader class of networks that compute functions similar to those specified by HG (Smolensky 2006a). In such a network, a linguistic representation can be realised by a numerical pattern of activity over a set of representational units (in a

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

Hopfield network, these are limited to binary patterns – strictly 0's and 1's – although similar properties hold for networks that allow for continuous activation values; Hopfield 1984). A set of weights or connections links each of these units. These connections can be thought of constraints on the pairwise activity of units. For example, a negative weight signifies that the two units should not have the same activation value; a positive weight signifies that they should agree in activation value. (The strength of the weight represents the strength of this constraint.) During processing, units update their activation values so as to maximise the degree to which the activation pattern satisfies the constraints of the network. For example, if units A and B have a positive weight, and unit A is activated, over the course of processing unit B will become active – satisfying the constraint that both units have the same activation level.

As shown by Smolensky & Legendre (2006a), HGs can be implemented within this type of connectionist network (for implementations of specific HGs, see Legendre *et al*. 2006, Soderstrom *et al*. 2006).[5] The constraints of HG, as well as their relative strengths, are encoded in the weights of these networks. Our analyses assume that HG constraints reflect the structure of 'classical' OT (Prince & Smolensky 1993) which consists of faithfulness constraints (which refer solely to the relationship between the underlying and surface form) and markedness constraints (which refer solely to properties of the surface form). We therefore assume that the HG is instantiated by connections between units representing the underlying and surface form (encoding faithfulness) as well as between units representing the surface form (encoding markedness).[6]

To concretely illustrate the translation of HGs to connectionist networks, consider the HG fragment illustrated in tableau (3). This makes use of the constraints defined above, but focuses on a single feature of the underlying and surface representations (voicing).

(3) *HG tableau illustrating a grammar fragment*

weight	2	1	$\mathcal{H}$
/−voi/	IDENT[voi]	*LAR	
☞ a. [−voi]			0
b. [+voi]	−2	−1	−3

[5] Note that for finite networks, the maximum size of representations must be fixed. In connectionist networks, symbolic structures (such as the input and output of an HG) are represented as activation patterns over simple processing units (Smolensky & Legendre 2006). A finite set of units can faithfully represent only a finite number of distinct patterns.

[6] In addition to simple pairwise connections among units, complex constraints may require additional 'hidden' units that can compute more complex functions (see Soderstrom *et al*. 2006 for examples of several such constraints). For example, the Hopfield networks depicted in Figs 1–5 contain a bias unit that serves to instantiate context-free markedness constraints. In the analyses here, we treat such 'hidden' structures as part of the surface form of the grammar (analogous to metrical structure that is not unambiguously recoverable from the acoustic signal; Tesar 1999).

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

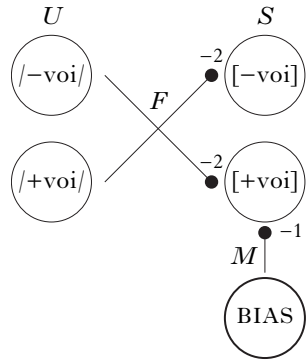


Figure 1
Hopfield network instantiating the grammar fragment illustrated in tableau (3).

We can instantiate this grammar fragment within the Hopfield network shown in Fig. 1 below. Nodes on the left denote elements of the underlying representation (U) while those on the right denote the surface representation (S). These representations are 'distributed' in the sense that a single feature [±voice] corresponds to two distinct units [+voice] and [−voice]. The surface representation also contains a bias unit which is always active (note that active units are shown with thick lines; inactive units are depicted with thin lines). Two negative weights (with strength −2) connect each underlying unit to the surface unit with the opposite feature specification. These weights (denoted by F) instantiate the faithfulness constraint IDENT[voice]; as discussed above, a negative weight signifies that the two connected units should not take on similar values (e.g. when /−voice/ is active, [+voice] should be inactive). The negative weight from the bias unit to the surface feature [+voice] (denoted by M) instantiates the markedness constraint *LAR. Since the bias unit is always active, this weight signifies that regardless of the input, the [+voice] representational unit should not be active. Note that although we assume the representations are fully interconnected, weights with strength 0 (e.g. connecting /+voice/ to [+voice] or the bias unit to [−voice]) are not depicted.

The Hopfield network computes the underlying → surface mapping specified by the HG through updating the activation of representational units. The activations of units corresponding to the underlying form are fixed (or 'clamped') and the activations of units representing the surface form are updated in such a way as to maximally satisfy the constraints encoded by the network.[7] Specifically, as shown by Smolensky (1986,

[7] The algorithm of Hopfield (1982) is only guaranteed to find local maxima (the 'nearest' forms that best satisfy the constraints of the HG). However, closely

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

2006a), a numerical measure (HARMONY) can be defined which characterises the degree to which different activation patterns satisfy the constraints of these networks. The algorithms that update activation values serve to maximise this numerical measure. The end result is that the activation of the output units corresponds to the most harmonic surface form for the given input. As the output of the HG is the surface form that maximally satisfies the constraints of the grammar, the input–output pairs computed by the Hopfield network are therefore identical to the input–output pairs specified by the HG.

Figure 2 illustrates harmony values for the network depicted in Fig. 1. The upper panel shows harmony values for a voiceless input (i.e. where /−voice/ is active, as shown by the thickened lines for this units, and [+voice] is inactive), while the lower panel shows the corresponding values for a voiced input. The left column depicts the harmony value of a voiceless output while the right depicts a voiced output. The harmony calculation is shown within a text box in each column.[8] Harmony is simply the sum, over all weights, of the product of the activation values of the two units and the weight (Smolensky 1986, 2006a). For example, in the right column of the upper panel, /−voice/ and [+voice] both have activation of 1. They are connected by a weight of −2. This weights contribution to harmony is therefore $1 \times -2 \times 1 = -2$.

Comparison of the harmony values across columns shows that this yields the same underlying–surface mapping as the HG illustrated above. For a voiceless input, a voiceless output has higher harmony (0 *vs.* −3); for a voiced input, a voiced output has higher harmony (−1 *vs.* −2). Note the asymmetry across inputs in the harmony difference between the correct output and its competitor (the HARMONY ADVANTAGE);[9] this difference becomes critical in the next section.

4.2 Disruptions to connectionist processing cause a markedness asymmetry in errors

4.2.1 *Disruptions to connectionist processing yield the full range of error types.* To disrupt processing within the Hopfield network, we assume that random noise is added to the connections that instantiate constraints.[10] This noise varies randomly from trial to trial within an

related computational architectures (that mimic the essential features of the Hopfield network) are guaranteed to find the global maximum (the best form overall; Aarts & Korst 1989). For our analysis, we merely assume that the connectionist processing mechanism is sufficiently powerful enough to find the global maximum specified by the network constraints.

[8] In the examples discussed here, we have excluded the harmony contribution of the symmetric weights (e.g. the reciprocal weight connecting [+voice] to BIAS, which has the same value as the BIAS to [+voice] weight).

[9] This is equivalent to Boersma & Pater's (2008) 'margin of separation of harmony', a critical component of their theory of HG learning.

[10] See Bíró (2006) for an alternative approach to disrupting grammar processing. We believe this approach is not general enough to account for the full range of speech

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

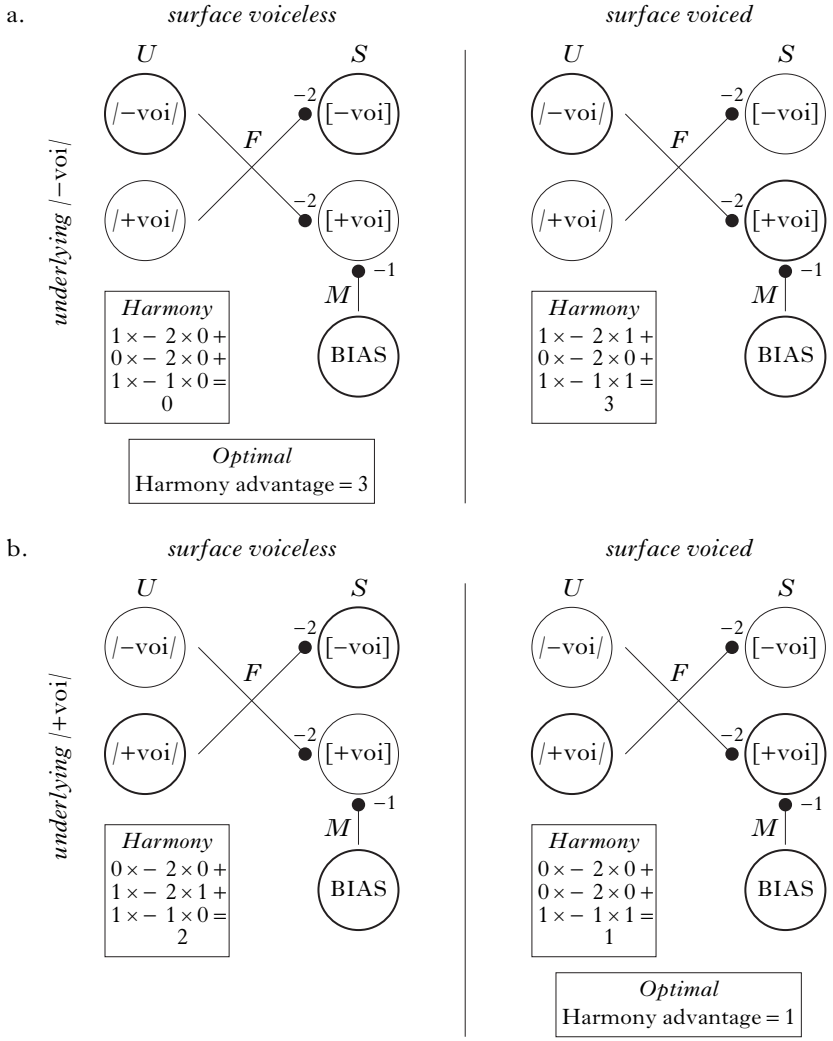


Figure 2
Harmony in the Hopfield network instantiating the HG fragment.

individual. This is a common means of disrupting connectionist processing to simulate speech errors (e.g. Dell *et al.* 1997). Within a Hopfield network, the consequence of such random noise will be to distort the

errors. Bíró's account predicts that errors violating harmonic ascent will only occur when forms violating *vs.* respecting harmonic ascent are distant in representational space. As shown by the errors in §4.3.3, violations of harmonic ascent can occur when the alternating forms differ by only a single feature (i.e. the anti-harmonic form is not a local maximum).

