

Chapter 6

Dutch Voice Assimilation with SA-OT

6.1 The magic square

In this chapter, we are discussing a set of linguistic variations that one could call the *magic square*. Their common characteristic is that two related features vary in a synchronous way. The basic structure is represented in Figure 6.1, where + and – are the possible values of the two consecutive features.

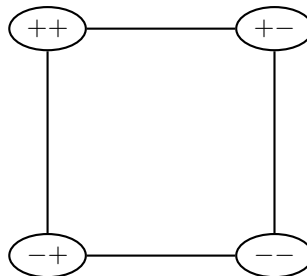


Figure 6.1: The “magic square”

The prototypical example, which we shall presently use in the discussion, is *voice assimilation*: if two neighbouring stops have different [voice] features, either regressive or progressive assimilation takes place, yielding a homogeneous sequence with respect to this articulative feature. Hence, candidates ++ and -- are favoured over candidates +- and -+. Steriade, Lombardi, Joe Pater, Eric Bakovic and others have argued that it is very uncommon (impossible) for a language to follow a third strategy, such as inserting an epenthetic vowel in order to avoid clash in voicing (e.g. Lombardi, 2001), even though epenthesis is a frequently used strategy to avoid prohibited consonant clusters, in general.¹

¹Thus, in Brazilian Portuguese, *football* translates to *futebol* and English *handball* to *handebol*. For schwa epenthesis in Modern Hebrew, see for instance Bíró and Hamp (2002). But epenthesis is claimed not to be a repair strategy for syllable codas having an unwanted [voice] feature. Not only can an epenthetic vowel never intervene between two stops with dissimilar [voice] features, but suffixing a final schwa is also not an option in languages prohibiting voiced consonants in a word-final position.

If + in Fig. 6.1 stands for [+voice], and – for [-voice], then out of the four possibilities, only ++ and -- may be grammatical in many languages. Furthermore, although for some input usually only one of ++ and -- is grammatical, yet sometimes the other may surface as an alternate form. Our paradigmatic example will be the Dutch word pair *op die* ('in this...'), where the clash in the voice feature can be solved by assimilation in either of the two possible directions. Dutch phonology requires *regressive* voice assimilation, that is, *o[bd]ie*, and yet, often *o[pt]ie* emerges as the result of *progressive* voice assimilation.

Further examples can be also found that exhibit a similar *magic square*. The Dutch word *partij* ('(political) party') is sometimes pronounced as [ptɛi],² forming an otherwise prohibited open consonant cluster with the deletion of two segments. Now + represents the presence of the segment and – its absence in the rhyme, and again the same diagonally opposed forms alternate: the grammatical form ++ with the alternative form --. The two other forms involving only partial deletion in the rhyme, +- and -+, are not allowed.

An analogous situation is used by Bíró and Gervain (2006): the *resyllabification* of the [z] in the Hungarian definite article *a / az*. The choice between the two allomorphs depends on whether the next word begins with a consonant or with a vowel:

$$\begin{array}{ll} az \text{ énekesnő} & \text{'the soprano'}, \\ a \text{ kopasz énekesnő} & \text{'the bald soprano'}. \end{array} \quad (6.1)$$

The definite article is also prone to undergo resyllabification turning the [z] into the onset of the subsequent syllable. In other words, the pause between the article and the subsequent word can drop, and therefore the segment [z] can be perceived as belonging to the next word, sometimes leading to misunderstanding in the case of minimal pairs, and sometimes to language games.

Judit Gervain performed a controlled psycholinguistic experiment measuring the frequency of this phenomenon. Hungarian distinguishes at least five speech levels on the basis of the rate/speed of speech, and she tested two of them: (i) motherese or infant-directed speech, characterised by a rather slow pace and emphatic, exaggerated prosody, (ii) and fluent, casual, conversational style with a medium speech rate. The hypothesis, which had been already described theoretically but never measured empirically (Kiefer, 1994), claims that more resyllabification occurs with the acceleration of the speech rate. The experiments confirmed this hypothesis by measuring the overall length and the presence of pauses in critical minimal pairs (e.g. *az ár* 'the price' vs. *a zár* 'the lock') excised from test sentences pronounced by three female native speakers

I am thankful to everybody who answered my question on Linguist List in February 2005. That is where I was referred, among others, to the following urls:

<http://roa.rutgers.edu/view.php?id=29>,

http://www.linguistics.ucla.edu/people/steriade/papers/P-map_for_phonology.doc,

<http://people.umass.edu/pater/pater-balantak.pdf>,

http://camba.ucsd.edu/bakovic/work/bakovic_wilson_lar.pdf.

²At least, it was pronounced so by the late Dutch prime minister Joop den Uyl within the context *Partij van de Arbeid*, 'Labour Party'. Otherwise, the native speaker would not judge [ptɛi] necessarily better than [patei] or [prtɛi], or would also allow the insertion of a schwa ([p@tɛi]). Note, however, that, similarly to Boersma (2004b), in SA-OT grammaticality judgements need not correlate with production: a form might be produced by the performance model (SA-OT) even if it is judged absolutely out of question by the underlying competence model (OT).

in both conditions. In the production of the slower infant-directed speech, resyllabification (*az ár* pronounced as *a zár*) happens about in 40% of the cases, which raises to about 80% in conversational style. On the perception side, she tested whether a naïve group of native speakers could identify which of the minimal pairs were pronounced by the speakers. When the segmentation cue—the pause—was present, subjects identified the words with a 85% accuracy, while they were at chance when the words were pronounced without a pause.

Now, the + and – values of the *magic square* describes the presence and the absence of the segment [z] on either side of the syllable boundary. Candidate +- corresponds to the input obtained by concatenating the lexical items (*az.énekesnő*), whereas -+ is the resyllabified form (*a.zénekesnő*). According to Gervain, the -- form (*a.énekesnő*) appears in children’s speech.

In an alternative analysis, + corresponds to the preferred syllable structures (an empty coda and an onset filled with [z]), and – to the disfavoured ones (a coda filled with [z] and an empty onset). Now, the original form *az.énekesnő* is -- and the resyllabified form *a.zénekesnő* is ++, whereas +- (*a.énekesnő*) and -+ (*az.zénekesnő*) could be but are non attested—similarly to the +- and -+ candidates of all of the previous examples. Observe that the resyllabified form is the best with respect to syllable structure, whereas the original form is the worst one among the four theoretical possibilities. And yet, if with respect to some other factor forms +- and -+ are worse, candidate -- becomes a local optimum thanks to the two candidates separating it from the global optimum in Fig. 6.1. We shall return to this phenomenon in section 7.1, and present a detailed analysis based on a subsequent experiment of Judit Gervain.

As an example from syntax, I take the favourite one of Modern Hebrew linguist purists. The “correct” form for ‘three shekels’ would be *šloša škalim*, with agreement both in gender (morphologically visible on the cardinal number) and in number (the [-im] plural suffix on the noun). Nonetheless, most speakers use *šaloš šekel*, omitting both agreement features.³ This magic square is formed by the presence (+) or absence (–) of agreement in number (first position) and in gender (second position). The grammar, in general (for noun+adjective pairs), requires again candidate ++ to win, but in some special cases (namely, with numerals) speaker might also produce --, but not candidates +- and -+.

Notice that in many of these examples, form --, the alternative one, occurs in frequent (“semi-lexicalised”) constructions. Progressive voice assimilation in Dutch would be unconceivable in nouns such as *zakdoek* or *duikboot*, only in bi-grams of unstressed function words. The pronunciation [pteɪ] of the word *partij* was characteristic to Joop Den Uyl, the late prime minister of the Netherlands (D. Gilbers and M. Schreuder, personal communication) in the expression *Partij van de Arbeid* (‘Labour Party’). The lack of double agreement in Modern Hebrew occurs only in frequently used expressions of quantity.

An Optimality Theoretical account of these phenomena should include at least two constraints. The first constraint C1 requires surface homogeneity, punishing the heterogeneous forms +- and -+. The lower ranked constraint

³Some speakers in colloquial Hebrew omit the agreement in gender for all cardinal + noun constructs. In their case, however, one may argue that the masculine forms of the numerals have been removed from the language altogether for the sake of paradigm uniformity, due to their counter-intuitiveness. Namely, in semitic languages the gender morphemes on numerals are the opposite of the gender morphemes generally found in the grammar.

C2 prefers ++ to --, whereas -+ and +- may either satisfy or violate C2. In the following tableau, the well-known \mathbb{E} points to the grammatical form, whereas the \sim symbol shows the alternative form:

	C1	C2
\mathbb{E} ++		
\sim --		*
+-	*!	?
-+	*!	?

(6.2)

Indeed, tableau (6.2) together with the candidate space topology in Fig. 6.1 will turn ++ into the global optimum, and -- into the only alternative local optimum (see Fig. 6.4 on page 172 for a concrete example).

To tell the truth, each story is a little bit more complex if we want to stay linguistically correct. Still, the basic structure of the tableaux remains similar. A crucial property of these tableaux is that +- and -+ are defeated in an earlier stratum, while the difference between ++ and -- appears only lower in the hierarchy.

In case of the word *partij* pronounced [ptɛi], one may propose using two faithfulness constraints. Constraint C1 is FAITHFULNESS[RHYME]: each syllable rhyme in the input is identical to its image in the output form, *if* there is such an image. The rhyme /ar/ in the input /par.tei/ is identical to its image [ar] in the candidate [partɛi], but different from [a] and [r] in candidates [patɛi] and [prtɛi] respectively. That rhyme, however, has no correspondent in the string [ptɛi] for it has been deleted altogether, therefore the constraint is satisfied vacuously.⁴ Subsequently, constraint C2 is a faithfulness constraint on segments: deleting each segment increases the number of violation marks by one. This constraint then favours [partɛi] with zero violation marks to [ptɛi] with two violation marks. Candidates [prtɛi] and [patɛi] are assigned only one violation mark each, but they have been already put out of the game previously by constraint C1.

In Hungarian resyllabification, C1 is a constraint which requires a strictly alternating vowel-consonant sequence, which is satisfied both by *a.zénekesnő* and by *az.énekesnő*, but not by *a.énekesnő* or by *az.zénekesnő*. C2 can be derived from constraints ONS and NOCODA known from Basic Syllable Structure Theory (Prince and Smolensky, 2004): by their sum, each empty syllable onset and each filled coda incurs one violation mark.

Using a “pseudo-minimalist” approach in the case of syntactic agreement in Modern Hebrew, constraint C1 in (6.2) can be said to require agreement features

⁴Similarly to what will be said on agreement in Hebrew, we could differentiate between a candidate [p.tei] in which the correspondence relation is not defined on the underlying rhyme /ar/ (thus, no image in the candidate), and a candidate [p∅∅.tei] in which the correspondence relation maps the rhyme /ar/ to an empty string, incurring two violation marks. Introducing the second candidate does not have any effect, for it is always a loser, because it is harmonically bounded by other candidates. A different, maybe more convincing but less elegant solution is to use two constraints to eliminate candidates [prtɛi] and [patɛi]. The first candidate can be easily eliminated by using highly ranked syllable structure constraints that do not allow a complex onset [prt] (being too complex and violating sonority requirements), and do not allow for the syllabification of [r] as a nucleus either. The second candidate may be eliminated by using a simpler and more convincing version of FAITHFULNESS[RHYME]: a rhyme in the surface form has to correspond to a rhyme in the underlying form. This second version of FAITHFULNESS[RHYME] is satisfied by [par.tei], [ptɛi], [prtɛi] and [pr.tei], but not by [pa.tei], for the rhyme [a] does not correspond to the rhyme /ar/ in the input form.

to be either checked or unchecked: *in the case* checking does take place overtly, then no feature may be left unchecked (one violation mark per feature left unchecked). The form *—* (*šaloš šekel*) satisfies this constraint automatically because no feature checking occurs. (An alternative candidate, identical on the surface, would violate this constraint twice if it involves feature checking, but then both gender and number are left unchecked.) Forms *+—* and *—+* (*šloša šekel* and *šaloš skalim*) do involve feature checking, but not all features are checked, which leads to violating constraint C1. Subsequently, constraint C2 requires features to be checked, so any unchecked (not agreeing) feature incurs one violation mark. Hence *šaloš šekel* with two unchecked features is worse than *šloša skalim* (both features checked) for C2. The two constraints are almost identical, the only difference being that the lower ranked constraint requires features be checked always, whereas the higher ranked constraint requires it only if feature checking is performed in general.