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

preferences of the constraints encoded by the undisrupted network. The relative harmony of surface forms is therefore altered. In some cases, these disruptions will be sufficient to cause the target form to be less harmonic than some competitor. Because network processing always maximises harmony, this results in an error. In general, if disruption serves to make the target less harmonic than a competitor, the competitor will be the output.

Being blind to the nature of constraints, the disruptions caused by random noise will not be constrained to produce only those mappings that can specified by the grammar. This is critical, because it will allow for the possibility of violations of harmonic ascent. This is because random disruptions will not only distort the relative strength of various constraints; they can also result in disruptions to the constraints themselves.

To illustrate the wide range of possible mappings that disruption can produce, we will consider two instances of disruption to single weights in the Hopfield network above. Figure 3 illustrates a case where disruption has changed the strength of the weight connecting the bias unit to [+voice] from −1 to −3 (a weight change of −2). This has a transparent interpretation in HG terms; the weighting of *LAR has increased to 3 from 1. Since *LAR now dominates IDENT[voice], we would expect devoicing to occur. As shown in the figure, this is what occurs; the voiceless input is correct, but the voiced output yields a devoicing error.

However, since random noise is blind to the structure and content of constraints, it can also yield underlying–surface mappings that would not be produced by a typical HG. This is illustrated in Fig. 4. Here, noise has again disrupted a single weight in the network – in this instance, the weight linking underlying /−voice/ to surface [+voice]. Our assumption is that noise is randomly and evenly distributed. This allows for weight values to decrease (as shown above) as well as increase. The latter case is depicted here; instead of making the weight more negative, noise has flipped its sign (a weight change δ of +4). This has the effect of converting a faithfulness constraint into an *anti*-faithfulness constraint. This weight now signifies that when /−voice/ is active, [+voice] should also be active – in other words, underlying voiceless should be mapped to surface voiced. As shown in Fig. 4a, this yields a mapping violating harmonic ascent. This stands in contrast to other mechanisms of producing variation in HGs. Typically, HG theories restrict constraint weightings to a particular range (e.g. Potts *et al*. 2008 restrict constraints to positive weightings; see Boersma & Pater 2008 for discussion). Even if such restrictions hold for undisrupted networks, the 'blind' nature of disruption at the level of individual weights can cause violations of such restrictions.

Furthermore, since disruption occurs at the level of individual weights, it need not uniformly target a constraint. For example, since in this network the feature [voice] is realised in a distributed fashion (over two representational units), the constraint IDENT[voice] is instantiated by two separate weights. Disruption may not affect all of these weights equally. In this case, the weight instantiating IDENT[voice] for voiced inputs is

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

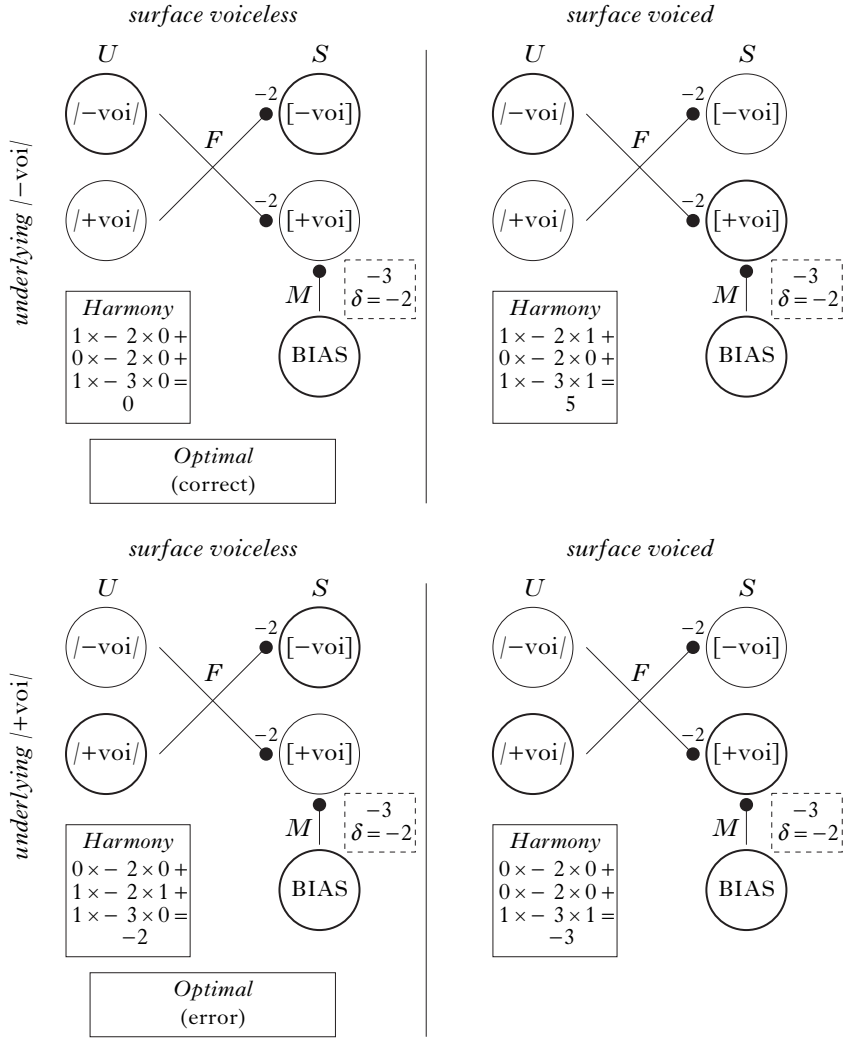


Figure 3
Disruption to the Hopfield network yielding a devoicing error. Disrupted weight is shown by dotted lines.

unaffected; therefore, there is no error on a voiced target (as shown in Fig. 5b). In essence, disruption has 'split' IDENT [voice] into two constraints – an anti-faithfulness constraint specific to voiceless inputs and a standard faithfulness constraint specific to voiced inputs.[11] This

[11] It should be noted that we are not aware of any empirical data documenting such a pattern.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

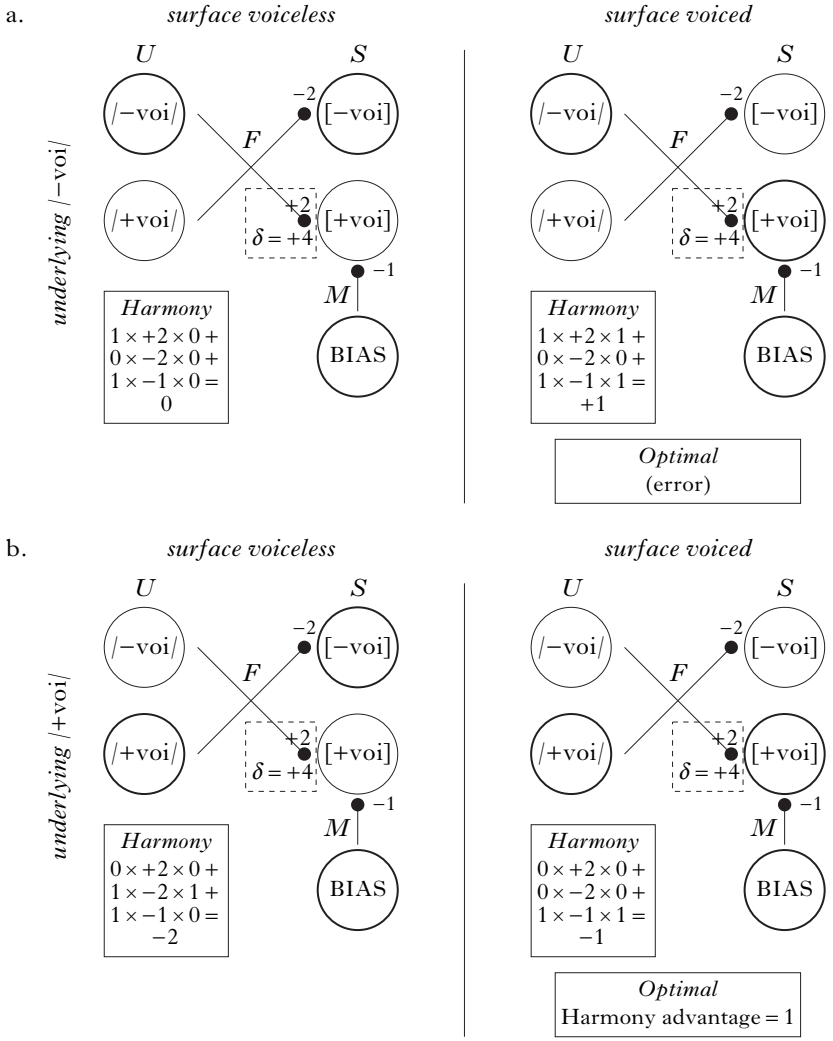


Figure 4
Disruption to the Hopfield network yielding a voicing error. Disrupted weight is shown by dotted lines.

stands in contrast to other mechanisms for producing variation in HG. These typically allow random variation only in the relative weighting of different constraints (Boersma & Pater 2008).