Finally, in the voice assimilation example, phonology would propose a markedness constraint $[\alpha\text{voice}][\alpha\text{voice}]$, requiring a homogeneous sequence with respect to the [voice] feature; as well as a faithfulness constraint that punishes any change of the value in the [voice] feature compared to the input form. These two constraints would not distinguish however between *o[bd]ie* and *o[pt]ie*, for both satisfy markedness and both violate faithfulness once. Therefore, the regressiveness of the assimilation should be also incorporated into the analysis. In addition, we also would like to consider forms with epenthesis in a subsequent approach, thus the constraint DEP is required to punish epenthetic forms. In the following section, we work out the details of this analysis.

6.2 Voice assimilation in Dutch

Voice assimilation in general, and regressive voice assimilation of neighbouring stops in particular, is an extremely widespread phenomenon across languages. Not surprisingly, we can also observe it in Dutch, a language that tends to neutralise the [voice] feature of obstruents in other contexts, as well, such as in the word-final position.

The middle consonant cluster in words such as *duikboot* ('submarine') or *zakdoek* ('handkerchief') exemplifies *regressive voice assimilation*: in these cases we obtain [gb] and [gd] respectively. The coda of the previous syllable assimilates to the onset of the subsequent syllable. The traditional way to account for this phenomenon in Optimality Theory is to assume two constraints, namely a faithfulness constraint overranked by a markedness constraint. The constraint FAITH[VOICE] requires the value of the [voice] feature be kept unchanged in the output, whereas ASSIMILATE[VOICE] punishes adjacent stops not sharing their [voice] feature in the surface form. The need for the faithfulness constraint is supported by hypercorrect (or extremely careful) pronunciation yielding *za[kd]oek* and *dui[kb]oot*: in this register FAITH[VOICE] is promoted above the markedness constraint ASSIMILATE[VOICE], due to which assimilation may not take place.⁵

⁵As Paul Boersma pointed out, the word *handboek* ('hand book') is pronounced as *han[tb]oek* in equally careful speech, violating both FAITH[VOICE] and ASSIMILATE[VOICE]. This case might be influenced by another factor, such as an Output-Output Correspondence to the word *han[t]*.

Actually, careful vs. careless speech is often seen as a parameter orthogonal to speech rate (e.g., Kiefer, 1994), the first factor being dependent upon the social context, while the second being determined by time pressure on the individual speaker. Fast careful speech may have different characteristics from careless speech, which itself can also have different speech rates. Hence, they might have to be modelled separately. This is why employing constraint reranking to account for these extremely careful or hypercorrect forms does not contradict our agenda of using SA-OT for speech rate dependent phenomena. In turn, constraint reranking—either performed categorically, or in a Stochastic OT-style—reflects the intuitive view that extremely careful or hypercorrect speech is indeed about faithfulness to the underlying form; or, more precisely, to the written form in literate languages with a prescriptive tradition. Hypercorrectness could be seen as a separate register (or language), thus stipulating a separate hierarchy is not in conflict with our previous criticism about supposing separate hierarchies for different speech rates.

After these considerations, we can focus on phenomena that are typical to speech rate (or other factors), and we may ignore the hypercorrect forms. Dutch features an additional variation: the preposition *op* followed by *die* (only as a demonstrative pronoun or an article, such as in *op die manier* ‘in that way’) may sometimes involve *progressive* voice assimilation, and result in the consonant cluster [pt], besides the form [bd] yielded by regressive assimilation.⁶

Progressive voice assimilation between stops seems to contradict our belief in a homogeneous Dutch phonological system, because exclusively regressive assimilation is allowed everywhere else. In order to save the uniform phonology, as part of the supposed linguistic competence of the native speaker, we shall try explaining the form *o[pt]ie* as a performance phenomenon and reproducing it with simulated annealing. Our strategy here is to exile exceptions from competence (the static mental representation of the language), and to use the performance (or computational-production) model to account for them. If it works, we can keep the competence model simple and still account for all observed data.

Two models will be introduced, and these two models will demonstrate the capabilities and restrictions of SA-OT. Indeed, the real goal of this chapter is to further analyse what SA-OT is able to do, rather than to account for Dutch progressive assimilation in particular. The latter is taken as a mere example out of the analogous phenomena listed in the previous section, and ongoing and future work (such as Bíró and Gervain, 2006) should collect more empirical data that our simulations ought to reproduce.

The first model uses a finite (actually, quite restricted) search space, and is only able to account for a 50%-50% distribution of the forms *o[pt]ie* ([pt] in short) vs. *o[bd]ie* ([bd], henceforth), independently of the parameter settings.

⁶As I have been informed by my readers, for further references on the subject see for instance: Wim Zonneveld: Lexical and phonological properties of Dutch voice assimilation, in: Van der Broecke *et al.* (eds.): *Sound Structures, Studies for Antonie Cohen*, Floris, Dordrecht, 1983:297-312; or Mirjam Ernestus: *Voice Assimilation and Segment Reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface*, PhD thesis, Vrije Universiteit, Amsterdam, Amsterdam, 2000. Wim Zonneveld claims that the double forms are limited to clitic-like non-lexical categories, so *op deze lijst* ‘on this list’ can be realised both as [bd] and [pt], but *op dikke boeken* ‘on thick books’ must be [bd].

Adam Albright pointed out that Northeastern Yiddish displays a similar progressive voice assimilation that is basically limited again to a function word, namely, to the reflexive pronoun *zikh*. For instance, *golt zikh* ‘shave-3sg’ becomes *gol[ts]ikh*.

The model will be a slightly more complicated version of the toy example presented in section 2.3.2. The lesson is that SA-OT does not necessarily converge towards maximal precision. Instead of interpreting this observation as a failure of SA-OT, I propose to see it as a source of hope for the frequent cases where simple and elegant linguistic models have had to be turned into very complex ones just because of some few annoying exceptions.

Based on this model, I argue that linguistic data might be reproduced by keeping the competence model simple and by leaving the dirty job to such performance models. And since this family of performance models always predicts errors, independently of the parameter settings, one cannot distinguish *a priori* between phenomena related to competence in its narrow sense and between phenomena constantly introduced by the second level on Table 2.1 (page 43). This case is contrasted to the situation presented in Chapter 5, where the output frequencies depended on the parameters, and therefore the *allegro* form, whose frequency increased at higher speech rate, could be identified as the performance error form.

In contrast to the first one, the second model will allow tuning the frequencies of candidates [bd] (*o[bd]ie*) and [pt] (*o[pt]ie*) by varying the parameters. True, this second model necessitates a constraint which may not meet the expectations of all phonologists, for it is a markedness constraint referring also to the underlying form. However, the model turns to be illuminating about the possibilities of Simulated Annealing Optimality Theory, whereas further research may replace the problematic constraint with a less controversial one.

6.3 The building blocks of Simulated Annealing

First, we have to define the candidate set with respect to a given underlying form. Let the underlying form be a pair of stops $\sigma_1\sigma_2$. Now, σ'_1 denotes the stop that has the same features as σ_1 , but the [voice] feature is different; similarly for σ'_2 . The candidate set will then be a set of strings beginning with either σ_1 or σ'_1 , ending with either σ_2 or σ'_2 , and having zero or more epenthetic vowels (say, schwas) in-between—the simple regular language $\{\sigma_1, \sigma'_1\} \times @^* \times \{\sigma_2, \sigma'_2\}$.

We have not really argued for the need to epenthesise yet, but we advance it here for the sake of the second model to be presented in this chapter. After all, epenthesis is always a possibility for a phonologist, who would never prevent GEN from producing candidates including epenthetic segments, footnote 1 on page 161 notwithstanding.

To simplify notation, let us replace σ_1 and σ_2 by [p] and [d] (from the input *op die*). We write then the underlying (input) form as /pd/, and the output forms (candidates) as [pd], [bd], [pt], [bt], [p@d], [b@@t], etc. The @ symbol will refer to the epenthetic vowel, and a superscript may refer to its repetition n times (e.g. [p@ⁿd]), zero or more times (Kleene-star: [p@*d]), or one or more times (Kleene-plus: [p@⁺d]).

As we follow the usual steps of introducing the building blocks of SA-OT (see page 45 or page 129), we have to define next the neighbourhood structure on this set. We shall regard two candidates as neighbours if and only if one candidate can be reached from the other by performing exactly one of the following *basic steps*:

- Insert or delete exactly one epenthetic vowel (from $\sigma_1@^n\sigma_2$ to $\sigma_1@^{n\pm 1}\sigma_2$).

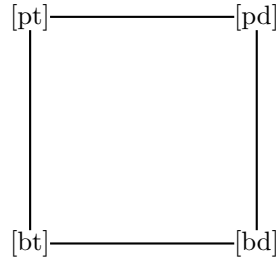


Figure 6.2: **Search space used in the first model for voice assimilation:** neighbours are connected by a line.

- Change the value of the [voice] feature of exactly one of the bordering stops (from $\sigma_1 @^n \sigma_2$ either to $\sigma'_1 @^n \sigma_2$ or to $\sigma_1 @^n \sigma'_2$).

In our first model the candidate set will be restricted to the four pairs of stops without allowing any epenthesis ($n = 0$) (Fig. 6.2). By adding the possibility of iterative epenthesis, we arrive at the structured candidate set appearing in Fig. 6.3, the one to be used in the second model.

As for the *a priori* probabilities, that is, the second part of the definition of a topology on the search space, we simply give equal probability to each neighbour of a candidate (Eq. (2.5) on page 49).

Now, let us move forward, defining the constraints. Suppose \mathcal{C}_w is the *correspondence relation* (see also section 4.1.5), that is a partial bijection⁷ mapping (some of) the segments (tokens) of the input string onto (some of) the segments of candidate w fulfilling *contiguity*,⁸ as defined by *Correspondence Theory* (cf. McCarthy and Prince (1993b) p. 67).

In our case, \mathcal{C}_w maps the first and the second underlying stops onto the first and the last stops in the candidate w , respectively. The epenthetic vowels are not contained in the range of \mathcal{C}_w .

We shall use the following constraints (the definition provided is more general than needed for the present case):⁹

- DEP (DEPENDENCY, “don’t epenthesise!”): one violation mark assigned to each segment in the candidate that does not correspond to a segment in the input string. In other words, a candidate w is assigned as many violation marks as the number of its segments, minus the cardinality of the range of \mathcal{C}_w
- ASSIMILATE[VOICE]: one violation mark to each pair of segments (σ_1, σ_2) in the candidate such that σ_1 immediately precedes σ_2 (in the candidate

⁷I shall call a relation $\mathcal{R} \subset A \times B$ a *partial bijection* if and only if \mathcal{R} is a bijection—a one-to-one mapping—between its domain and its range, even if its domain and its range may be a proper subset of A and B respectively.

⁸In the present model, *contiguity* requires that for all segments σ_1 and $\sigma_2 \in \text{Domain}(\mathcal{C}_w)$, segment $\mathcal{C}_w(\sigma_1)$ is left of $\mathcal{C}_w(\sigma_2)$ in the candidate string if and only if σ_1 is left of σ_2 in the input string.

⁹For historical reasons, constraints are typically defined in terms of what criteria must be met: what is the structure that does not incur any violation mark. However, Optimality Theory constraints are functions on the candidates that have not necessarily Boolean (true / false) values. Very often, the *number* of violation marks assigned plays a crucial role. This is why I repeatedly argue that constraints should be defined positively, by giving the *number* of violation marks they assign to a given candidate.

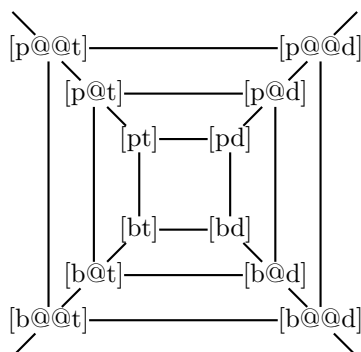


Figure 6.3: **Search space used in the second model for voice assimilation:** the character @ stands for the epenthetic schwas.

string), both have a [voice] feature, and yet, these features have a different value.

- STOPASSIMILATION=REGRESSIVE: one violation mark to each pair of segments (σ_1, σ_2) in the underlying representation such that σ_1 immediately precedes σ_2 (in the *underlying form*), both are elements of the domain of \mathcal{C}_w , furthermore σ_2 , $\mathcal{C}_w(\sigma_1)$ and $\mathcal{C}_w(\sigma_2)$ are stops with a [voice] feature, and yet these three voice features do not have the same value.
- FAITH[VOICE]: one violation mark is assigned to each segment σ in the underlying form such that $\mathcal{C}_w(\sigma)$ exists, both σ and $\mathcal{C}(\sigma)$ have a [voice] feature, and yet these features have a different value.