4.2.2 *Disruptions probabilistically respect markedness asymmetries.* The blind nature of random noise allows it to produce errors that radically

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

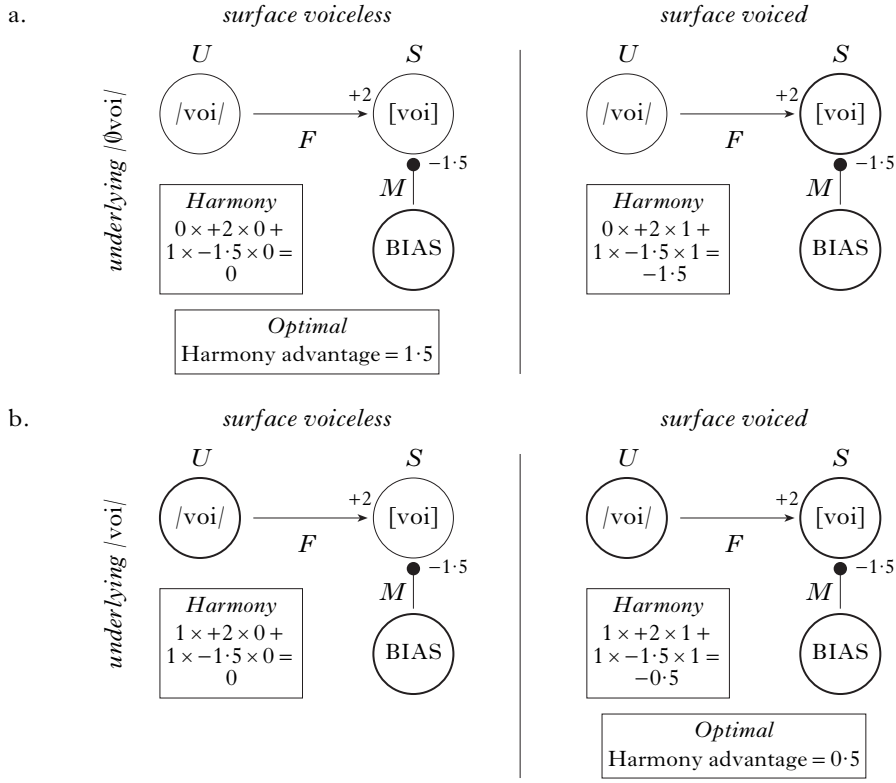


Figure 5
Smaller Hopfield network, instantiating HG with a positive faithfulness constraint and a privative voicing feature.

depart from the markedness preferences of the intact network/HG. However, just because the noise itself is blind does not entail that it favours all outcomes equally. Errors that respect the markedness preferences of the intact network – i.e. errors respecting harmonic ascent – will arise more easily than errors that do not. This is because the amount of disruption required to produce an error is a function of its harmony relative to the correct target form. If the error is unmarked along many dimensions (e.g. it satisfies markedness constraints violated by the target), then its harmony will be relatively close to the target; the harmony advantage of the target will be relatively small. A relatively small amount of random disruption to the harmony function will be sufficient for the harmony of this error to surpass that of the target. In contrast, an error that is marked along most (or all) dimensions will have a harmony far less than that of the target; a large amount of disruption will be required to produce it as an error.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

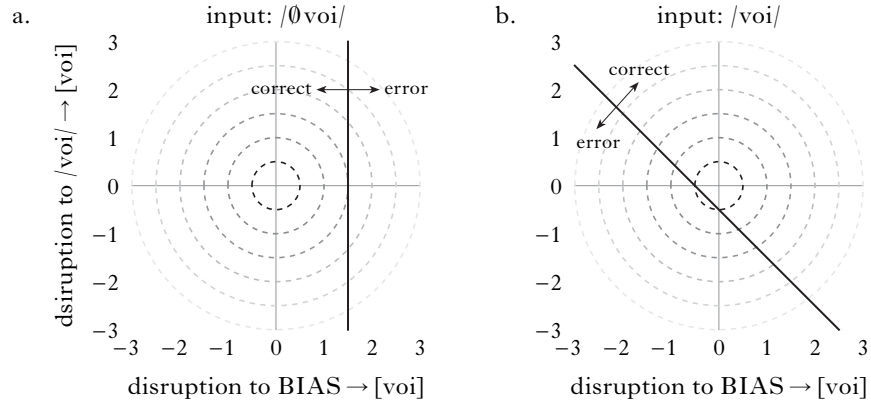


Figure 6
Consequences of disruption to weights of the smaller Hopfield network (see Fig. 5).

A specific instance of this asymmetry can be seen in the instances of disruption discussed above. In Fig. 1, we see that the harmony advantage of a /−voice/ input is larger (3) than a /+voice/ input (1). Comparing Figs 3 and 4, we can see that the amount of disruption yielding a devoicing error is smaller (two units on a single weight) than the amount of disruption yielding a voicing error (four units on a single weight).

This asymmetry is not specific to these two instances of disruption, but holds more generally. This is difficult to illustrate with the network discussed above; given that there are more than two weights in the network, it is difficult to visualise the effects of disruption. Instead consider the Hopfield network shown in Fig. 5. This is distinguished from the previous example in two ways. It makes use of a privative specification of voicing (i.e. [voice] active represents voiced, [voice] inactive represents voiceless). Second, violating the convention typically adopted in both OT and HG, we make use of a faithfulness constraint that rewards satisfaction rather than penalising deviation. A positive weight connects the underlying /voice/ feature to the surface [voice] feature, signifying that both units should take on similar activation values. The weighting adopted here produces the same underlying–surface mappings as in the network above (i.e. underlying voicing is faithfully preserved in the intact network) and exhibits a qualitatively similar asymmetry in harmony advantages (i.e. the voiceless target has a harmony advantage greater than that of the voiced target).

Figure 6 depicts the consequences of disruption to the weights in the network above. The *x*-axis represents disruption to the weight instantiating the markedness constraint; the *y*-axis represents disruption to the weight instantiating the faithfulness constraint. The thick line in each

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

panel shows the boundary between correct and incorrect outputs for varying degrees of disruption to both weights. As shown in (a), for a voiceless input errors occur when the weight connecting the bias unit to the surface [voice] feature becomes positive. This corresponds to a flip in the sign of the constraint – changing it from a markedness constraint that disprefers voiced outcomes to an anti-markedness constraint which *prefers* such structures.[12] As shown in (b), errors for voiced inputs are influenced by two factors. First, if the faithfulness constraint is weakened (i.e. negative changes in the weight), the influence of the markedness constraint will yield voiceless outcomes. Similarly, if the markedness constraint becomes more powerful (i.e. negative changes in the weight), it will tend to overpower the influence of faithfulness.

As noted above, we assume that disruption is randomly and evenly distributed along each weight. This is depicted by the concentric circles in each figure (where the centre depicts the intact network). The centremost circle contains the smallest disruptions, which have the highest probability. The ever-lighter circles surrounding it indicate decreasing probability as the level of disruption increases. Inspection of the figures reveals that it is easier for disruption to yield a devoicing error than a voicing error. Consider the region denoted by the two innermost circles. This entire area falls within the correct region for a voiceless input; in contrast, a significant portion of this area of disruption falls outside the correct region for a voiced input. As this asymmetry holds for all levels of disruption, errors for voiced inputs are more likely than errors for voiceless inputs. In general, then, if disruption is equally and randomly distributed, it will be more likely to produce an error respecting *vs.* violating harmonic ascent – matching the broad empirical patterns reviewed above.

The Appendix contains a proof showing that this asymmetry holds more generally for disruption to Hopfield networks realising HGs. We specifically compare the probability of two errors that are equal distortions of the target (e.g. /+voice/ → [−voice] and /−voice/ → [+voice] both involve the change of one feature value). One of these error outcomes better respects the markedness preferences of the intact network/HG; this error outcome is harmonically closer to the target than the other error outcome (e.g. /+voice/ → [−voice]). We then show that if an HG is instantiated within a Hopfield network, and noise is randomly and evenly distributed across all connections within the network (as in the concentric circles shown in Fig. 6 above), the probability of an error is a function of the harmony advantage of the target. This yields the desired result; errors that better respect the markedness preferences of the intact HG (e.g. /+voice/ → [−voice]) are more likely than those that do not (e.g. /−voice/ → [+voice]).

[12] Since for this case the input unit is inactive, the harmony contribution of the faithfulness constraint is always 0.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

Unlike many purely grammatical accounts of variation, our proposed mechanism for modelling speech errors allows for the production of errors that violate harmonic ascent. However, this does not mean that errors completely disrespect harmonic differences; they are biased to result in less *vs.* more marked outcomes. This allows our proposal to match the broad empirical patterns identified above. Note that this holds even though disruption is completely blind to the structure and content of grammatical constraints. A process that operates solely at the level of 'implementation' of the grammar can nonetheless exhibit sensitivity to grammatical distinctions.

4.3 Linking grammatical theories and speech-error distributions

Although our proposal provably accounts for the general markedness-asymmetry pattern, a more concrete link between particular grammatical theories and predicted speech-error distributions would be desirable. Unfortunately, the method utilised in the Appendix for precisely characterising the speech-error distribution is computationally and analytically quite difficult. In supplementary online materials we outline more efficient methods and provide simulation implementations for use in modelling speech-error distributions.[13]

4.3.1 *An illustration: initial consonant errors in English.* To illustrate how grammatical theories and speech-error distributions can be linked, we consider initial singleton consonants in English. Goldrick (2002) used a tongue-twister task to induce speech errors involving six pairs of such consonants. These consonant pairs differed by a single distinctive feature, and were selected such that one member of the pair was less marked than the other (with respect to both the appropriate statistics of English as well as cross-linguistic markedness distinctions). For example, [d] and [z] differ with respect to the feature [continuant]; [d] is both more frequent and less marked than [z].

The average distribution of errors across all pairs (see (6) below) fit the empirical generalisation above (e.g. errors were more likely to result in less marked forms, although both types of errors occurred). However, it should be noted that several individual pairs deviated from this overall pattern. The tongue-twister task – analysed through broad transcription – is not an ideal source of data for tapping exclusively into phonetic encoding processes. It causes disruptions at multiple levels of sound-structure processing (e.g. motor/articulatory processes; see note 1 for discussion). Errors arising at these other levels may distort the distribution of errors arising at phonetic encoding. However, since many of the errors in this task arise during phonetic encoding, we believed these data were sufficient to illustrate the approach outlined above.

[13] Available at http://journals.cambridge.org/issue_Phonology/Vol26No01.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

To model these data we constructed a simple HG fragment. With respect to representations, we utilised three distinctive features to distinguish among the consonants in this study: [voice], [continuant] and [place] (which was associated with one of three privative specifications). The features used for the eight consonants in this study are shown in (4) (notice that there are not twelve consonants in total, as some consonants were repeated across pairs). The underlying and surface representations were triples of these features (e.g. /z/: [+voice], [+continuant], [coronal]).

(4) *Features of the simple Harmonic Grammar*

	[voice]	[cont]	[labial]	[dorsal]	[coronal]
[p]	−	−	+		
[b]	+	−	+		
[k]	−	−		+	
[g]	+	−		+	
[t]	−	−			+
[d]	+	−			+
[s]	−	+			+
[z]	+	+			+

There were seven markedness constraints of the HG which referred to the seven values of the three features (two each for [voice] and [continuant], plus three for place; see (5) below). Three faithfulness constraints compelled identity between the feature values of each of the three features (e.g. IDENT[voice]).

As all of these consonants surface faithfully as initial singleton onsets, the faithfulness constraints dominate all of the markedness constraints (we assumed they all had an equal weighting at the top of the hierarchy). To determine the relative ranking of the markedness constraints, we assumed that well-formedness is proportional to log frequency. We calculated the log frequency of each feature value by summing the type frequencies for each singleton English onset, using the frequency counts and feature matrices of Hayes & Wilson (2008).[14] As the relative ranking of markedness constraints specifies *ill*-formedness, we assumed that their ranking would be proportional to the inverse of log frequency. The specific weightings were assigned by arbitrarily setting the faithfulness constraints to ten, and each markedness constraint to ten times the inverse of the log frequency of the relevant feature. The resulting weightings for the markedness constraints are shown in (5).

[14] Note that Hayes & Wilson (2008) assume both contrastive and privative underspecification; we assumed that only segments that are specified for some feature value contribute to its frequency (e.g. the nasals did not contribute to the frequency of [+voice]).

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

(5) *HG markedness constraints and weightings for English initial singleton onset grammar*

*[dorsal]	2·84	*[−cont]	2·49
*[+voice]	2·65	*[−voice]	2·47
*[+cont]	2·61	*[coronal]	2·40
*[labial]	2·56		

The harmony differences between target and error for each of the six experimental pairs was then calculated. We assumed that disruption to the connections within the network implementing this HG were independently and identically distributed according to a normal distribution. Using the procedure described in the online supplemental materials, we estimated the standard deviation of this normal distribution for each individual error.[15] We then averaged these estimated standard deviations to obtain an overall estimate of the standard deviation of the disruption distribution (specifically, 1·66). This average standard deviation was then used to generate the predicted rate of error for each consonant pair (using the Monte Carlo procedure described in the online supplementary materials, with 100,000 iterations per pair).

As shown in Table I, the HG was able to capture the average pattern quite closely. In the actual data, on average errors that resulted in more marked structures (e.g. /d/ → [z]) occurred at a rate of 4·9%; errors resulting in less marked structures occurred at a higher rate, 5·9%. This illustrates the markedness asymmetry discussed above. On average, the simulated error distribution matched this quite closely (deviating 0·1% in each case). Although the average fits across all pairs were quite good, the model's predictions were quite far off on individual pairs; we return to this below.

Of course, since the parameter for generating errors (the standard deviation of the disruption function) was selected by averaging the parameter estimates for each error pair, it is not surprising that the mean predictions of the model track the mean error rates quite closely. What is more surprising is the ability of this HG to match the average error rates within each error category. This is not guaranteed to occur when using the procedure outlined above. To illustrate this, we constructed a second HG where the weights of the markedness constraints were proportional to the inverse of raw type frequency (not log frequency).[16] Utilising the same procedure outlined above, on average the model predicted a rate of 3·8% for errors resulting in more marked structures *vs.* 7·5% for errors resulting in less marked structures. Note the average of these two rates (5·7%) is

[15] The number of non-zero representational elements in the underlying representation was three (i.e. the number of distinctive features); for both the target and error, the number of distinct elements in the surface representation was one (all pairs were distinguished by a single feature) and the number of common elements was one (assuming that a bias unit is used to instantiate the context-free markedness constraints).

[16] Specifically, $10{,}000 \times 1/\text{frequency of each feature}$.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

error outcome: marked	observed	predicted	error outcome: unmarked	observed	predicted
/t/ → [d]	5·5%	5·0%	/d/ → [t]	5·2%	5·7%
/t/ → [s]	5·4%	5·1%	/s/ → [t]	6·7%	5·7%
/p/ → [k]	4·6%	4·9%	/k/ → [p]	3·8%	5·8%
/k/ → [g]	5·5%	5·1%	/g/ → [k]	7·6%	5·7%
/d/ → [z]	3·7%	5·1%	/z/ → [d]	8·2%	5·6%
/b/ → [g]	4·7%	4·8%	/g/ → [b]	4·1%	5·9%
mean	4·9%	5·0%	mean	5·9%	5·8%

Table I

Modelling of initial consonant errors in English. 'Observed' refers to average rate of errors by participants in Experiment 2 in Goldrick (2002).

quite close to the average of the observed error rates (5·4%). This is simply a by-product of the parameter-selection procedure. What the procedure cannot do is guarantee that the model will make the correct predictions for particular classes of errors; such matches occur only if the grammar has the appropriate architecture. In this case, the poor performance of this latter model suggests that, as in many other domains, log frequency, not raw frequency, better reflects distinctions in cognitive processing. Specifically, raw frequency overestimates the well-formedness of frequent structures.

Although we believe the Harmonic Grammar outlined above captures some aspects of English speakers' knowledge of sound structure, it is worth underscoring that the model is clearly inadequate as a comprehensive theory of English phonological knowledge. It is restricted to single consonants that can be distinguished by three distinctive features. This is woefully impoverished with respect to the consonantal inventory of English (much less the full complexities of English phonology). The space of constraints is also far too limited. As noted by a reviewer, it is well known that place and voice interact in determining markedness (Gamkrelidze 1978). Constraints must refer to feature conjunctions to encode such interactions. The poor fit of the model to error rates on individual consonants may reflect some of these simplifications. (Alternatively, as noted above, deviations may reflect the contamination of the data from errors arising in processes other than phonetic encoding.) However, the fact that such a simple model is capable of a capturing general trends in these data suggests that this approach holds promise.

4.3.2 *Extensions to the model of disruption.* It should be noted that all of the methods analysed above have assumed that errors are generated by evenly distributed disruption to the weights encoding the HG. It is likely

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

that this assumption will not provide a fully general account of all types of speech errors. In particular, we believe a full model must allow for the possibility that disruption targets particular dimensions of phonological representations. For example, Goldrick & Rapp (2007) and Buchwald *et al.* (2007) report two cases of acquired impairment to phonetic encoding processes. The former individual has a somewhat higher overall segment accuracy (roughly 95% of segments produced correctly, compared to 83% correct for the individual reported in Buchwald *et al.*). This slight difference in overall accuracy was enormously magnified for clusters. The former individual produced 91% of clusters correctly, whereas the latter was only 28% correct. This particular difficulty with this type of phonological structure may reflect disruption that preferentially affects those network mechanisms involved in the representation of clusters. In future work, we hope to compare and contrast the predictions of evenly *vs.* non-evenly disrupted models of disruption.

We have also assumed that errors arising in phonetic encoding are induced solely by disruptions to phonetic encoding mechanisms. While we believe this to be an appropriate model of errors arising in single-word tasks, it is not clear if it generalises to tasks where errors are believed to be induced by the priming of competitors. For example, as pointed out by a reviewer, tongue twisters induce errors via the production of similar sounds (in similar environments) in close temporal proximity to target sounds. Future work should examine how our approach can be extended to model such effects.

5 Conclusions

Characterising markedness distinctions among phonological structures is a central concern of phonological grammars. Although grammatical theories have been primarily motivated by internal evidence, a wealth of studies suggest that markedness influences online behaviour. Specifically, speech errors are biased to result in less *vs.* more marked structures. To account for this markedness asymmetry in errors, we propose that speech errors are generated by random disruptions to the connectionist mechanisms that encode the constraints of the grammar. This method provably accounts for the markedness asymmetry in speech-error distributions. Furthermore, by making certain assumptions about the distribution of random disruptions to connectionist mechanisms, we can link specific grammatical theories with speech-error data (as illustrated by the application to initial consonant errors in English).

There are numerous benefits to establishing tighter connections between phonological grammars and online speech behaviour. Theories of language processing can greatly benefit from the formal insights of grammatical theories. For example, many processing theories adopt overly simplistic notions of markedness (e.g. position-specific mono- and bi-phone frequency; Vitevitch & Luce 1999). Other accounts rely on trained connectionist networks; the internal structure of such networks is

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

notoriously opaque (see e.g. Goldrick 2007 for discussion in the context of speech-error research). Grammatical theories often an attractive third path; they are complex, but analytically transparent. They may therefore provide a productive framework for illuminating the complex internal structure of phonological processing mechanisms. With respect to grammar, establishing links to speech-error data allows for a richer empirical base for theory development and testing. More generally, work of this sort is a critical part of a larger programme of research that aims to make grammars a central component of theories of language acquisition and processing (e.g. Smolensky *et al*. 2006).