Constraint DEP, originally FILL in Prince and Smolensky (1993), is the standard constraint in Optimality Theory that prohibits inserting elements into a candidate that were not present in the input form. Constraint ASSIMILATE[VOICE], a straightforward way to compel stops to agree in voicing, is called AGREE by Lombardi (1995). She refers to FAITH[VOICE] as IDENT(laryngeal). Yet, her constraint IDENTONSET(laryngeal), which causes voice assimilation to be regressive by punishing unfaithful onsets but not codas, is not going to be useful in our analysis. Indeed, our fourth constraint, STOPASSIMILATION=REGRESSIVE might be claimed to be the most questionable.¹⁰

Notice that constraint STOPASSIMILATION=REGRESSIVE punishes all underlyingly adjacent pairs of stops that do not assimilate regressively (either do not assimilate at all or assimilate progressively, if they are different underlyingly), even if they are not adjacent on the surface. This constraint sounds quite weird to the ears of a phonologist, for markedness constraints should refer to properties of the surface form alone, without touching upon the underlying form. Nevertheless, we shall need it for our second approach.¹¹ In the first approach, the following simpler alternative may replace it:

¹⁰My impression was that finding the constraint corresponding to STOPASSIMILATION=REGRESSIVE is also the most difficult task when applying this model to the analogous phenomena mentioned earlier.

¹¹What the model requires is tableau (6.5). Alternative formulations of this constraint might be possible, which will assign violation marks with no significant change. In the footsteps of Lombardi (1995)'s constraint IDENTONSET(laryngeal), we could for instance give the following definition: one violation mark is assigned to each stop in an onset position that (i) is not

- STOPASSIMILATION=REGRESSIVE2: one violation mark goes to each pair of segments (σ_1, σ_2) in the candidate string iff σ_1 immediately precedes σ_2 (in the *candidate string*), both are elements of the range of \mathcal{C}_w , furthermore $\sigma_1, \sigma_2, \mathcal{C}_w^{-1}(\sigma_1)$ and $\mathcal{C}_w^{-1}(\sigma_2)$ are stops with a [voice] feature, assimilation has taken place (σ_1 and σ_2 share the same [voice] feature), and yet σ_1 has the same [voice] feature as $\mathcal{C}_w^{-1}(\sigma_1)$, whereas σ_2 differs from $\mathcal{C}_w^{-1}(\sigma_2)$ in this feature. (In short, progressive assimilation has occurred.)

As we have to check whether assimilation has been progressive or regressive, we probably cannot avoid referring to the underlying form, at least in some hidden way. Yet, this second formulation differs in two respects from the first one. Firstly, it does not punish progressive assimilation anymore if epenthetic vowels intervene between the stops in the surface form: this is exactly the point making the first definition unattractive to a phonologist but necessary for the second model to be presented. Secondly, the second formulation does not assign any violation mark to candidates where no assimilation has taken place (vacuous application), whereas the first formulation punished them for missing the occasion of assimilating (in a regressive way). Consequently, candidates [pd] and [bt] vacuously fulfil STOPASSIMILATION=REGRESSIVE2, whereas they violate STOPASSIMILATION=REGRESSIVE. This difference will not have any effect in the models, since these two candidates are already defeated by their neighbours due to a higher ranked constraint, namely ASSIMILATE[VOICE]. The second model will, nevertheless, crucially exploit the fact that [b@+d] are the only candidates with epenthesis satisfying this constraint, consequently that model necessitates that [p@+d] and [b@+t] violate it, too.

The last step in constructing our Simulated Annealing Optimality Theory model is to define the hierarchy. As explained in the introduction, the faithfulness constraint has to be demoted below the markedness constraints, otherwise no assimilation will take place. Constraint DEP, which will play a role only in the second model, should be ranked high in order to avoid forms with epenthesis becoming successful. Similarly, the relative ranking of ASSIMILATE[VOICE] and STOPASSIMILATION=REGRESSIVE is determined by the fact that *o[pt]ie* should emerge as an alternative form, and not *o[pd]ie*. In summary, the following ranking is the most likely to help us:

$$\begin{aligned} \text{DEP} \gg \text{ASSIMILATE[VOICE]} \gg \\ \gg \text{STOPASSIMILATION=REGRESSIVE} \gg \text{FAITH[VOICE]} \end{aligned} \quad (6.3)$$

Before going on with the analysis of Simulated Annealing, let us review the tableaux produced by this constraint hierarchy. The well-known \mathfrak{E} symbol refers to the optimal candidate, whereas \sim will refer again to the alternative form. In the first model, we may use STOPASSIMILATION=REGRESSIVE2, yielding the following chart (*vac* meaning that the constraint is satisfied vacuously):

faithful in its [voice] feature if the following syllable nucleus is original; (ii) is faithful to the input form in its [voice] feature if the following syllable nucleus is epenthetic.

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]2	FAITH[VOICE]
☞ [bd]				*
~ [pt]			*!	*
[pd]		*!	vac	
[bt]		*!	vac	**

(6.4)

For the second model, we need to use the original formulation of the constraint STOPASSIMILATION=REGRESSIVE and to enlarge our candidate set. The @ symbol refers to the epenthetic vowel (for instance, a schwa), and the exponent n multiplies the preceding character (in the tableau $n > 0$).

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]	FAITH[VOICE]
☞ [bd]				*
~ [pt]			*!	*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	* ⁿ !			*
[p@ ⁿ t]	* ⁿ !		*	*
[p@ ⁿ d]	* ⁿ !		*	
[b@ ⁿ t]	* ⁿ !		*	**
...

(6.5)

6.4 Model 1: Finite search space

In the first approach, the search space (the structured candidate set) is restricted to the four candidates appearing in Fig. 6.2.

What does the landscape of the search look like? The landscape—represented in three dimensions in Fig. 6.4—is determined by the difference in the violation profiles of the *neighbouring* candidates. As this difference depends only on the highest ranked constraint distinguishing between the two profiles, phonologists can replace constraint STOPASSIMILATION=REGRESSIVE with STOPASSIMILATION=REGRESSIVE2, and both tableau (6.4) and the first rows of tableau (6.5) may be used. The global optimum is above [bd], and another local optimum, diagonally opposed to it, above [pt]. At the two ends of the other diagonal, [pd] and [bt] represent peaks.

Let us run simulations under the usual conditions. Temperature drops from above the highest constraint to much below the lowest constraint, so that enough time is given both to walk freely around the search space initially, and to find the local optimum (relax) finally. The domains containing the constraints follow

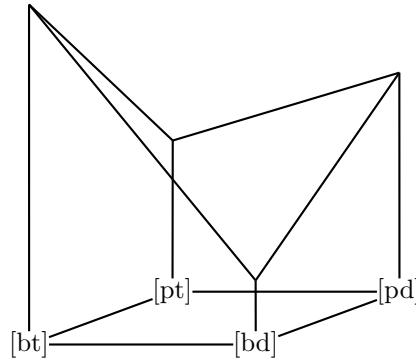


Figure 6.4: **3-D landscape of the first model for voice assimilation**: more harmonic candidates are drawn lower than the less harmonic ones (searching for the optimum corresponds to searching for the minimum). Candidate [bd] is the global optimum, [pt] is another local optimum, whereas [bt] is the least harmonic candidate.

each other, $K_{step} = 1$, and the real part of the temperature drops from $T_{max} = 3$ to $T_{min} = 0$ in equal steps T_{step} .

Based on our previous experience, we would predict Simulated Annealing Optimality Theory to produce the following behaviour: with slow simulation, the global optimum [bd] is easily found, whereas accelerated simulation may be stuck in the erroneous local optimum [pt]. The faster the simulation, the more frequently [pt] is expected to be returned—according to our intuition.

How very wrong we are! Implementing the model shows that [bd] and [pt] are returned with equal probability, 50% each, with some random dispersion. This happens so independently of the parameter setting!

The reason for this surprising result is easy to understand, and lies in the symmetry of the landscape. We face a case similar to those discussed in section 2.3.2. In fact, there is a greater symmetry than one would initially suppose. On the one hand, the chance of leaving the two local optima are equal at every moment of the simulation. To leave either of them, temperature has to allow violating ASSIMILATE[VOICE] once, which is possible at high temperatures, impossible at low temperatures, and has a chance between 0 and 1, when temperature is exactly in the domain of this constraint. The fact that [pt] violates a lower ranked constraint not violated by [bd] does not influence anything. On the other hand, from [pd] and [bt] both local optima are chosen with an *a priori* probability of 50%. Once chosen, the random walker moves always, both to [bd] and [pt], with a transition probability of 1. In turn, nothing guarantees that the random walker will prefer moving to [bd] to moving to [pt] from candidates [pd] and [bt]. In brief, both local optima are reached with equal probability and are left with equal probability—independently of the parameters of the simulation.

How could we break the symmetry of the search space just described, which results in the two local optima being found with equal probability? A first idea might be to increase the probability for the random walker to move from [pd] and [bt] to [bd], and to decrease the chance of moving to [pt]. The second idea will be to enlarge the search space in an asymmetric way, as will be demonstrated in the second model.

As both [bd] and [pt] are more harmonic than [pd] and [bt], it would con-

tradict the idea of simulated annealing not to have the transition probability $P(w \rightarrow w') = 1$, once a neighbour w' of $w = [\text{pd}]$ or $w = [\text{bt}]$ has been chosen. So, one may alter rather the *a priori* probabilities determining the choice of the neighbours. So far, each neighbour has been chosen with an equal probability, but this symmetry can be deformed *ad hoc*. One may also reconsider simulated annealing, and connect the horizontal structure of the landscape to the vertical one, although it is quite unclear to me how this could be done in the general case. This direction would involve taking the possible gain in the harmony function (the vertical structure) into consideration when determining the *a priori* chance to pick a neighbour (the horizontal structure): the more you can gain in harmony, the more it is probable that you will consider the possibility to move to this neighbour.¹²

In any case, experiments (for $p = 0.67$ and $p = 0.8$) have shown that this direction is still fruitless. If the *a priori* chance of picking $[\text{bd}]$ (as opposed to choosing $[\text{pt}]$) when the random walker is in $[\text{pd}]$ or $[\text{bt}]$ is increased, say, to $p = 2/3$, then the random walker will indeed prefer moving to $[\text{bd}]$. In the next step, however, leaving $[\text{bd}]$ has still the same probability as leaving $[\text{pt}]$. In general, if the probability of moving from $[\text{pd}]$ or $[\text{bt}]$ to $[\text{bd}]$ is p , and to $[\text{pt}]$ is q (with $p + q = 1$), then the simulation will return $[\text{bd}]$ with probability p and $[\text{pt}]$ with probability q —independently of the cooling schedule! Such a model can describe empirical data with a distribution different from 50% - 50%, and yet the interpretation of why p and q have some specific values would still be missing. Likewise missing is the interpretation explaining how and why these parameters of the horizontal structure are tuned by different speech situations.¹³

What would lead to success is a model in which the random walker is less likely to leave $[\text{bd}]$ than to leave $[\text{pt}]$ —at least, in some phase of the simulation. Once in $[\text{bd}]$, it is captured there, while leaving $[\text{pt}]$ is still possible. In order to end up in $[\text{pt}]$, the system has to choose to move always back to $[\text{pt}]$ —and never to $[\text{bd}]$ —each time the system has escaped from $[\text{pt}]$. The slower the cooling schedule, the more often such a decision has to be made. If $[\text{pt}]$ is chosen with probability q , then a cooling schedule offering n such decisions would return $[\text{pt}]$ with a probability of q^n , and $[\text{bd}]$ with a probability of $1 - q^n$. A slower cooling schedule means a higher n , resulting in a lower q^n with a higher $1 - q^n$. As discussed in section 2.3.2, however, it is unclear how the difference of two violation profiles could be defined in order to have the system escape from $[\text{pt}]$ more easily than from $[\text{bd}]$. The last resort is obviously the introduction of a new, highly ranked constraint satisfied by $[\text{bd}]$ and violated by the other three candidates, so that once temperature has dropped below this new constraint, escaping from $[\text{bd}]$ is not possible any more, but escaping from $[\text{pt}]$ still has some chance.

¹²In the present case, for instance, moving from $[\text{pd}]$ or $[\text{bt}]$ to $[\text{pt}]$ can be seen as an “improvement of one constraint level”, because the highest violation mark incurred is assigned by constraint `STOPASSIMILATION=REGRESSIVE[VOICE]` instead of `ASSIMILATE[VOICE]`. Similarly, moving from $[\text{pd}]$ or $[\text{bt}]$ to $[\text{bd}]$ is an “improvement of two constraint levels”, for $[\text{bd}]$ ’s highest violation mark originates from `FAITH[VOICE]`. The *a priori* chance to pick a neighbour with an improvement of two constraint levels may, in turn, be assigned double weight, as opposed to the chance of picking a neighbour with an improvement of only one constraint level.