Does this approach offer insight to phonological behaviour beyond speech errors? As noted above, a number of grammatical mechanisms have been proposed within OT and HG to account for synchronic variation. In accounting for speech errors (a type of probabilistic behaviour), we have pursued a novel approach (stochastic disruption to connectionist mechanisms that encode the grammar). Having established that our approach can account for speech errors, we can ask if it is a fully general mechanism for accounting for other types of variation. Specifically, can stochastic disruption account not just for speech errors but also synchronic variation? We suspect this is not generally true. As noted above, available data suggest that synchronic variation respects harmonic ascent. Since stochastic disruption will regularly violate this principle, it is likely that this mechanism is too unrestrictive for a general theory of grammatical variation. However, we consider it plausible that stochastic disruption may provide a better account of performance in online speech production tasks than purely grammatical mechanisms – a possibility that should be explored in future work.

If our theories adopt multiple mechanisms for producing variation (e.g. floating constraints as well as stochastic disruption), a natural question that arises is how to determine which mechanism is responsible for any particular case of variation we observe empirically. One criterion may be stability across social groups. Variationist phonological research has repeatedly emphasised that such variation is specific to particular social groups and must be acquired by its members (e.g. Roberts 1997). Since processing mechanisms are presumably roughly equally distributed across such groups, the existence of contrasting patterns of variation across these populations suggests that these patterns do not reflect properties of processing mechanisms (e.g. stochastic disruption). Alternatively, if some pattern of variation is influenced by increasing memory load or attentional demands while holding social factors constant, we may conclude that this pattern is at least partially influenced by the properties of processing mechanisms. Of course, as with any behavioural data, it is likely that variation reflects multiple, interacting mechanisms. For example, changes in constraint rankings (and consequent variation in outputs) might reflect both floating constraints as well as stochastic noise on connection weights. The challenge for future work is to examine how these multiple mechanisms interact to produce complex patterns of variation.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

Appendix

The primary goal of this appendix is to develop the theoretical and mathematical machinery to account for speech errors using the Harmonic Grammar formalism. The basic idea is to add noise to the weight matrix that instantiates the mapping from underlying to surface representations.

There are a number of formal proofs, interspersed with brief prose descriptions. The prose descriptions are intended to guide the reader through the general logic of the proofs, including a prose description of what is to be proved, an explanation of why it needs to be proved, and an overview of the logic of the proof. Although the notation is at times baroque, most terms correspond closely with standard concepts of generative phonology. The first original result is Theorem 2, which gives a formula to calculate the probability of a speech error for an arbitrary Harmonic Grammar. The final destination is Corollary 3, which accounts for the markedness asymmetry in speech errors (e.g. [g] is more marked than [k], so /k/ → [g] errors are less frequent than /g/ → [k] errors). In addition, we prove that every Harmonic Grammar can be embedded in a Hopfield network (Hopfield 1982), a result first obtained by Smolensky (1986). The result is re-proved here to ensure that Theorem 2 and Corollary 3 still apply when a grammar is Hopfield-embedded.

This appendix is organised into four sections: §1 lays out the formal foundations of harmonic grammars, §2 defines speech errors as the outcome of additive noise, §3 introduces geometry to handle high-dimensional noise and §4 calculates error probabilities.

1 Foundations

This section gives formal definitions for standard concepts of generative phonology. Of especial interest is Theorem 1, which shows how any Harmonic Grammar can be embedded in a Hopfield network.

Definition 1: A harmonic grammar G consists of a set of underlying representations $U \in \{0, 1\}^m$, a set of surface representations $S \in \{0, 1\}^n$, a set of faithfulness constraints $F \in \Re^{m \times n}$ and a set of symmetric markedness constraints $M \in \Re^{n \times n}$ satisfying $M_{ij} = M_{ji}$ and $M_{ii} = 0$. The associated harmony function $H^G : U \times S \Rightarrow \Re$ (Smolensky 1986, 2006a) is defined by:

$$H^G(x, y) = xFy^t + yMy^t$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} F_{ij} \cdot x_i y_j + \sum_{i=1}^{m}\sum_{j=1}^{n} M_{ij} \cdot y_i y_j$$

It will prove convenient to separate the representational terms from the grammatical terms, so define

$$G = \begin{pmatrix} F \\ M \end{pmatrix},\ R(x, y) = \begin{pmatrix} x^t y \\ y^t y \end{pmatrix}$$

so that

$$H^G(x, y) = G \cdot R(x, y)$$

where $\cdot$ is the matrix dot product $A \cdot B = \sum_i \sum_j A_{ij} \cdot B_{ij}$.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

Definition 2: A surface representation $y \in S$ maximises the harmony given an underlying representation $x \in U$ if $H^G(x, y) \geq H^G(x, z)\ \forall z \in S$. The grammar G maps x to y, written $G(x) = y$, if there is a unique harmony-maximising surface y representation for x.

Theorem 1: A harmonic grammar G can be instantiated in a Hopfield network N realising the same mapping as G.

Proof: Let

$$W = \begin{pmatrix} 0 & F^t \\ F^t & M + M^t \end{pmatrix}$$

be the weight matrix for a Hopfield network N and $\forall x_i \in U$, define $y_i \equiv G(x_i)$, $z_i \equiv x_i \oplus y_i$. By definition, the harmony function of N is $H^W(z) = \frac{1}{2} z W z^t$. Then $H^W(z_i) = H^W(x_i \oplus y_i)$

$$= \tfrac{1}{2}(x_i\ y_i) \begin{pmatrix} 0 & F^t \\ F^t & M + M^t \end{pmatrix} \begin{pmatrix} x_i^t \\ y_i^t \end{pmatrix}$$

$$= \tfrac{1}{2}(x_i F y_i^t + y_i F^t x_i^t + y_i M y_i^t + y_i M^t y_i^t)$$

$$= \tfrac{1}{2}(x_i F y_i^t + (x_i F y_i^t)^t + y_i M y_i^t + (y_i M y_i^t)^t)$$

$$= \tfrac{1}{2}(2 x_i F y_i^t + 2 y_i M y_i^t)$$

$$= x_i F y_i^t + y_i M y_i^t$$

$$= H^G(x_i, y_i)$$

Because W is symmetric, the network is guaranteed to converge to a harmony maximum when some input pattern is clamped (Hopfield 1982; see also note 6). In particular, whenever the first m units are clamped to x_i, N will converge to the vector $z = x_i \oplus y$ with highest harmony. Since by definition $y_i = G(x_i)$ maximises the harmony given x, the remaining n units of the network will converge to $G(x_i)$. Thus N instantiates the same mapping as G.

We regard the Hopfield network as the neural mechanism instantiating a Harmonic Grammar. Thus, our goal in re-proving Theorem 1 is to assure ourselves that the Hopfield embedding of a Harmonic Grammar does not impose additional assumptions that are incompatible with the speech-error theory developed below. In other words, we want to be sure that the two behave the same under noise.

There are two potential issues that stem from the convergence properties of Hopfield networks. First, convergence is not guaranteed if the network contains 'self-connections', i.e. if there are non-zero weights from units to themselves. This is why we require $M_{ii} = 0$ in Definition 1. Second, convergence is not guaranteed if the weight matrix is asymmetric. This condition is enforced in Theorem 1 by making the bottom right block of W equal to the sum of M and M^t, so Theorem 1 applies even if $M_{ij} \neq M_{ji}$. We nonetheless require M to be symmetric because the theory of speech errors developed below is more mathematically tractable if the connection between surface representation units i and j is regarded as a single entity. This has implications for the way that noise is introduced into the weight matrix, so we will return to this point below when stochastic disruption is introduced.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

2 Speech errors

The first fundamental concept introduced in this section is the HARMONY ADVANTAGE of one surface representation over another, for a given input (see Fig. 2 and accompanying text). For example, the mapping /k/ → [k] is more harmonic than the unfaithful mapping /k/ → [g], so relative to the input /k/, [k] has a positive harmony advantage over [g]. The second concept introduced here is DISRUPTION. The term disruption is used to refer to additive noise that affects the constraint weights of the grammar.

The fundamental insight formalised in this section is that a speech error occurs when disruption overcomes the harmony advantage of the normal output candidate (which would otherwise be selected in the undisrupted grammar). That is, /k/ maps to [g] when the grammar is disrupted in a way that renders the harmony advantage of [k] over [g] negative.

Definition 3: For an underlying representation x, the harmony advantage of surface representation y over z is

$$\Delta H_x^G(y, z) = H^G(x, y) - H^G(x, z)$$

As before, it will prove useful to separate the representational and grammatical components, so define

$$\Delta R_x(y, z) = R(x, y) - R(x, z)$$

so that

$$\Delta H_x^G(y, z) = G \cdot \Delta R_x^G(y, z)$$

Finally, define the normalised representational difference and normalised harmony advantage

$$\Delta \hat{R}_x(y, z) = \frac{\Delta R_x(y, z)}{|\Delta R_x(y, z)|}$$

$$\Delta \hat{H}_x^G(y, z) = G \cdot \Delta \hat{R}_x(y, z)$$

Definition 4: The effect of stochastic disruption

$$\delta G = \begin{pmatrix} F \\ M \end{pmatrix}$$

on a grammar is linear (see Figs 3 and 4). That is, the harmony function of the (disrupted) grammar $G + \delta G$ is $H^{G+\delta G}(x, y) = (G + \delta G) \cdot R(x, y)$.

Corollary 1: $\Delta H_x^{G+\delta G}(y, z) = \Delta H_x^G(y, z) + \Delta H_x^{\delta G}(y, z)$

Proof: $\Delta H_x^{G+\delta G}(y, z) = H^{G+\delta G}(x, y) - H^{G+\delta G}(x, z) = H^G(x, y) + H^{\delta G}(x, y) - H^G(x, z) - H^{\delta G}(x, z) = H^G(x, y) - H^G(x, z) + H^{\delta G}(x, y) - H^{\delta G}(x, z) = \Delta H_x^G(y, z) + \Delta H_x^{\delta G}(y, z)$.

Definition 5: Disruption δG causes an $x \xrightarrow{y} z$ error if $G(x) = y$ but $(G + \delta G)(x) = z \neq y$.

Corollary 2: δG causes an $x \xrightarrow{y} z$ error if and only if $\Delta H_x^{\delta G}(y, z) < -\Delta H_x^G(y, z)$.

Proof: δG causes an $x \xrightarrow{y} z$ error if and only if

$$\Delta H_x^{G+\delta G}(y, z) < 0$$

$$\Delta H_x^G(y, z) + \Delta H_x^{\delta G}(y, z) < 0$$

$$\Delta H_x^G(y, z) < -\Delta H_x^{\delta G}(y, z)$$

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

To return briefly to the issue raised in the previous section, we assume that δM obeys the same constraints as M, i.e. the symmetry ($\delta M_{ij} = \delta M_{ji}$) and zero-diagonal ($\delta M_{ii} = 0$) constraints. It would perhaps be more realistic to assume that each weight is disrupted independently, i.e. that disruption does not respect network symmetry. However, this would necessitate a separate, less mathematically tractable proof of convergence. Thus this symmetry assumption can be regarded as an idealisation, whose practical effect is to restrict the degrees of freedom in δG to $N = mn + (n(n-1)/2)$.

Although it is an idealisation, we do not regard as an especially troubling one. This is because a number of studies have found that asymmetric Hopfield networks usually have similar convergence properties to symmetric ones (e.g. Xu *et al.* 1996). Thus, although there exist asymmetric networks exhibiting non-convergent (oscillatory) behaviour, it seems to be the exception rather the rule; in particular, such a network is unlikely to result from adding a small amount of noise to a network that was completely symmetric to begin with.

3 Stochastic hypergeometry

§2 shows that when the weights of a harmonic grammar are disrupted by additive noise δG, the disruption may or may not cause an error. The general purpose of this section is to introduce the mathematical tools that are necessary to characterise the values that cause an error, and in particular to calculate the probability that disruption causes an error.

Because disruption is instantiated as a high-dimensional random variable, it is necessary to analyse it with the tools of high-dimensional geometry. Since high-dimensional objects are normally impossible to visualise, the reader may find it helpful to refer back to Fig. 6, which illustrates the three-dimensional case. In this figure, the x-axis represents disruption to the connection between the BIAS and [voice], and the y-axis represents disruption to the connection between /voice/ and [voice]. Thus the origin represents zero disruption, and the dashed circles represent disruptions of varying magnitudes. The distance from the origin to the bold line represents the magnitude of disruption that is required to cause an error. The distance to the bold line is larger for the unmarked input, so the probability of a marked error is correspondingly lower. The higher-dimensional case is exactly analogous, except that there are more constraint weights that can be disrupted, and thus more dimensions. This section formalises these ideas probabilistically.

The core tool that is introduced is the function S_r^n, which gives the surface area of an n-dimensional hypersphere of radius r. (This is simply the n-dimensional generalisation of the formula for the circumference of a circle: $S_r^2 = 2\pi r$). The primary result of this section is a formula to calculate the surface area of a SPHERICAL CAP. (Presumably we are not the first to derive this formula, but we were unable to find it in standard reference materials on hypergeometry. Thus we refer to Lemma 1 as a 'result', because we derived it ourselves.)

A spherical cap is obtained by cutting a hypersphere with a hyperplane. The reader may understand this by analogy with a three-dimensional case, namely the globe. Each line of latitude corresponds to cutting the globe with a plane. For example, the Tropic of Cancer is a latitude line in the northern hemisphere that cuts the globe at about 23½°. The area between the Tropic and the North Pole is the spherical cap.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

The core idea in this calculation is to divide up the spherical cap into slices and integrate the 'circumference' of these slices. The key fact is that each slice is itself a sphere of a lower dimension. For example, to continue with the globe analogy, the Tropic of Cancer is not only a slice through a sphere in three dimensions, but it is also a 'sphere' in two dimensions, better known as a circle.

Definition 6: Let B_r^n denote the n-dimensional ball of radius r, and ∂B_r^n denote its boundary, i.e. the hypersphere of radius r in n dimensions, $\partial B_r^n = \{x \in \Re^n | |x| = r\}$. The subscript r may be omitted to indicate the unit hypersphere. Further define the set of points in ∂B_r^n whose first coordinate is fixed to α, $\partial B_r^n(\alpha) = \{x \in \partial B_r^n | x_1 = \alpha\}$. Finally, define the spherical cap $cap_r^n(\alpha) = \cup_{\beta<\alpha} \partial B_r^n(\beta)$.

Definition 7: Let S_r^n be the area of ∂B_r^n, $S_r^n = nC(n)r^{n-1}$, where

$$C(n) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}$$

is the spherical coefficient function. Let $S_r^n(\alpha)$ be the area of $cap_r^n(\alpha)$. Finally, define the rectified arc cosine function

$$\overline{\arccos}(x) = \begin{Bmatrix} \pi & x < -1 \\ \arccos(x) & -1 \le x \le -1 \\ 0 & 1 < x \end{Bmatrix}$$

Lemma 1: $S_r^n(\alpha) = S_r^{n-1} \int_0^{\overline{\arccos}(-\alpha)} \sin^{n-2} \phi \, d\phi$

Proof: To calculate the area of a spherical cap, change to a polar representation. Let ϕ denote the angle from the first axis. Then the spherical cap subtends the range $\phi \in [\overline{\arccos}(\alpha), \pi]$, or equivalently, $\phi \in [0, \overline{\arccos}(-\alpha)]$. Thus this area can be calculated by integrating over the area of the sub-hyperspheres whose first coordinate is fixed to $\cos \phi$, which is itself an $(n-1)$-dimensional hypersphere with radius $r \sin \phi$, $\partial B_r^n(\cos \phi) = \partial B_{r \sin \phi}^{n-1}$. The area of this sub-hypersphere is $S_{r \sin \phi}^{n-1} = (n-1)C(n-1)(r \sin \phi)^{n-2} = (n-1)C(n-1)\, r^{n-2} \sin^{n-2} \phi = S_r^{n-1} \sin^{n-2} \phi$. Integrating, $S_r^n(\alpha) = \int_0^{\overline{\arccos}(-\alpha)} S_r^{n-1} \sin^{n-2} \phi \, d\phi$. Since S_r^{n-1} is constant with respect to ϕ, it can be pulled out of the integral.

Definition 8: A random variable $\delta G \in \Re^n$ is spherically distributed according to μ, written $\delta G \sim sph(\mu, n)$, if the magnitude of δG is distributed according to μ and the direction is uniformly distributed (on the unit hypersphere).

Corollary 3: If $\delta G \sim sph(\mu, n)$ then $\forall \alpha \in \Re \; \forall x \in \Re^n$, $p[\cos(x, \delta G) < \alpha] = S_r^n(\alpha)$.

Proof: $\cos(x, \delta G) = \hat{x} \cdot \delta \hat{G}$, which is simply the component of δG in the direction of x. Because $\delta \hat{G}$ is uniformly distributed on the hypersphere, this probability must be independent of the actual direction of $\hat{x}$. Thus, without loss of generality, suppose that $\hat{x}$ points along the first axis. Then $p[\cos(x, \delta G) < \alpha] = p[\delta \hat{G}_1 < \alpha]$, where $\delta \hat{G}_1$ is the first component of $\delta \hat{G}$. Again, because δG is uniformly distributed on the hypersphere, this probability is simply the proportion of the hypersphere on which the first coordinate is less than α. This is precisely the area of the spherical cap, $S_r^n(\alpha)$.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

4 Markedness asymmetry in speech errors

The purpose of this section is to derive a formula for the probability of an error, and then use this formula to derive qualitative predictions about the relative frequency of different errors. In particular, the goal is to show how the Harmonic Grammar formalism explains the markedness asymmetry in speech errors, in which marked → unmarked errors (e.g. /g/ → [k]) are more frequent than unmarked → marked errors (e.g. /k/ → [g]).

The core result of this section is Theorem 2, which derives the formula for the probability of an error. After Theorem 2, this section formalises the notion of a minimal pair. Finally, Theorem 3 derives the markedness asymmetry.

Theorem 2 derives the formula for the probability of an error. The proof consists of three fundamental steps. The first step is to calculate the probability of an error for a given magnitude of disruption, which is derived from the spherical cap area function. The next step is to integrate this probability over all possible disruption magnitudes. The final step is to show that a larger harmony advantage leads to a smaller error probability, i.e. that the function is decreasing.

Theorem 2: If G is subject to spherically distributed disruption $\delta G \sim sph(\mu, N)$, then the probability of an $x \xrightarrow{y} z$ error is

$$\mathrm{p}[\delta G \text{ causes } x \xrightarrow{y} z] = \int_0^\infty \mu(\delta) S^N(-\frac{1}{\delta}\,\Delta \hat{H}_x^G(y, z))d\delta$$

where $N = mn + (n(n-1)/2)$ is the number of degrees of freedom in δG. Furthermore, this probability is a non-increasing function of the normalised harmony advantage of y over z.

Proof: For δG of fixed magnitude, Corollary 2 implies that $p[\delta G \text{ causes } x \xrightarrow{y} z]$

$$= p[\Delta H_x^{\delta G}(y, z) < -\Delta H_x^G(y, z)]$$
$$= p[\delta G \cdot \Delta\ R_x(y, z) < -G \cdot \Delta R_x(y, z)]$$

Dividing by $|\delta \mathrm{G}|$ and $|\Delta R_x(y, z)|$,

$$= p[\frac{\delta G \cdot \Delta R_x(y, z)}{|\delta G|\ |\Delta R_x(y, z)|} < \frac{G \cdot \Delta R_x(y, z)}{|\delta G|\ |\Delta R_x(y, z)|}]$$

$$= p[\cos(\delta G, \Delta R_x(y, z)) < -\frac{1}{|\delta G|}\, G \cdot \Delta \hat{R}_x(y, z)]$$

$$= S^N(-\frac{1}{|\delta G|}\,\Delta \hat{H}_x^G(y, z)]$$

where $N = mn + nn$ is the number of entries in G.