¹³In general, how should we interpret the tuning of the probabilities related to the *horizontal* structure of the landscape? Nevertheless, by supposing that the horizontal structure may be slightly different for each individual (yet constant within a person), we can account for variations among speakers, dialects or sociolects.

To summarise, this model is analogous to the situations described in subsection 2.3.2 discussing the cases where SA-OT does not work. The present model with our four candidates is a variant of that basic situation. If traditional simulated annealing converges asymptotically, but SA-OT does not, should we conclude that simulated annealing has not been applied properly? The discussion in Chapter 3 aimed at demonstrating that peculiarities of SA-OT follow directly from the core of Optimality Theory. Thus, it is only to be hoped that the divergence between simulated annealing and SA-OT can be interpreted within linguistics and OT.

Consequently, I have a good and a bad bit of news: which do you want to hear first? The bad news is that this approach can only produce an equal distribution of the two forms, which is too strong a prediction. It is quite unlikely that empirical research would report on an exactly 50%-50% distribution. Stochastic approaches—and simulated annealing is one of them—aim at reproducing quantitative phenomena; why shall we content ourselves, then, with the qualitative result that both candidates can be reproduced? Just as in good-news-bad-news jokes, however, the good news will also resolve the bad news.

Now, the good news. Well, this sounds initially also as a bad news: we have to give up our expectations about the precision of SA-OT converging to 1, as simulated annealing is performed slower. And yet, this is a good piece of news. Optimality Theory Simulated Annealing is claimed to be a performance model on top of OT as a competence model (Table 2.1), and we know that performance is indeed always full of errors. Why do we actually expect it to be precise asymptotically, then?

Being even more radical, I suggest reformulating some basic ideas in linguistics. So far, phenomena independent of external factors (such as speech rate) were supposed to belong to competence, to the core of linguistic knowledge deeply encoded in the brain (or, at least, in the physiology of the speech production-perception system). However, many phenomena may steadily persist in language, even if they “contradict” the (static) mental representation of the given language, because they are necessarily introduced by the (dynamic) computational production process. In the present case, even though competence in its narrow sense would require regressive assimilation (hence, its model, the OT grammar, yields exclusively [bd] as optimal); and yet, the computational production process, modelled by SA-OT, cannot help but also return the [pt] form displaying progressive assimilation.

The fact that the ratio of the “erroneous” form is constant and does not depend on speech rate makes it impossible to argue *a priori* for a certain form to be the performance error. Earlier, namely, we could identify the form whose frequency increases in fast speech as the performance error, based on the assumption that fast speech cannot be more correct than normal speech. Our aim was to reproduce this behaviour using SA-OT. Now, however, it is only the model that turns a certain form into the grammatical form (by having it as the globally optimal candidate), and other forms as performance errors (local optima), and not pre-theoretical observations. The only hint was that the form with progressive assimilation seemed to be an exception from the general trend displaying only regressive assimilation.

I am convinced that models of the mental representation of languages could be kept simpler if many “ugly cases” were exiled to the production-computation process. The present example has shown us the way: without reformulating

Dutch phonology, we could reproduce the exceptional progressive assimilation in *o[pt]ie*.

Obviously, the question arises why the same pressures do not apply to *zakdoek* ('handkerchief') or to *duikboot* ('submarine'). This brings us back to the bad news: if we were able to modulate frequencies by tuning the parameters, we could simply argue that the unaccented frequent function words constituting *op die* are produced much more quickly—that is, with a different parameter setting—than relatively infrequent nouns such as *zakdoek* and *duikboot*. This is why we have to confront the second model, whose parameters will influence again the output frequencies.

The last good piece of news then is that the second model and the mathematical challenges posed by its formal analysis (which can be safely skipped by the reader less interested in math) will turn out to be illuminating about the techniques offered by SA-OT.

6.5 Model 2: Infinite search space

6.5.1 Enlarging the search space

The second model involves enriching the search space with new candidates, and in this way breaking its symmetry. The candidate set becomes huge—actually infinite. The four candidates of the previous model (Fig. 6.2) form but the central zone of the new search space (already advanced in Fig. 6.3). As the periphery of the latter does not exhibit the same symmetry as the centre, the two local optima may be returned with probabilities different from 50% each. The more use the system makes of the periphery, the more significant the difference from the 50%-50% distribution will be.

Importantly, the periphery will be less optimal than the central valley, and therefore we can get farther in the periphery only in the *first phase* of the simulation. This is, when temperature is still higher than the highest ranked constraint. In other words, to distance the system from the 50%-each distribution, we have to allow many iterations in the first phase. Hence the novelty of this model: unlike in the different uses of SA-OT so far, all of which included but a finite search space, the parameter K_{max} is assigned now a leading role.

Parameter K_{max} is starring in the present model also for a second reason, which is similarly related to the fact that the candidate set is infinite. Due to this fact, we cannot launch the simulation from each of the candidates with equal probability, as we have done before. One option would be to define a probability distribution on the candidate set; but we leave this option open to future research, and we rather launch the simulation always from one of the four candidates in the central basin. This is why K_{max} will determine how far from the central basin the random walker can get, and thereby, how much of the asymmetry of the search space's external regions we can make use of.

After this introduction, the question is raised: how can we enrich the search space? A straightforward direction, copying the classical paradigms in OT, is to allow epenthesis: let us insert an epenthetic vowel (a schwa) between the two consonants. Indeed, vowel epenthesis is frequently employed by natural languages to break up unwanted consonant sequences, even if not necessarily to

resolve clashes in voice.¹⁴

The possibility of inserting only one schwa has not proven to be fruitful. Inserting any number of schwas recursively is more interesting (Fig. 6.3 depicting the structured candidate set is repeated here as Fig. 6.5). This is so even if forms with more than one epenthetical vowel are—most probably—not attested in any language.

This paradox, namely the fact that forms not attested in natural languages render the model fruitful, is worth emphasising here. Following Bíró and Gervain (2006), we could call this phenomenon the “*Bald Soprano*” effect, or even the *Godot effect*: there are characters in the play who never appear openly on the scene, and yet, they influence importantly the whole story line. In fact, the present study refutes a possible criticism to Optimality Theory in general: why should a model include an infinite set of candidates, if not for the sake of simplicity and of mathematical beauty? OT’s main goal is to account for linguistic typology and typologies include only a very restricted number of types, whence one would expect a very restricted finite candidate set. Do the candidates that can never win (the *losers* according to Samek-Lodovici and Prince, 1999) play any role in Optimality Theory at all? We shall see presently that they do, at least in SA-OT.

6.5.2 The landscape

After such a long introduction, let us enter the linguistic details of the new model. The infinite candidate set and its topology have already been defined in section 6.3, so we turn our attention to the constraints.

Now the constraint STOPASSIMILATION=REGRESSIVE in its first formulation will play a role. Recall tableau (6.5) in section 6.3, repeated here below. All forms with an epenthesis violate DEP (as many times as the number of epenthetical vowels included), and satisfy ASSIMILATE[VOICE]. The third most important constraint is STOPASSIMILATION=REGRESSIVE, which is satisfied by [b@⁺d] (a positive number of epenthetical vowels surrounded by [b] and [d]) and, crucially, violated by the other candidates with epenthesis.

¹⁴See footnote 1 on page 161. For a concrete example of schwa insertion, consider Modern Hebrew. (Notice that the word “schwa” originates from the concept of *schwa mobile* coined by the Biblical Hebrew grammarians.) A clash occurs when the past tense singular 2nd masculine suffix [-ta] is added to a verb ending in a [d], such as *lamad* ‘learn, study’. In such cases, two forms may emerge: the first one involves regressive assimilation ([lamatta]), while the second one inserts an epenthetic schwa ([lamad@ta]). In fact, this case serves as an example for the prohibition of homorganic consonant clusters in general [Schwarzwald (2001, pp. 11-12.); for further examples, see Bíró and Hamp (2002)]. Still, the behaviour of Modern Hebrew with respect to homorganic consonant clusters can be only described by using a candidate set that includes forms with epenthetical vowels, as well as by constraint DEP overruled by a markedness constraint *_[αPLACE][αPLACE]. In sum, Modern Hebrew—among, most probably, a huge number of further languages—does support the need to include the new candidates into the candidate set. If one requires such a support at all, as most phonologists view GEN as a black box generating literally “everything”.

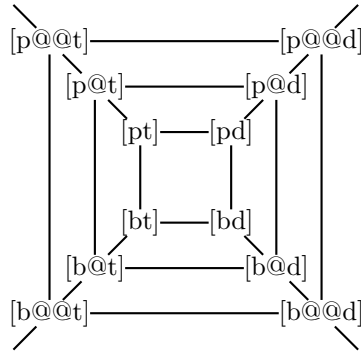


Figure 6.5: The second search space used for *op die*. The character @ stands for the epenthetic schwas.

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]	FAITH[VOICE]
[bd]				*
~ [pt]			*!	*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	* ⁿ !			*
[p@ ⁿ t]	* ⁿ !		*	*
[p@ ⁿ d]	* ⁿ !		*	
[b@ ⁿ t]	* ⁿ !		*	**
...

(6.6)

Consequently, the landscape looks as follows: in the “middle” we find a *central basin*, formed by the four candidates without epenthesis and having the form already discussed (in section 6.4, and especially in Fig. 6.4), which is in turn surrounded by ever rising hills (the higher, the less optimal). This picture is the result of promoting DEP, the constraint penalising recursion, to the highest position. Furthermore, this “radial” structure of the landscape is modulated by a “tangential” structure, to be presented soon in Fig. 6.6. In each concentric circle outside the central basin, [b@ⁿd] ($n > 0$) is lower (more harmonic) than [p@ⁿd], [p@ⁿt] and [b@ⁿt], due to constraint STOPASSIMILATION=REGRESSIVE. This search space can thus be visualised as a circular crater with a smaller radial valley formed by a river that runs down in a centripetal direction towards the central basin.

Our goal has been exactly to create this channel [b@ⁿd], and this is why we require the first definition of constraint STOPASSIMILATION=REGRESSIVE. Imagine the water falling on such a landscape, which sooner or later reaches some deepest valley in the landscape by flowing down the slope. The valley

collects most of the water and streams it to the basin in the form of a river or channel. Even though an initial rain has spread the water, say, uniformly in a larger region, still water will concentrate more and more in the river, and later in the central basin, as time passes. Remembering this metaphor might help better understand the behaviour of our SA-OT system.

In the first, unhindered stage of the simulation, the freely roaming random walker may be found, more or less, everywhere in the landscape. The likelihood $P_0(w)$ of the random walker being at a certain point w of the search space by the end of this first phase (we will be calculated exactly later) resembles the quantity of water in w after the initial rain. The dispersion is not necessarily even (that is, $P_0(w_1) = P_0(w_2)$ does not necessarily hold for any w_1 and w_2), but “smooth”. Additionally, the total amount of water is unity, corresponding to the fact that $\sum_{w \in \text{Gen}(UR)} P_0(w) = 1$ must hold.

Now the water starts flowing; that is, the probability distribution $P_t(w)$ of the random walker being in w changes as time t advances in each time step of the simulation. Obviously, the total amount of water ($\sum_{w \in \text{Gen}(UR)} P_t(w)$) remains the same over time. As temperature reaches the domain of the highest ranked constraint, DEP, not all moves are equally likely anymore. In particular, centrifugal moves increasing the number of the epenthetic vowels become blocked in this stage. Once moving upwards in the landscape becomes difficult for the random walker, the water—the probability $P_t(w)$ —will be collected and streamed to the central basin by the structure of the landscape, and especially by channel [b@⁺d]. By the end, $P_\infty(w)$ (the “amount of water collected in w ”) gives you the probability of the algorithm returning candidate w : usually 0, unless w is a local optimum.

How does channel [b@⁺d] work? Suppose the random walker is “out in the hills”, that is, not in the central basin, when temperature drops to the domain of the constraint STOPASSIMILATION=REGRESSIVE. At this moment, some tangential moves—moves changing the [voice] feature of the stops—are not free anymore either: the transition probability of stepping from [b@ⁿd] to either [p@ⁿd] or [b@ⁿt] becomes less than 1—and this probability quickly diminishes to zero—because such steps would require incurring a violation mark by this constraint. In turn, [b@ⁿd] serves as a trap for the tangential component of the random walk. The “water” is collected by channel [b@⁺d] during the tangential steps, and the channelled water has no other option but to flow towards the central basin through a series of centripetal steps (deleting the epenthetic vowels, but not altering the voiced feature of the consonants).