Then the total probability of of an $x \xrightarrow{y} z$ error can be found by integrating over all error magnitudes:

$$\mathrm{p}[\delta G \text{ causes } x \xrightarrow{y} z] = \int_0^\infty \mu(\delta) S^N(-\frac{1}{\delta}\,\Delta \hat{H}_x^G(y, z))d\delta$$

To see that this probability is non-increasing in $\Delta \hat{H}_x^G(y, z)$, let $\Delta \hat{H}_x^G(y_1, z_1) = \alpha < \beta = \Delta \hat{H}_x^G(y_2, z_2)$. Then $-\alpha/\delta > -\beta/\delta$, and so $S^N(-\alpha/\delta) > S^N(-\beta/\delta)$. The term $\mu(\delta)$, being a probability density, is always positive, and so $\mu(\delta)S^N(-\alpha/\delta) > \mu(\delta)S^N(-\beta/\delta)$. Integration (forward) preserves this inequality, so $p[\delta G$ causes $x \xrightarrow{y} z]$ is non-increasing in the normalised harmony advantage of y over z.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

The following definition formalises the notion of a phonological MINIMAL PAIR. We focus on such pairs so as to examine errors that are matched in terms of their representational distance from their targets.

Definition 8: Consider a pair of mappings $G(x) = y$ and $G(w) = z$. The underlying representations x and w contrast if $x \neq w$, and G preserves this contrast if $y \neq z$. The two mappings are equifaithful if $xFy^t = wFz^t$ and $xFz^t = wFy^t$. The surface representation y is less marked than z if $yMy^t > zMz^t$. The mappings are representationally equidistinct if $|\Delta R_x(y, z)| = |\Delta R_w(z, y)|$. The mappings form a contrast-preserving minimal pair with markedness asymmetry ($y > z$), written $x \rightarrow y > w \rightarrow z$, if they are equifaithful, contrast-preserving, representationally equidistinct, and y is less marked than z.

Theorem 3 derives the markedness asymmetry in speech errors. It follows straightforwardly from applying Theorem 2 to the definition of minimal pair, although some algebra is required.

Theorem 3: If $x \rightarrow y > w \rightarrow z$ then $\Delta\hat{H}^G_x(y, z) > \Delta\hat{H}^G_w(z, y)$.

Proof: By definition of markedness asymmetry, $yMy^t > zMz^t$. Doubling and rearranging terms,

$$yMy^t - zMz^t + yMy^t - zMz^t > 0$$

By equifaithfulness, $xFy^t = wFz^t$ and $xFz^t = wFy^t$, so $xFy^t - wFz^t = 0 = wFy^t - xFz^t$. Adding these terms,

$$xFy^t + yMy^t - xFz^t - zMz^t + wFy^t + yMy^t - wFz^t - zMz^t > 0$$

Rewriting,

$$H^G(x, y) - H^G(x, z) + H(w, y) - H(w, z) > 0$$

$$\Delta H^G_x(y, z) + \Delta H^G_w(y, z) > 0$$

$$G \cdot \Delta R_x(y, z) + G \cdot \Delta R_w(y, z) > 0$$

Because $\Delta R_w(y, z) = -\Delta R_w(z, y)$,

$$G \cdot \Delta R_x(y, z) > G \cdot \Delta R_w(z, y) > 0$$

By representational equidistinctness, $|\Delta R_x(y, z)| = |\Delta R_w(z, y)|$, so both sides can be divided by this magnitude:

$$G \cdot \frac{\Delta R_x(y, z)}{|\Delta R_x(y, z)|} > G \cdot \frac{\Delta R_w(z, y)}{|\Delta R_w(z, y)|}$$

Therefore,

$$G \cdot \Delta\hat{R}_x(y, z) > G \cdot \Delta\hat{R}_w(z, y)$$

So

$$\Delta\hat{H}^G_x(y, z) > \Delta\hat{H}^G_w(z, y)$$

Corollary 3: If G is subject to spherically distributed disruption and $x \rightarrow y > w \rightarrow z$ then $p[\delta G \text{ causes } w \xrightarrow{z} y] \geq p[\delta G \text{ causes } x \xrightarrow{y} z]$.

Proof: By Theorem 3, $\Delta\hat{H}^G_x(y, z) > \Delta\hat{H}^G_w(z, y)$. By Theorem 2, the probability of an error is non-increasing in the normalised harmony advantage. Thus $p[\delta G \text{ causes } w \xrightarrow{z} y] \geq p[\delta G \text{ causes } x \xrightarrow{y} z]$.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

REFERENCES

Aarts, Emile H. L. & Jan Korst (1989). *Simulated annealing and Boltzmann machines : a stochastic approach to combinatorial optimization and neural computing*. Chichester: Wiley.

Abd-El-Jawad, Hassan & Issam Abu-Salim (1987). Slips of the tongue in Arabic and their theoretical implications. *Language Sciences* **9**. 145–171.

Anttila, Arto (1997). Deriving variation from grammar. In Frans Hinskens, Roeland van Hout & W. Leo Wetzels (eds.) *Variation, change and phonological theory*. Amsterdam & Philadelphia: Benjamins. 35–68.

Béland, Renée & Carole Paradis (1997). Principled syllabic dissolution in a primary progressive aphasia case. *Aphasiology* **11**. 1171–1196.

Béland, Renée, Carole Paradis & Monique Bois (1993). Constraints and repairs in aphasic speech: a group study. *Canadian Journal of Linguistics* **38**. 279–302.

Berg, Thomas (1998). *Linguistic structure and change : an explanation from language processing*. Oxford: Clarendon.

Bíró, Tamás S. (2006). *Finding the right words : implementing Optimality Theory with simulated annealing*. PhD dissertation, University of Groningen.

Blumstein, Sheila E. (1973). *A phonological investigation of aphasic speech*. The Hague: Mouton.

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

Boersma, Paul & Bruce Hayes (2001). Empirical tests of the Gradual Learning Algorithm. *LI* **32**. 45–86.

Boersma, Paul & Joe Pater (2008). Convergence properties of a gradual learning algorithm for Harmonic Grammar. Ms, University of Amsterdam & University of Massachusetts, Amherst. Available as ROA-970 from the Rutgers Optimality Archive.

Buchwald, Adam B., Brenda Rapp & Maureen Stone (2007). Insertion of discrete phonological units: an articulatory and acoustic investigation of aphasic speech. *Language and Cognitive Processes* **22**. 910–948.

Butterworth, Brian & Harry Whittaker (1980). Peggy Babcock's relatives. In Georg E. Stelmach & Jean Requin (eds.) *Tutorials in motor behavior*. Amsterdam: North Holland. 647–656.

Chomsky, Noam (1980). *Rules and representations*. New York: Columbia University Press.

Clements, G. N. (1990). The role of the sonority cycle in core syllabification. In John Kingston & Mary E. Beckman (eds.) *Papers in laboratory phonology I : between the grammar and physics of speech*. Cambridge: Cambridge University Press. 283–333.

Coetzee, Andries W. (2006). Variation as accessing 'non-optimal' candidates. *Phonology* **23**. 337–385.

Dell, Gary S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review* **93**. 283–321.

Dell, Gary S., Kristopher D. Reed, David R. Adams & Antje S. Meyer (2000). Speech errors, phonotactic constraints, and implicit learning: a study of the role of experience in language production. *Journal of Experimental Psychology : Learning, Memory, and Cognition* **26**. 1355–1367.

Dell, Gary S., Myrna F. Schwartz, Nadine Martin Eleanor M. Saffran & Deborah A. Gagnon (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review* **104**. 801–838.

Frisch, Stefan (2000). Temporally organized lexical representations as phonological units. In Michael B. Broe & Janet B. Pierrehumbert (eds.) *Papers in laboratory*

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

phonology V : acquisition and the lexicon. Cambridge: Cambridge University Press. 283–298.
Frisch, Stefan & Richard Wright (2002). The phonetics of phonological speech errors: an acoustic analysis of slips of the tongue. *JPh* **30**. 139–162.
Gamkrelidze, Thomas V. (1978). On the correlation of stops and fricatives in a phonological system. In Joseph H. Greenberg, Charles A. Ferguson & Edith A. Moravcsik (eds.) *Universals of human language*. Vol. 2: *Phonology*. Stanford: Stanford University Press. 9–46.
Goldrick, Matthew (2002). *Patterns of sound, patterns in mind : phonological regularities in speech production*. PhD dissertation, Johns Hopkins University.
Goldrick, Matthew (2004). Phonological features and phonotactic constraints in speech production. *Journal of Memory and Language* **51**. 586–603.
Goldrick, Matthew (2007). Constraint interaction: a lingua franca for stochastic theories of language. In Carson T. Schütze & Victor S. Ferreira (eds.) *The state of the art in speech error research : Proceedings of the LSA Institute Workshop*. *MIT Working Papers in Linguistics* **53**. 95–114.
Goldrick, Matthew & Sheila E. Blumstein (2006). Cascading activation from phonological planning to articulatory processes: evidence from tongue twisters. *Language and Cognitive Processes* **21**. 649–683.
Goldrick, Matthew & Meredith Larson (2008). Phonotactic probability influences speech production. *Cognition* **107**. 1155–1164.
Goldrick, Matthew & Brenda Rapp (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition* **102**. 219–260.
Goldstein, Louis, Marianne Pouplier, Larissa Chen, Elliot Saltzmann & Dani Byrd (2007). Dynamic action units slip in speech production errors. *Cognition* **103**. 386–412.
Goldwater, Sharon & Mark Johnson (2003). Learning OT constraint rankings using a Maximum Entropy model. In Jennifer Spenador, Anders Eriksson & Östen Dahl (eds.) *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*. Stockholm: Stockholm University. 111–120.
Greenberg, Joseph H. (1966). *Language universals, with special reference to feature hierarchies*. The Hague & Paris: Mouton.
Guy, Gregory & Charles Boberg (1997). Inherent variability and the Obligatory Contour Principle. *Language Variation and Change* **9**. 149–164.
Hayes, Bruce & Colin Wilson (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI* **39**. 379–440.
Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences* **79**. 2554–2558.
Hopfield, J. J. (1984). Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences* **81**. 3088–3092.
Jäger, Gerhard (2007). Maximum entropy models and stochastic Optimality Theory. In Annie Zaenen, Jane Simpson, Tracy Holloway King, Jane Grimshaw, Joan Maling & Chris Manning (eds.) *Architectures, rules, and preferences : variations on themes by Joan W. Bresnan*. Stanford: CSLI. 467–479.
Jäger, Gerhard & Anette Rosenbach (2006). The winner takes it all – almost: cumulativity in grammatical variation. *Linguistics* **44**. 937–971.
Jakobson, Roman (1941). *Kindersprache, Aphasie, und allgemeine Lautgesetze*. Translated 1968 by Allan R. Keiler as *Child language, aphasia and phonological universals*. The Hague & Paris: Mouton.
Jarosz, Gaja (2006). *Rich lexicons and restrictive grammars : maximum likelihood learning in Optimality Theory*. PhD dissertation, Johns Hopkins University.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

Johnson, Mark (2002). Optimality-theoretic lexical functional grammar. In Paola Merlo & Suzanne Stevenson (eds.) *The lexical basis of sentence processing : formal, computational and experimental issues*. Amsterdam & Philadelphia: Benjamins. 59–74.

Kupin, Joseph J. (1982). *Tongue twisters as a source of information about speech production*. Bloomington: Indiana University Linguistics Club.

Laver, John (1980). Slips of the tongue as neuromuscular evidence for a model of speech production. In Hans W. Dechert & Manfred Raupach (eds.) *Temporal variables in speech : studies in honour of Frieda Goldman-Eisler*. The Hague: Mouton. 21–26.

Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990). Harmonic Grammar: a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Hillsdale: Erlbaum. 388–395.

Legendre, Géraldine, Antonella Sorace & Paul Smolensky (2006). The Optimality Theory–Harmonic Grammar connection. In Smolensky & Legendre (2006b: vol. 2). 339–402.

Levitt, Andrea G. & Alice F. Healy (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language* **24**. 717–733.

Lombardi, Linda (1999). Positional faithfulness and voicing assimilation in Optimality Theory. *NLLT* **17**. 267–302.

MacKay, Donald G. (1972). The structure of words and syllables: evidence from errors in speech. *Cognitive Psychology* **3**. 210–227.

Maddieson, Ian (1984). *Patterns of sounds*. Cambridge: Cambridge University Press.

Marshall, Jane (2006). Jargon aphasia: what have we learned? *Aphasiology* **20**. 387–410.

McMillan, Corey T., Martin Corley & Robin J. Lickley (2009). Articulatory evidence for feedback and competition in speech production. *Language and Cognitive Processes* **24**. 44–66.

Meyer, A. S. (1992). Investigation of phonological encoding through speech error analyses: achievements, limitations, and alternatives. *Cognition* **42**. 181–211.

Moreton, Elliott (2004). Non-computable functions in Optimality Theory. In John J. McCarthy (ed.) *Optimality Theory in phonology : a reader*. Malden, Mass.: Blackwell. 141–163.

Motley, Michael T. & Bernard J. Baars (1975). Encoding sensitivities to phonological markedness and transitional probabilities: evidence from spoonerisms. *Human Communication Research* **1**. 353–361.

Mowrey, Richard A. & Ian R. A. MacKay (1990). Phonological primitives: electromyographic speech error evidence. *JASA* **88**. 1299–1322.

Ouden, Dirk-Bart den (2002). *Phonology in aphasia : syllables and segments in level-specific deficits*. PhD dissertation, University of Groningen.

Pater, Joe (to appear). Weighted constraints in generative linguistics. *Cognitive Science*.

Pierrehumbert, Janet B. (2002). Word-specific phonetics. In Carlos Gussenhoven & Natasha Warner (eds.) *Laboratory Phonology 7*. Berlin & New York: Mouton de Gruyter. 101–139.

Potts, Chris, Joe Pater, Rajesh Bhatt & Michael Becker (2008). *Harmonic Grammar with Linear Programming : from linguistic systems to linguistic typology*. Ms, University of Massachusetts, Amherst. Available as ROA-984 from the Rutgers Optimality Archive.

Pouplier, Marianne (2007). Tongue kinematics during utterances elicited with the SLIP technique. *Language and Speech* **50**. 311–341.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

Pouplier, Marianne (2008). The role of a coda consonant as error trigger in repetition tasks. *JPh* **36**. 114–140.
Pouplier, Marianne & William Hardcastle (2005). A re-evaluation of the nature of speech errors in normal and disordered speakers. *Phonetica* **62**. 227–243.
Prince, Alan (2007). The pursuit of theory. In Paul de Lacy (ed.) *The Cambridge handbook of phonology*. Cambridge: Cambridge University Press. 33–60.
Prince, Alan & Paul Smolensky (1993). *Optimality Theory: constraint interaction in generative grammar*. Ms, Rutgers University & University of Colorado, Boulder. Published 2004, Malden, Mass. & Oxford: Blackwell.
Rapp, Brenda & Matthew Goldrick (2006). Speaking words: contributions of cognitive neuropsychological research. *Cognitive Neuropsychology* **23**. 39–73.
Reynolds, William (1994). *Variation and phonological theory*. PhD dissertation, University of Pennsylvania.
Roberts, Julie (1997). Acquisition of variable rules: a study of (-t, d) deletion in preschool children. *Journal of Child Language* **24**. 351–372.
Romani, Cristina & Andrea Calabrese (1998). Syllabic constraints in the phonological errors of an aphasic patient. *Brain and Language* **64**. 83–121.
Romani, Cristina & Claudia Galluzzi (2005). Effects of syllabic complexity in predicting accuracy of repetition and direction of errors in patients with articulatory and phonological difficulties. *Cognitive Neuropsychology* **22**. 817–850.
Romani, Cristina, Andrew Olson, Carlo Semenza & Alessia Granà (2002). Patterns of phonological errors as a function of a phonological versus an articulatory locus of impairment. *Cortex* **38**. 541–567.
Shattuck-Hufnagel, Stefanie & Dennis H. Klatt (1979). The limited use of distinctive features and markedness in speech production: evidence from speech error data. *Journal of Verbal Learning and Verbal Behavior* **18**. 41–55.
Smolensky, Paul (1986). Information processing in dynamical systems: foundations of Harmony Theory. In D. E. Rumelhart, J. L. MacClelland & The PDP Research Group (eds.) *Parallel Distributed Processing: explorations in the micro-structure of cognition*. Vol. 1. Cambridge, Mass.: MIT Press. 194–281.
Smolensky, Paul (2006a). Optimization in neural networks: harmony maximization. In Smolensky & Legendre (2006b: vol. 1). 345–392.
Smolensky, Paul (2006b). Computational levels and integrated connectionist/symbolic explanation. In Smolensky & Legendre (2006b: vol. 2). 503–592.
Smolensky, Paul & Géraldine Legendre (2006a). Formalizing the principles II: optimization and grammar. In Smolensky & Legendre (2006b: vol. 1). 207–234.
Smolensky, Paul & Géraldine Legendre (2006b). *The harmonic mind: from neural computation to optimality-theoretic grammar*. 2 vols. Cambridge, Mass.: MIT Press.
Smolensky, Paul, Géraldine Legendre & Bruce Tesar (2006). Optimality Theory: the structure, acquisition, and use of grammar. In Smolensky & Legendre (2006b: vol. 1). 453–544.
Soderstrom, Melanie, Donald W. Mathis & Paul Smolensky (2006). Abstract genomic encoding of universal grammar in Optimality Theory. In Smolensky & Legendre (2006b: vol. 2). 403–471.
Stemberger, Joseph Paul (1983). *Speech errors and theoretical phonology: a review*. Bloomington: Indiana University Linguistics Club.
Stemberger, Joseph Paul (2004). Neighbourhood effects on error rates in speech production. *Brain and Language* **90**. 413–422.
Taylor, Conrad F. & George Houghton (2005). Learning artificial phonotactic constraints: time course, durability, and relationship to natural constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**. 1398–1416.
Tesar, Bruce (1999). Robust interpretive parsing in metrical stress theory. *WCCFL* **17**. 625–639.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press

Vitevitch, Michael S. & Paul A. Luce (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* **40**. 374–408.
Vousden, Janet I., Gordon D. A. Brown & Trevor A. Harley (2000). Serial control of phonology in speech production: a hierarchical model. *Cognitive Psychology* **41**. 101–175.
Wan, I-Ping & Jeri Jaeger (1998). Speech errors and the representation of tone in Mandarin Chinese. *Phonology* **15**. 417–461.
Warker, Jill A. & Gary S. Dell (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology : Learning, Memory, and Cognition* **32**. 387–398.
Warker, Jill A., Gary S. Dell, Christine A. Whalen & Samantha Gereg (2008). Limits on learning phonotactic constraints from recent production experience. *Journal of Experimental Psychology : Learning, Memory, and Cognition* **34**. 1289–1295.
Wilson, Colin (2006). Learning phonology with substantive bias: an experimental and computational study of velar palatalization. *Cognitive Science* **30**. 945–982.
Xu, Zong-Ben, Guo-Qing Hu & Chung-Ping Kwong (1996). Asymmetric Hopfield-type networks: theory and applications. *Neural Networks* **9**. 483–501.
Zipf, George (1935). *The psycho-biology of language : an introduction to dynamic philology*. Boston: Houghton Mifflin.
Zwicky, Arnold M. (1980). 'Internal' and 'external' evidence in linguistics. In *PSA : Proceedings of the Biennial Meeting of the Philosophy of Science Association*. Vol. 2: *Symposia and invited papers*. Chicago: University of Chicago Press. 598–604.

https://doi.org/10.1017/S0952675709001742 Published online by Cambridge University Press