Now, the clue to this model is the fact that this channel enters the central basin at [bd]. This is crucial, since all the “water” (probability of the random walker being there) channelled by the river or channel will be stuck in [bd], cannot end up in [pt], for [bd] is a local optimum. The *channelling effect* starts when temperature falls to the domain of STOPASSIMILATION=REGRESSIVE. At this stage of the simulation, escaping from the two local optima is not possible anymore, because escaping would require incurring a violation mark by ASSIMILATE[VOICE], which is higher than the actual temperature.

On the other hand, the water that has not been collected by the channel may end up in [pt], which is also a trap, a local optimum, due to tableau (6.6). The water reaching the basin in [pt] from [p@t] gets caught there; whereas the water arriving into [pd] and [bt] (from [p@d] and [b@t]) is equally divided between

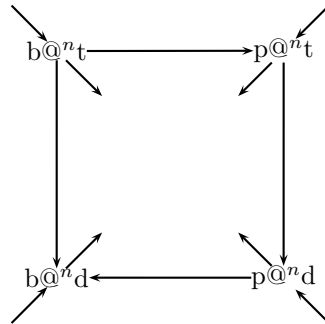


Figure 6.6: “**Channelling**” effect in the infinite search space, in the n th ($n > 0$) concentric layer, which is formed by the candidates with n epenthetical vowels. The arrows point to the more harmonic candidate, based on tableau (6.5). If temperature is below the lowest ranked constraint, the random walker can move exclusively towards the centre, or towards candidate $b@^n d$, the most harmonic candidate among the candidates with n epenthetical vowels. In this sense, we can speak of the $b@^n d$ valler or channel.

the two local optima, as explained in the context of the first model.

6.5.3 Tuning the output of the model

Consequently, moving away from the 50%-50% distribution between the two local optima of the four-candidate model is possible by channelling as much “water” as possible into channel $[b@^+d]$, and thereby increasing the probability of our simulation returning $[bd]$. Notice that decreasing it is not feasible in our model. Nonetheless, it is more likely that accounting for empirical data would require increasing the probability of $[bd]$ rather than decreasing it.¹⁵

What we need to do then is to disperse initially the constant amount of water on a region as large as possible, by providing the walker the possibility to move much without any obstacles (i.e., a long initial phase in the simulation). Why is this so? In short, once the temperature has reached the domain of DEP, centrifugal moves become prohibited, and a competition starts between centripetal and tangential moves.

The competition is about the water reaching the channel first or the basin first. If the channel is reached, $[bd]$ will certainly be the output, otherwise $[bd]$ and $[pt]$ have equal chance. The more “water” reaches the channel, the higher the likelihood of $[bd]$. Additionally, observe that from a larger distance, more centripetal steps are required to arrive at the central basin, which increases the chance to reach the channel first by performing a few tangential steps.

In sum, the farther the random walker is from the central basin at the end of the unhindered phase, the more likely it is for $[bd]$ to be returned by the algorithm. This is the technique we can have our SA-OT model returning the two outputs with different probabilities.

¹⁵Decreasing the probability of $[bd]$ is possible by using constraints that define a very similar landscape but with a channel $[p@^+t]$. Then, the more water that is channelled, the higher the frequency of output $[pt]$. Observe that such a model is possible even if $[bd]$ is the global optimum, and not $[pt]$. In other words, here we see an example of an SA-OT model with the global optimum being returned in less than or equal to half of the cases.

At this point an additional issue arises, the choice of the initial candidate. We must remember that the search space is infinite, unlike in our previous models. So far, we could choose each candidate with equal probability to be the starting point of the random walk, but now, we have to come up with a different solution. For instance, a Gaussian-style distribution could be defined so that the likelihood of a candidate w_0 with n epenthetical vowels ($[C_1 @^n C_2]$) being the initial candidate of the walk be proportional to $e^{-n^2/2\sigma^2}$. Then, a larger σ will disperse the “rain” over a wider region, resulting in a higher frequency of output [bd], due to the channelling effect.

Instead of introducing an additional parameter σ to the model, however, we rather leave this idea to future research, and prefer exploiting the already existing parameters of the SA-OT Algorithm. Probably the most natural choice is to employ exclusively the four basic candidates of the central basin ([pd], [pt], [bd] and [bt]) as initial candidates, with 25% chance each. In turn, having “the initial rain covering a wide region” corresponds to allowing the random walker to get really far away from the initial candidate in the first, unhindered phase of the simulation. Using the water metaphor, the water is poured in the four central candidates, before it splashes to the initially unhindering mountains.

Then, a last point remains to be clarified in our train of thought: the way we lengthen the initial phase of the simulated annealing, that is, the phase in which the random walker moves freely even to worse neighbours. From the parameters of the algorithm (see Fig. 2.8 on page 64), two are the most straightforward candidates: K_{max} and T_{step} . In other words, we either add extra upper domains that the temperature has to traverse before reaching the domain of the highest ranked constraint (increase K_{max}); or increase the number of steps to be performed within each domain in order to have more steps also in the domain(s) superior to the domain of the highest ranked constraint. The second strategy can be realised in the simplest way by decreasing T_{step} , and this is the technique we have used the most often so far. So, should we increase K_{max} or decrease T_{step} , if we would like to have more iterations in the first, unhindered phase of the simulation?

In contrast to our previous models, it turns out that simply decreasing T_{step} does not work now.¹⁶ The reason, in short, is that increasing the number of steps within one domain will also increase the number of steps while temperature is between the domains of DEP and STOPASSIMILATION=REGRESSIVE. The reason, in detail, requires some mathematical discussion, which can be skipped without losing the general train of thought of my dissertation.

6.5.4 The interaction of K_{max} with T_{step}

The present subsection aims at presenting a formal analysis of how the parameters of the model influence the probabilities of returning [pt] and [bd]. First,

¹⁶Note this major difference between the present case and stress assignment in fast speech. For stress assignment, increasing K_{max} alone would not have any effect: the candidate set is finite, and not only can the random walker rove around the whole search space in the initial phase of the simulation, but also each point has an equal chance to be the starting point of the random walker. The phenomenon arises from changing the number of steps in the *second* phase of the simulation, that is, when temperature has already reached the domains of the constraints. In the present case, however, the search space is infinite, we start the simulation from a small subset of it (the four central candidates), and the goal is to have the random walker also visit candidates as remote as possible.

let us introduce a few notations:

$$\tau = \left\lfloor \frac{T_{max} - T_{min}}{T_{step}} \right\rfloor + 1 \approx \frac{T_{max} - T_{min}}{T_{step}} = \frac{3}{T_{step}} \quad (6.7)$$

$$k = K_{max} - K_{highest} = K_{max} - 3 \quad (6.8)$$

where $K_{highest}$ is the index associated with the highest ranked constraint, 3 in the present case. Further, τ stands for the number of repetitions performed by the inner loop of the SA-OT algorithm, that is, the number of iterations while temperature decreases one domain. Here, we employ our standard values, $T_{max} = 3$ and $T_{min} = 0$, and $\lfloor x \rfloor$ represents the integer part of x .

K_{max} is located k domains above constraint DEP, so temperature traverses k domains in the first phase of the simulation. In the period when temperature is exactly in the domain of DEP, centrifugal moves become banned only gradually: in the beginning of this period, they are almost free, and later almost impossible. Let us approximate this gradual effect by supposing that the random walker can freely move away from the centre as long as the temperature crosses the first $k + 0.5$ domains, and this direction becomes maximally prohibited immediately afterwards, from that point onwards when temperature enters the lower part of the domain of DEP. Consequently, the number of steps performed by the random walker in the first (unhindered) phase of the simulation is:

$$N = (k + 0.5) \cdot \tau \quad (6.9)$$

Remember that a candidate $[C_1 @^n C_2]$ (with $n > 0$) has four neighbours, two in a tangential direction (that is, also including n epenthetical vowels: $[C_1' @^n C_2]$ and $[C_1 @^n C_2']$), and two in a radial direction ($[C_1 @^{n+1} C_2]$ and $[C_1 @^{n-1} C_2]$). As each neighbour has an equal *a priori* probability of 0.25, the number of radial steps among these first N steps can be approximated by

$$N_{radial} = N_r \approx \frac{N}{2} = (k + 0.5) \cdot \frac{\tau}{2} \quad (6.10)$$

Estimating $\pi_N(n)$

Now, we calculate the probability $\pi_N(n)$ of being exactly at a distance n from the central valley by the end of the first phase, that is, of starting the “competition” from some candidate with exactly n epenthetical vowels ($[C_1 @^n C_2]$). The radial component of this first phase is a one-dimensional *Brownian motion* with equal probability of moving in both directions (centripetal and centrifugal). One flips a symmetrical coin N_r times, with head corresponding to the insertion of a @, and tail to the deletion of a @. Ending up with n epenthetical vowels requires exactly $\frac{N_r + n}{2}$ heads and $\frac{N_r - n}{2}$ tails, supposing that N_r and n have the same parity. Consequently, $\pi_N(n)$ can be approximated with a binomial distribution:¹⁷

¹⁷Observe that in each prefix of the insertion-deletion (head-tail) sequence, the number of deletions must not exceed the number of insertions, as we launch our algorithm from a candidate with no epenthetical vowel. Yet, we can overcome this problem by employing a trick. Observe that when the number of epenthetical vowels is zero, we still may flip our coin, but both head and tail should correspond to insertion, with deletion having zero probability. So, if flipping the coin returns then tail, we reverse the roles of heads and tails: from now on

$$\pi_N(n) = \begin{cases} 0 & \text{if } n \not\equiv N_r \pmod{2} \\ \binom{N_r}{N_r/2} \cdot 0.5^{N_r} & \text{else if } n = 0 \\ 2 \binom{N_r}{(N_r+n)/2} \cdot 0.5^{N_r} & \text{else} \end{cases} \quad (6.11)$$

Using the basic properties of the binomial coefficients, one can quickly check that $\sum_n \pi_N(n) = 1$. What we shall need is actually the sum of $\pi_N(n)$ over a large range of its argument n in function of N (i.e., of N_r), so we can render our life simpler by “smoothing” $\pi_N(n)$ (dividing the probabilities among $\pi_N(2k)$ and $\pi_N(2k \pm 1)$):

$$\begin{aligned} \pi_N(n) &= \begin{cases} \binom{N_r}{(N_r+n)/2} \cdot 0.5^{N_r} & \text{if } n \equiv N_r \pmod{2} \\ \binom{N_r}{(N_r+n+1)/2} \cdot 0.5^{N_r} & \text{if } n \not\equiv N_r \pmod{2} \end{cases} \\ &\approx \frac{1}{\sqrt{2\pi}} \frac{2}{\sqrt{N_r}} e^{-\frac{n^2}{2N_r}} \end{aligned} \quad (6.12)$$

Here, we have employed the well-known fact that a binomial distribution ($p = q = 0.5$ in our case) can be approximated with a normal distribution, namely¹⁸

$$\binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{npq}} \varphi\left(\frac{k - np}{\sqrt{npq}}\right) \quad (6.13)$$

for large n , where $\varphi(x)$ is the standard normal distribution:

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (6.14)$$

Equation (6.12) makes clear that $\pi_N(n)$, the position of the random walker by the end of the first phase of the simulation, follows approximately the positive part of a Gaussian distribution, centred around the origin, with a standard deviation¹⁹

head will correspond to deletion and tail to insertion. Therefore, each of the 2^{N_r} different head-tail series can be made a legitimate one, and each of them has an equal probability.

This more complicated interpretation of heads and tails can be visualised if head is still seen as always moving one unit to the positive direction of the scale, and tail as moving one to the negative one; but we allow moving to both directions of the origin, with both positions $+n$ and $-n$ corresponding to n insertions in the candidate string. Indeed, once the coin returns tail when no deletion is possible, we move from 0 to -1 on the scale, which corresponds to an insertion (-1 also meaning one insertion) together with the reversion of the roles of heads and tails from the viewpoint of the number of epenthetical vowels (e.g. a second tail would bring to -2 , i.e. to a second insertion).

As arriving at both $-n$ and $+n$ corresponds to n insertions, $\pi_N(n)$ has to be multiplied by 2 if $n \neq 0$, as compared to the standard binomial distribution. In short, the two halves of the scale are folded around the origin.

Note finally that this train of thought would work correctly only if the *a priori* probabilities corresponding to the elements of the central valley had been defined in a slightly different way. Namely, by assigning a 50% chance to insertion, and 25% chance to changing the [voice] feature of one of the stops, similarly to the *a priori* probabilities of the other candidates. Now that each of the three neighbours has an equal probability of 1/3, the following formula is but an approximation.

¹⁸See e.g. <http://mathworld.wolfram.com/NormalDistribution.html>.

¹⁹It would have sufficed to refer to the fact that in a *Brownian motion* the expected value of the squared displacement is proportional to the number of steps performed. Let us take a

$$\sigma = \sqrt{N_r} = \sqrt{(k + 0.5) \cdot \frac{\tau}{2}} \tag{6.15}$$

Estimating n_0

In what follows, we estimate the distance n_0 beyond which channelling is expected to take place. If the random walker has reached this distance by the end of the first phase, then it will most probably end up in candidate [bd]; otherwise, it has an equal chance to return us [pt] or [bd]. Thus, once n_0 has been estimated, we predict candidate [pt] to be returned with probability

$$\mathcal{P}([pt]) = \frac{1}{2} \sum_{i=0}^{n_0} \pi_N(i) \approx \int_0^{n_0} \frac{1}{\sqrt{N_r 2\pi}} e^{-\frac{t^2}{2N_r}} dt \tag{6.16}$$

Let us repeat here tableau (6.6):

/pd/	DEP	ASSIM[VOICE]	STASS=RGR[VC]	FAITH[VOICE]
☞ [bd]				*
~ [pt]			*!	*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	* ⁿ !			*
[p@ ⁿ t]	* ⁿ !		*	*
[p@ ⁿ d]	* ⁿ !		*	
[b@ ⁿ t]	* ⁿ !		*	**
...

(6.17)

Imagine that temperature is just crossing the domain of ASSIMILATE[VOICE], and the random walker is located somewhere in the epenthetical hills. Say, at [C₁@ⁿC₂]. The four neighbours ([C'₁@ⁿC₂], [C₁@ⁿC'₂], [C₁@ⁿ⁺¹C₂] and [C₁@ⁿ⁻¹C₂]) are chosen with equal probability. The centrifugal move (inserting an extra @) is prohibited for $T \ll \text{DEP}$, whereas choosing the centripetal step (deleting one @, which is always possible) results in bringing you towards the central basin with transition probability 1. The two other neighbours are chosen *a priori* with 0.5 chance, and moving in this tangential direction is still fully free. Since temperature is high, channelling does not occur. A race with time starts: the centripetal steps performed in the approximately 1/4 of the

one-dimensional random walk (*Brownian motion*) starting from $x_0 = 0$, with p and q being the probabilities of stepping one unit to the right ($x_{i+1} = x_i + 1$) and to the left ($x_{i+1} = x_i - 1$) respectively. The expected value of the location of the walker after N steps is $\overline{x_N} = N(p - q)$, whereas the dispersion is $(x_N - \overline{x_N})^2 = 4Npq$. See for instance Hubbey (1999, p. 229) and references therein, or <http://scienceworld.wolfram.com/physics/BrownianMotion.html> and references there. A very creative derivation is found in Reif (1965, pp. 13-16). In our case, $p = q = 0.5$.

iterations should not bring you back to the central basin until temperature has reached constraint `STOPASSIMILATION=REGRESSIVE` so that channelling can be effective.

As an approximation, let us say that there are 2τ iterations—while temperature drops from the middle of the domain of `DEP` to the middle of the domain of `STASS=RGR[VC]`—during which centrifugal moves are prohibited, but tangential and centripetal moves are free. On average, a quarter of these time steps are used to bring us closer to the central basin. This is why the random walker has to reach a distance larger than $n_1 = \tau/2$ by the end of the first phase, if it should not be probable for the random walker to reach the central basin until channelling is effective (until temperature reaches constraint `STASS=RGR[VC]`).

Once the `[b@nd]` channel becomes visible, a few more steps are still needed for the random walker to reach it, and not the central basin. The expected number of tangential steps for the random walker to reach the channel is $\kappa = 2.5$, which is slightly altered whenever constraint `FAITH[VOICE]` becomes also active.²⁰

While the system performs κ tangential steps, it also tries—on average— $\frac{\kappa}{2}$ centrifugal steps (in vain), and performs $\frac{\kappa}{2}$ centripetal steps. In other words, if the random walker has been not farther than $n_2 = \frac{\kappa}{2}$ from the central valley, there is a chance of reaching the central valley at a random point (that is, yielding outputs `[bd]` and `[pt]` with equal chance), and not through channelling. If, however, the random walker has been farther away, the random walker will have reached the `[b@nd]` channel, before entering the central valley in `[bd]` *due to* the channelling effect.

In sum, at the end of the first phase the random walker has to reach at least a distance of

$$n_0 = n_1 + n_2 = \frac{\tau}{2} + \frac{\kappa}{2} + 1 \quad (6.19)$$

for the channelling effect to take place significantly (remember $\tau = \frac{3}{T_{step}}$ and $\kappa = 2.5$). Even after having deleted n_1 epenthetical vowels while $T \approx \text{ASSIM}[\text{VOICE}]$, and having lost subsequently n_2 epenthetical vowels while trying to reach the already visible channel, there must be at least one `@` left.

²⁰Suppose that temperature is such that constraint `STASS=RGR[VC]` already prohibits leaving `[b@nd]` (the channel acts as a trap), but constraint `FAITH[VOICE]` is not yet active to block some of the other tangential moves.

Let us focus now on the tangential component of the moves, by projecting the search space onto a circle of four candidates, `[bd]`, `[pd]`, `[pt]` and `[bt]` (or `[b@nd]`, `[p@nd]`, `[p@nt]` and `[b@nt]`). Supposing that the random walk in this small space is free, but `[bd]` is a trap, how many steps are required on average for the walker to get stuck in `[bd]`?

Let k_w be the expected number of tangential steps that is required to reach `[bd]` from candidate w . From `[b@nt]` we either move to `[b@nd]` (1 step required to reach the channel; with probability 0.5), or we move to `[p@nt]` ($1 + k_{[pt]}$ steps required to reach the channel; with probability 0.5). Similarly for the other candidates, which yields the following equations:

$$\begin{aligned} k_{[bd]} &= 0 \\ k_{[bt]} &= 0.5 \cdot 1 + 0.5 \cdot (1 + k_{[pt]}) \\ k_{[pd]} &= 0.5 \cdot 1 + 0.5 \cdot (1 + k_{[pt]}) \\ k_{[pt]} &= 0.5 \cdot (1 + k_{[pd]}) + 0.5 \cdot (1 + k_{[bt]}) \end{aligned} \quad (6.18)$$

By solving these equations, we obtain $k_{[bd]} = 0$, $k_{[bt]} = 3$, $k_{[pd]} = 3$ and $k_{[pt]} = 4$. This is why $\kappa = 0.25k_{[bd]} + 0.25k_{[bt]} + 0.25k_{[pd]} + 0.25k_{[pt]} = 2.5$.

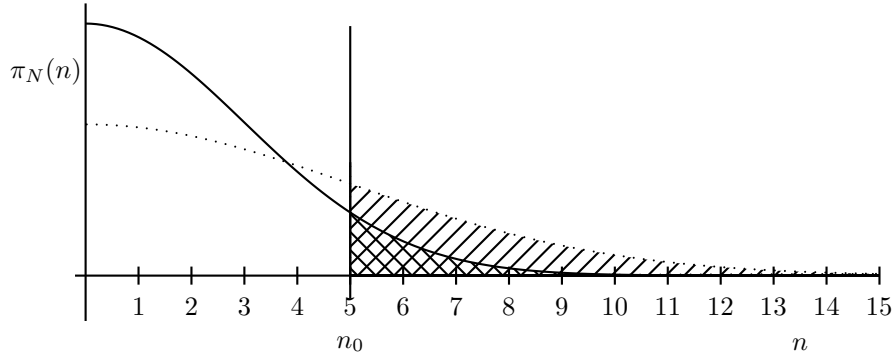


Figure 6.7: **Distribution** $\pi_N(n)$ of the random walker's position at the end of the first phase. The standard deviation of a distribution is $\sqrt{N_r}$. Here, the dotted distribution corresponds to a larger number of radial steps N_r (to a larger N) than the solid one. For a given n_0 , the chance of the random walker getting to a distance of at least n_0 increases as N_r grows larger.

If the random walker has reached this distance by the end of the first phase, the output will be most probably [bd]. If the random walker has not reached this distance by the end of the first phase, both [bd] and [pt] have equal chance to be returned. In brief, equation (6.19) defines the n_0 to be used in equation (6.16), which we repeat here:

$$\mathcal{P}([\text{pt}]) = \frac{1}{2} \sum_{i=0}^{n_0} \pi_N(i) \approx \int_0^{n_0} \frac{1}{\sqrt{N_r} 2\pi} e^{-\frac{t^2}{2N_r}} dt \quad (6.20)$$

Understanding the role of the parameters

Figure 6.7 helps us summarising what we have so far. We are interested in the impact of two parameters, namely K_{max} (or k , see equation (6.8)) and T_{step} (or τ , see equation (6.7)), on the output frequencies estimated by equation (6.20). This estimation includes two derived parameters, N_r and n_0 . According to equation (6.10), N_r depends on both K_{max} and T_{step} ; whereas n_0 depends exclusively on T_{step} by equation (6.19).

Thus, let us first fix T_{step} (hence, n_0), and consider the influence of K_{max} on the outputs. As Fig. 6.7 illustrates, a larger K_{max} (a larger k , a larger N_r) increases the chance of the random walker finishing up beyond the fixed n_0 (the curve has a thicker tail), thereby decreasing the probability of returning [pt] by equation (6.20). Experiments performed will support this prediction in the next subsection.

What happens if K_{max} is fixed and T_{step} varies? Our experience in earlier chapters has been that a larger T_{step} increases the probability of returning the suboptimal alternating form, [pt] in the present case. Will the same happen now, as well?

Let us transform equation (6.20) into the integral of the standard normal distribution by employing a replacement $u = t/\sqrt{N_r}$:

$$\mathcal{P}([\text{pt}]) \approx \int_0^{n_0} \frac{1}{\sqrt{N_r 2\pi}} e^{-\frac{t^2}{2N_r}} dt = \int_0^{n_0/\sqrt{N_r}} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du = \Phi\left(\frac{n_0}{\sqrt{N_r}}\right) \quad (6.21)$$

It becomes clear that the frequency of [bd] increases, that is, the frequency of [pt] decreases, if the argument of Φ —the integral of the standard normal distribution, a monotone increasing function—decreases; that is, if n_0 decreases and N_r increases. Increasing K_{max} and keeping T_{step} fixed is increasing N_r with n_0 kept unchanged. The influence of varying T_{step} (with K_{max} being constant), on the other hand, depends on the influence of T_{step} on the argument G of Φ :

$$G := \frac{n_0}{\sqrt{N_r}} = \sqrt{\frac{\tau}{2(k+0.5)}} + \sqrt{\frac{(\kappa+2)^2}{2\tau(k+0.5)}} \quad (6.22)$$

For small τ values, the second addend dominates, whereas for large τ , the first one does. As we increase τ (decrease T_{step}), the value of G will first decrease, and then, as the first addend turns dominant, G grows larger again. Decreasing G corresponds to decreasing the frequency of [pt] by equation (6.21), and increasing G brings the frequency of [pt] closer to 0.5.

By employing the fact that the geometrical mean is always less or equal to the arithmetic mean, we obtain ($\kappa = 2.5$):²¹

$$G \geq \sqrt{2 \frac{\kappa+2}{k+0.5}} = \frac{3}{\sqrt{k+0.5}} \quad (6.23)$$

and G is minimal iff $\tau = \kappa + 2 = 4.5$. That is, iff $T_{step} \approx \frac{3}{\kappa+2} = \frac{2}{3}$.

Notice, however, that by its definition, τ must be an integer (the number of steps in a domain), so in the case of our standard T_{max} and T_{min} values, we expect the turning point to be around $1 > T_{step} \geq 0.75$ (corresponding to $\tau = 4$). It will turn out that on the other side of the turning point—for T_{step} values corresponding to $\tau = 5$ ($0.75 > T_{step} \geq 0.6$)— G grows faster, so these parameters will produce more [pt] outputs than parameters corresponding to $\tau = 4$.

Another prediction is that for $T_{step} \ll 1$, when the second addend in (6.22) becomes negligible, different parameter settings will produce the same frequencies if $\frac{\tau}{k+0.5}$ (that is, if $T_{step} \cdot (k+0.5)$) is kept constant.

After such a long mathematical discussion, let us probe the pudding now!

6.5.5 Experiments

The results of a few experiments are summarised in Tables 6.1 and 6.2, as well as in Fig. 6.8. In each of the cases, one of the two parameters K_{max} and T_{step} is kept constant, while the other varies. For each parameter setting, an experiment consisted of running 100 000 simulations, that is launching the simulation 25 000 times from each of the four central candidates, and of calculating the frequencies of the outputs. By repeating this experiment two more times, we could also determine the mean and the $\sigma(n-1)$ error of the measured frequency. Finally, these tables also show the estimated frequencies based on equation (6.21).

²¹ $a + b \geq 2\sqrt{ab}$. Furthermore, $a + b = 2\sqrt{ab}$ if and only if $a = b$. For our purpose, take the two addends in (6.22) as a and b . This trick saves us calculating $\frac{\partial G}{\partial \tau}$.

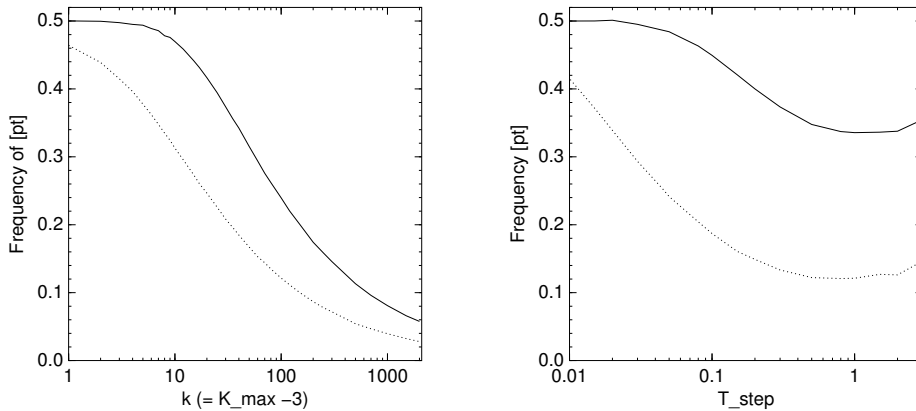


Figure 6.8: **Frequency of [pt] when varying either K_{max} or T_{step} .** The parameter varied is represented on a logarithmic scale. **Left box:** K_{max} changes, while $T_{step} = 0.05$ (solid line) or $T_{step} = 0.5$ (dotted line). Compare it to Table 6.1. **Right box:** T_{step} changes, while $K_{max} = 10$ (solid line) or $K_{max} = 100$ (dotted line). The frequencies are the same as those in Table 6.2.

The frequencies represented on the left panel of Figure 6.8, that is, in Table 6.1, confirm our prediction that the frequency of [pt] decreases as K_{max} (or k) grows larger, both for $T_{step} = 0.5$ and for $T_{step} = 0.05$. The curve corresponding to $T_{step} = 0.05$ runs higher than the one corresponding to $T_{step} = 0.5$. Not surprisingly, since the right panel of Figure 6.8 (that is, Table 6.2) demonstrates that the frequency of [pt] grows as T_{step} diminishes, supposing that $T_{step} < 0.8$. The turning point predicted by equation (6.23) can also be observed around $T_{step} \approx 1$, but we return to this point soon.

Subsequently, Fig. 6.9 presents the two dimensional *phase space*, that is, the behaviour of the system in function of both parameters. The radii of the circles are proportional to the difference of the frequencies of the two outputs. That is to say, the dots in the lower left corner correspond to the system returning [bd] and [pt] with (practically speaking) the same probability, whereas the large circles in the upper right corner visualise how [bd] becomes dominant. The largest circle is at $k = 27$ and $T_{step} = 0.8$, where the probability of [bd] reaches 79%.

One can also observe in Fig. 6.9 that circles of the same size are located, roughly speaking, on a diagonal straight line. As the figure uses logarithmic scales on both axes, such a diagonal straight line corresponds to a hyperbola on linear scales, and confirms our earlier prediction that for small T_{step} values keeping $T_{step} \cdot (k + 0.5)$ constant yields similar output frequencies.

In order to verify this observation in a more precise way, Table 6.3 presents a few parameter combinations that have been proven to yield [pt] with a chance of 0.25—that is, when channelling is effective in exactly half of the runs. This case corresponds to $G \approx 0.67$ by equation (6.21). If $T_{step} \ll 1$, the second addend in equation (6.22) becomes negligible:

$K_{max} - 3$	[pt] ($T_{step} = 0.05$)	<i>Pred.</i>	[pt] ($T_{step} = 0.5$)	<i>Pred.</i>
1	0.5001 ± 0.0014	0.5000	0.4641 ± 0.0031	0.4933
2	0.4996 ± 0.0018	0.4999	0.4389 ± 0.0013	0.4724
3	0.4976 ± 0.0013	0.4992	0.4149 ± 0.0016	0.4474
4	0.4950 ± 0.0030	0.4972	0.3963 ± 0.0019	0.4235
5	0.4938 ± 0.0004	0.4940	0.3776 ± 0.0003	0.4019
6	0.4889 ± 0.0022	0.4895	0.3621 ± 0.0020	0.3828
7	0.4858 ± 0.0003	0.4842	0.3478 ± 0.0015	0.3658
8	0.4782 ± 0.0024	0.4783	0.3348 ± 0.0016	0.3508
9	0.4758 ± 0.0008	0.4720	0.3244 ± 0.0030	0.3373
10	0.4697 ± 0.0014	0.4654	0.3128 ± 0.0003	0.3252
12	0.4585 ± 0.0006	0.4521	0.2955 ± 0.0001	0.3044
15	0.4417 ± 0.0010	0.4326	0.2737 ± 0.0004	0.2793
17	0.4314 ± 0.0019	0.4204	0.2605 ± 0.0011	0.2656
20	0.4165 ± 0.0009	0.4033	0.2464 ± 0.0008	0.2484
25	0.3943 ± 0.0011	0.3782	0.2258 ± 0.0022	0.2258
30	0.3737 ± 0.0014	0.3568	0.2084 ± 0.0006	0.2084
35	0.3562 ± 0.0024	0.3385	0.1960 ± 0.0009	0.1945
40	0.3423 ± 0.0021	0.3226	0.1848 ± 0.0019	0.1831
50	0.3152 ± 0.0014	0.2963	0.1671 ± 0.0001	0.1651
60	0.2940 ± 0.0013	0.2755	0.1529 ± 0.0017	0.1516
70	0.2758 ± 0.0013	0.2584	0.1440 ± 0.0003	0.1409
80	0.2621 ± 0.0010	0.2442	0.1349 ± 0.0003	0.1323
100	0.2395 ± 0.0012	0.2215	0.1214 ± 0.0007	0.1188
120	0.2203 ± 0.0006	0.2042	0.1115 ± 0.0012	0.1088
150	0.2003 ± 0.0013	0.1844	0.1000 ± 0.0004	0.0976
200	0.1742 ± 0.0012	0.1612	0.0867 ± 0.0003	0.0848
250	0.1583 ± 0.0001	0.1451	0.0778 ± 0.0001	0.0759
300	0.1458 ± 0.0011	0.1329	0.0718 ± 0.0006	0.0694
500	0.1132 ± 0.0004	0.1038	0.0542 ± 0.0001	0.0539
700	0.0960 ± 0.0010	0.0880	0.0468 ± 0.0004	0.0456
1000	0.0809 ± 0.0008	0.0738	0.0395 ± 0.0005	0.0382
1500	0.0660 ± 0.0006	0.0604	0.0324 ± 0.0001	0.0312
2000	0.0577 ± 0.0004	0.0524	0.0279 ± 0.0005	0.0270

Table 6.1: **Frequency of [pt] as a function of K_{max}** , while $T_{step} = 0.05$ (second column) and $T_{step} = 0.5$ (fourth column). Each frequency has been calculated by running 100 000 simulations trice. Error is $\sigma(n - 1)$. The figures in the first column are $k = K_{max} - 3$, that is, the number of strata above the highest ranked constraint. The third and the fifth columns show the estimations based on equation (6.21): the correspondence with the results of the experiments is often very good.

T_{step}	[pt] $K_{max} = 10$	$Pred.$	[pt] $K_{max} = 100$	$Pred.$
3	0.3549 ± 0.0025	0.4222	0.1466 ± 0.0007	0.1532
2	0.3377 ± 0.0023	0.3970	0.1262 ± 0.0009	0.1371
1.5	0.3363 ± 0.0018	0.3823	0.1269 ± 0.0005	0.1290
1	0.3356 ± 0.0013	0.3682	0.1212 ± 0.0008	0.1218
0.8	0.3372 ± 0.0022	0.3643	0.1207 ± 0.0006	0.1198
0.5	0.3476 ± 0.0008	0.3658	0.1222 ± 0.0015	0.1206
0.3	0.3735 ± 0.0023	0.3818	0.1334 ± 0.0012	0.1287
0.2	0.4000 ± 0.0009	0.4032	0.1489 ± 0.0020	0.1408
0.15	0.4208 ± 0.0011	0.4214	0.1608 ± 0.0018	0.1526
0.1	0.4495 ± 0.0018	0.4481	0.1870 ± 0.0013	0.1740
0.08	0.4632 ± 0.0005	0.4617	0.2043 ± 0.0010	0.1883
0.05	0.4842 ± 0.0023	0.4842	0.2415 ± 0.0012	0.2245
0.03	0.4951 ± 0.0008	0.4965	0.2934 ± 0.0010	0.2729
0.02	0.5011 ± 0.0007	0.4994	0.3390 ± 0.0018	0.3168
0.015	0.5001 ± 0.0008	0.4999	0.3716 ± 0.0014	0.3498
0.01	0.5000 ± 0.0008	0.5000	0.4152 ± 0.0015	0.3960

Table 6.2: **Frequency of [pt] as a function of T_{step}** ($[pt] \pm \sigma(n-1)$), while $K_{max} = 10$ and $K_{max} = 100$. Each frequency has been calculated by running 100 000 simulations trice. The third and the fifth columns show the estimations based on equation (6.21), not rarely matching the observed frequencies.

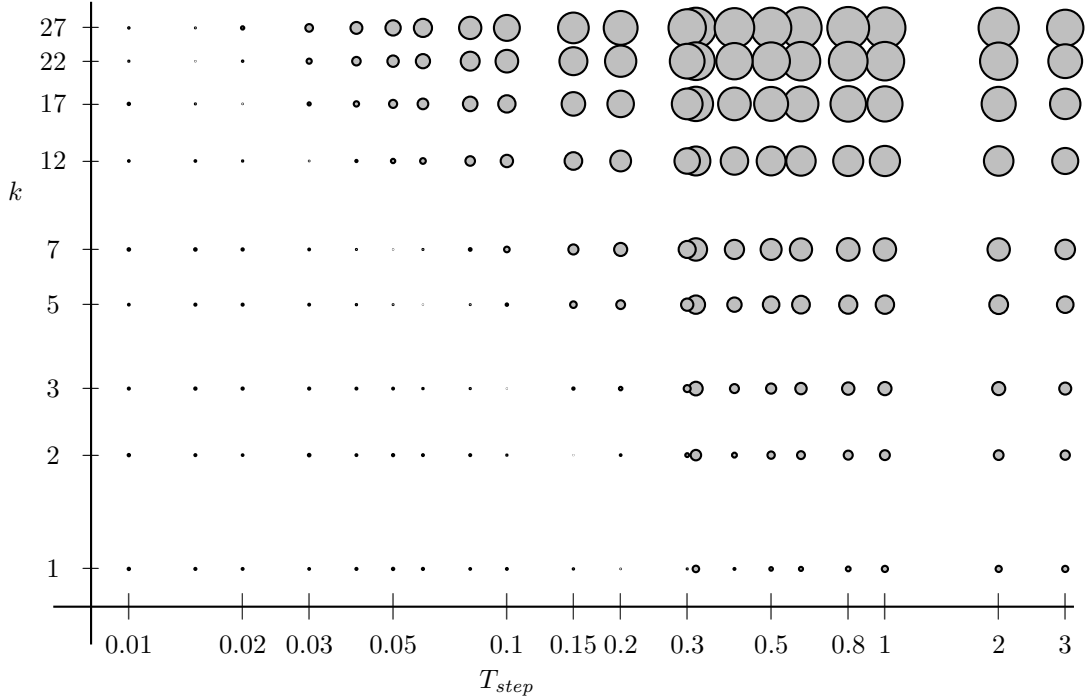


Figure 6.9: **The phase space:** the behaviour of the system in the function of the two parameters k (K_{max} minus the rank of the highest ranked constraint) and T_{step} , on a log-log scale. The radius of each circle is proportional to the difference of the probability of the two forms. Small dots represent (almost) 50%-50% distribution, whereas large circles correspond to [bd] dominating.

T_{step}	$K_{max} = k + 3$	Frequency [pt]	$T_{step} \cdot (k + 0.5)$	$\frac{T_{step}(k+0.5)}{(1+1.5T_{step})^2}$
0.0105	400	0.24996 ± 0.00379	4.17	4.05
0.0210	200	0.2493 ± 0.00725	4.15	3.89
0.0270	160	0.2501 ± 0.0029	4.25	3.93
0.045	100	0.2505 ± 0.0044	4.39	3.85
0.060	80	0.2495 ± 0.0029	4.65	3.91
0.105	50	0.2490 ± 0.0057	4.99	3.72
0.145	40	0.2506 ± 0.0043	5.44	3.67
0.220	32	0.2503 ± 0.0040	6.49	3.67
0.425	24	0.2503 ± 0.0038	9.14	3.408

Table 6.3: **Parameter settings producing [pt] with 25% chance:** (T_{step}, K_{max}) combinations that return candidate [pt] in 25% of the cases. We have noted that $T_{step} \cdot (k + 0.5)$ should be approximately constant for such parameter combinations. This prediction turns out to be correct only in a first approximation, and further factors (the second addend in equation (6.22) first of all) become gradually more important as T_{step} grows larger.

$$\begin{aligned}
G &= \sqrt{\frac{\tau}{2(k+0.5)}} + \sqrt{\frac{(\kappa+2)^2}{2\tau(k+0.5)}} = \\
&= \left(1 + \frac{\kappa+2}{\tau}\right) \cdot \sqrt{\frac{\tau}{2(k+0.5)}} = \\
&= (1 + 1.5T_{step}) \cdot \sqrt{\frac{\tau}{2(k+0.5)}} \approx \sqrt{\frac{\tau}{2(k+0.5)}} \quad (6.24)
\end{aligned}$$

If $G = 0.67$ and $\tau = 3/T_{step}$, then we predict $T_{step}(k + 0.5) \approx 3.34$. As Table 6.3 shows, the larger the value of T_{step} , the larger this product, for the second addend in (6.22) also contributes to G . Indeed, a better approximation following from equation (6.24) is that $\frac{T_{step}(k+0.5)}{(1+1.5T_{step})^2}$ must be constant.

Finally, let us check the prediction according to which there is a turning point in the frequencies around $T_{step} = 0.75$ if K_{max} is kept constant. Equation (6.23) gives a lower bound on G . Yet, the corresponding frequencies (that can be calculated by integrating equation (6.21) until this G value) cannot be reproduced, because G is predicted to be minimal for $\tau = \kappa + 2 = 4.5$, but τ is an integer. Consequently, we tried to find the minima in the frequencies of [pt] by varying T_{step} , for different K_{max} (Table 6.4). Since we required very accurate values, the number of iterations was very large: each piece of data in Table 6.4 originates from running 500 000 simulations trice in order to estimate also the error $\sigma(n - 1)$ of the frequencies.

The experiment confirms our predictions for $K_{max} \geq 16$: [pt] is produced the least frequently for $\tau = 4$ ($T_{step} = 0.8$ in our experiment²²), and these frequencies are only slightly larger than the minimal that would correspond to

²²Further results not reported here for lack of space show that different T_{step} values corresponding to the same τ (such as 2 and 1.5, or 1.2 and 1) have not produced significantly different frequencies. Probably many more runs are required in order to be able to demonstrate the role of the t values in the inner loop of the algorithm, as we have done in subsections 5.5.2 and 5.5.3.

K_{max}	$T_{step} = 1.5$	$T_{step} = 1.0$	$T_{step} = 0.8$	$T_{step} = 0.6$	$T_{step} = 0.4$	Pred.
150	1046 ± 0005	0994 ± 0001	0987 ± 0003	0990 ± 0006	1052 ± 0002	0975
100	1269 ± 0003	1213 ± 0002	1201 ± 0001	1212 ± 0004	1277 ± 0001	1194
70	1503 ± 0003	1439 ± 0002	1425 ± 0001	1433 ± 0006	1518 ± 0004	1425
50	1754 ± 0006	1689 ± 0005	1679 ± 0001	1691 ± 0003	1778 ± 0005	1683
40	1939 ± 0006	1870 ± 0007	1857 ± 0002	1877 ± 0005	1986 ± 0002	1879
35	2058 ± 0004	1984 ± 0008	1981 ± 0007	1997 ± 0010	2110 ± 0005	2006
30	2196 ± 0002	2124 ± 0010	2119 ± 0001	2141 ± 0011	2261 ± 0005	2163
25	2380 ± 0002	2307 ± 0009	2304 ± 0004	2329 ± 0003	2457 ± 0004	2364
22	2506 ± 0003	2444 ± 0004	2436 ± 0004	2466 ± 0004	2604 ± 0008	2515
20	2609 ± 0002	2544 ± 0003	2542 ± 0005	2576 ± 0002	2709 ± 0009	2633
18	2713 ± 0003	2660 ± 0003	2657 ± 0007	2699 ± 0004	2839 ± 0008	2769
16	2849 ± 0012	2786 ± 0003	2794 ± 0004	2824 ± 0008	2984 ± 0003	2929
14	2994 ± 0007	2947 ± 0009	2962 ± 0004	2994 ± 0004	3156 ± 0009	3118
12	3170 ± 0007	3127 ± 0009	3153 ± 0004	3186 ± 0002	3363 ± 0005	3348
10	3378 ± 0004	3353 ± 0003	3374 ± 0005	3423 ± 0004	3602 ± 0002	3633
8	3638 ± 0004	3629 ± 0004	3656 ± 0005	3718 ± 0006	3910 ± 0010	3996
7	3786 ± 0003	3792 ± 0006	3845 ± 0004	3888 ± 0003	4091 ± 0003	4213
6	3967 ± 0001	3979 ± 0002	4039 ± 0009	4097 ± 0008	4288 ± 0005	4456
5	4168 ± 0001	4202 ± 0003	4266 ± 0002	4327 ± 0003	4500 ± 0007	4711
4	4433 ± 0005	4480 ± 0008	4539 ± 0006	4590 ± 0007	4729 ± 0007	4928

Table 6.4: **The turning point around $\tau = \kappa + 2$:** the frequency of [pt] for different parameters, with the initial “0.” truncated due to lack of space. The turning point (a local minimum in the frequency of [pt]) predicted to be at $T_{step} = 0.8$ ($\tau = 4$) can be observed for larger k , even though $T_{step} = 1$ ($\tau = 3$) often produces frequencies that are not significantly different. Nonetheless, further factors become important for lower k values, and the turning point slowly shifts towards larger T_{step} : to $T_{step} = 1$ for $16 \geq K_{max} \geq 7$, and to $T_{step} \geq 1.5$ for $7 \geq K_{max} \geq 4$. Finally observe that wherever these further factors are not yet observable, our expectations on the lower bound predicted by equations (6.21) and (6.23) are met: the values in the fourth column ($T_{step} = 0.8$, $\tau = 4$) are but slightly larger than those in the last one (corresponding to $\tau = 4.5$).

$\tau = 4.5$ according to equation (6.21). However, our approximations are not good enough any more if $K_{max} \leq 16$. Most probably further factors have to be taken into considerations, or our approximations must be refined, in order to explain why the turning point shifts towards larger T_{step} , and why the observed frequencies are much lower than the predicted lower bound.

6.6 What have we learnt from [voice] assimilation?

The starting problem of the present chapter was Dutch voice assimilation in linguistic forms such as *op die* and *zakdoek*. The first model presented included four candidates with a topology of the form that we called a “magic square”. Similarly to some tableaux in the three-candidate search space of subsection 2.3.2, this magic square, together with the hierarchy we employed, demonstrates that SA-OT does not necessarily converge towards maximal precision as the number of iterations increases. The proposed version of simulated annealing for OT cannot avoid getting caught in a local optimum with a constant probability—independently of the cooling schedule.

As argued above, however, this phenomenon may help accounting for certain irregularities. Instead of making the model more complicated in order to include them, the model of the static mental representation (competence) can be kept simple, and irregularities are quarantined in the dynamic computational process. For instance, the model of Dutch phonology will include only regressive voice assimilation, that is, the only global optimum is *o[bd]ie*. Nevertheless, the local optimum *o[pt]ie* is also returned by the dynamic computational process as an irregular form. As the *o[bd]ie vs. o[pt]ie* alternation is most probably not a fast speech phenomenon, there is no need to tune the frequencies through the cooling schedule, as we have done in Chapter 5: the frequencies have to be the same under different speech conditions (for different speech rates).²³

From a methodological point of view, the difference between this variation and fast speech was that the observation that the frequency of the *andante* form diminishes at higher speech rate makes possible to point immediately to the *allegro* form as performance error. Whereas in the present case, only the theoretical model will decide which is the form that can be easily described (for instance, by having it globally optimal in OT), and which form has to be exiled to the dynamic computational process. Such an approach may prove to be advantageous even for language acquisition: a learning algorithm robust enough to deal with the inevitable noise will learn the simpler grammar faster, in which case the “performance effects” are realised for free.

But the situation is not so simple. The 50-50% distribution might turn to be incorrect empirically for *op die*; and is certainly false for other words such as *zakdoek* where only regressive assimilation may occur. We have, therefore, introduced a second model that involved an infinite search space.

However, the parameters influenced this model in a surprising way. Unlike earlier, decreasing T_{step} decreased the chance of returning the grammatical form *o[bd]ie*, while K_{max} also played an important role. And yet, this divergent be-

²³According to Paul Boersma, the voiceless variant might be more common in fast speech, which situation could be modelled using T_{step} values larger than the turning point.

haviour can be nicely interpreted. Fast speech phenomena were analysed using the parameter T_{step} in the previous chapter, whereas the present phenomenon is different, and is consequently analysed employing another parameter, namely K_{max} . If the $o[bd]ie \sim o[pt]ie$ variation is dialect dependent, speaker dependent or register dependent, then this variation should be driven by a parameter different from the one driving speech-rate dependent variations.

Furthermore, this observation also helps in explaining the difference between *op die* as opposed to *zakdoek*. An important point about Simulated Annealing Optimality Theory is that it is not only able to account for the presence and the absence of variation by tuning its parameters, but that the parameters can also be interpreted. Namely, parameter settings yielding variation often correspond to a faster production process than parameter settings yielding almost exclusively a given form. Now, production speed may not depend only on speech rate, but also on word frequency: frequent words, such as the unstressed function words in *op die*, are probably more quickly processed than relatively less frequent nouns. That is, processing *zakdoek* involves a much greater K_{max} , which causes the computation to take longer *even for the same speech rate*, and consequently the frequency of the regressive assimilation form to converge to 1. (Note, however, that our argument will be different for the Hungarian definite article, which is a phenomenon related to speech rate: there, we employ the fact that a larger K_{max} requires a longer computational time, and thus may be viewed as also corresponding to a slower speech rate.)

The second model also shows the usefulness of an infinite candidate set. Candidates that can never win are not only necessary for the mathematical consistency and beauty of the model, but they may also influence the search algorithm. In traditional OT, loser candidates (candidates winning for no constraint ranking) could be already excluded from GEN, but in SA-OT they play a role behind the scenes. Even if they are never returned as outputs, the system may rove through them, and it is exactly because the search space is infinite that the output frequencies can be tuned by varying K_{max} .

The analysis of this second model for voice assimilation with an infinite search space reveals an additional peculiarity. Let us alter slightly the definition of the constraints so that [pt] is the global optimum:

/pd/	DEP	ASSIM[VOICE]	C ₃	FAITH[VOICE]
~ [bd]			*!	*
☞ [pt]				*
[pd]		*!	*	
[bt]		*!	*	**
[b@d]	*!			*
[p@t]	*!		*	*
[p@d]	*!		*	
[b@t]	*!		*	**
...
[b@ ⁿ d]	*n!			*
[p@ ⁿ t]	*n!		*	*
[p@ ⁿ d]	*n!		*	
[b@ ⁿ t]	*n!		*	**
...

(6.25)

This model is expected to display the same behaviour as the original one. At small K_{max} values, both local optima, [pt] and [bd], have 50% chance to be returned; but channelling through the [b@ⁿd] river becomes significant as K_{max} increases, making [bd] more probable. Consequently, we have a model in which the global optimum, now [pt], can never be returned in more than half of the cases, and its frequency can even converge to zero.