Finding the right words

Bíró, Tamás Sándor

Publication date:
2006

Citation for published version (APA):
Bíró, T. S. (2006). *Finding the right words: implementing optimality theory with simulated annealing*. s.n.

Chapter 4

The Linguistic Context of SA-OT

4.1 A few words about the lexicon

The goal of the present section is three-fold. First, it aims at saying something about the way the lexicon can be seen from the point of view of Simulated Annealing Optimality Theory. Linguistics has failed to get round the questions related to the lexicon such as lexical exceptions, and interest has recently increased in lexicalist approaches. Within OT, the language specificity of the lexicon seems at first view to conflict with the *Richness of the Base* principle (all inputs are possible in all languages, cf. Prince and Smolensky, 2004, p. 225). According to another principle, *Lexical Optimisation* (*ibid*), the language learner should choose the input that corresponds to the most harmonic output among the possibilities given the surface form observed (cf. also the *Robust Interpretive Parsing* of Tesar and Smolensky, 2000).[1]

In particular, and this is the second goal of the present section, we point to the way that the complex lexicon model of Burzio (2002) could be realised in practice using SA-OT. Concrete realisation of this model drawn on physics is left to future work, nevertheless.

The main component of this model, *Output-Output Correspondence* (OOC, *Output-Output Faithfulness*) proposed by Benua and Burzio has been a widely used constraint within Optimality Theoretic phonology, and yet, it lacks a precise workable definition to my best knowledge. To be more precise, Burzio (2002) seems to have not fully worked out the details of his proposal, so that even though most linguists use OOC (successfully) in an even less formal way, this practice does not work for SA-OT which requires the exact number of violation marks assigned to any candidate. Consequently, and this is the section's third goal, a more formal definition of an Output-Output Correspondence-like constraint will be introduced, in order to employ it in Chapter 5.

Subsequently, the second section of the present chapter includes a few notes on learnability in SA-OT. Learnability issues have been successfully tackled both in standard OT (Tesar and Smolensky, 2000), as well as in Stochastic Optimality

[1] For the role of the lexicon in OT syntax, see for example van der Beek and Bouma (2004) and references therein.

Theory (Boersma and Hayes, 2001), a fact that provides a strong argument in favour of Optimality Theory, as opposed to many previous linguistic models. Therefore, the reader would naturally ask whether SA-OT has anything to say about learnability.

4.1.1 English Past Tense

One of the most investigated issues related to rules, *minor rules* and lexical exceptions is the case of English past tense. As is well known, the productive *major rule* is to add the suffix <*ed*>,[2] while *minor rules* may prescribe changing <*ing*> to <*ang*> (like in *sing – sang* and *ring – rang*) or the last coda to <*ought*> (e.g. in *bring*, *think*, *seek*). Some cases, such that of the verb *to be* or *to have* are fully irregular.

Several approaches have been presented to tackle the problem, starting with connectionist approaches (Rumelhart and McClelland, 1986), to output-output correspondence (Burzio, 2002) or ACT-R models (Taatgen and Dijkstra, 2003). Within OT, the first approaches have been proposed by Boersma (1998b) and by Burzio (1999). The latter was finally published as Burzio (2002), and we shall turn back to it soon.

In fact, a major debate emerged from this phenomenon, the so-called *Past Tense Debate*, when Pinker and Prince (1988) reacted to Rumelhart and McClelland (1986): the connectionist camp (McClelland, Plunkett, Seidenberg,...) argued for a single mechanism for both regular and irregular verbs, whereas the proponents of symbolic computation (Pinker, Ullman,...) fought for a dual route mechanism. For a recent, two-sided overview of the debate, see both Pinker and Ullman (2002) and McClelland and Patterson (2002). For recent neurolinguistic arguments for the dual route model based on double dissociation, see the work of William Marslen-Wilson and his colleagues (e.g. Tyler et al. (2002) and Stamatakis et al. (2004)). This debate goes much beyond the issue of English past tense, the latter being only a test case: the question of debate is the role of symbolic rules as opposed to connectionist approaches in language processing, or even in cognition in general.

Although its details may have been debated, a pattern called *U-shaped development* can be (more or less) observed in children's acquisition of English past tense forms (Brown (1973), pp. 333; Kuczaj (1977); Harley (2001), p. 96 and 125-126, and references therein, including an introduction to the Past Tense Debate). Even if using only a very restricted vocabulary, the youngest children perform quite well in producing the past tense of verbs. In a second stage, however, performance drops, before improving again in the third stage. Roughly speaking, we may say that the child memorises all forms in the first stage; later, the growing vocabulary allows making generalisations, and the drop in performance is due to over-generalisation, so forms such as **bringed* or *singed* appear beside the correct ones; in the third stage, nevertheless, these cases of over-generation are learned to be errors, that is, exceptions are (re)-learned.

A further interesting phenomenon is the acquisition of the so-called *minor rules*. For instance, such a minor rule, inferred from *sing – sang* and *ring – rang* may require changing the coda of a monosyllabic verb from <*ing*> into <*ang*>.

[2] For the sake of ease, I use the written form of the segment strings, and not the underlying representation or some surface allomorphs.

Child speech indeed produces, although with a very low frequency, forms such as **brang*, which can be seen as the result of overgeneration from this minor rule (Xu and Pinker, 1995; Taatgen and Dijkstra, 2003).

One may speculate about minor rule forms corresponding to local optima in the SA-OT search space; but only future work can tell whether such an approach to account for these phenomena—including those in acquisition—will turn to be fruitful.

4.1.2 Burzio's physical model of the mental lexicon

Burzio (2002) attempts at giving an Optimality Theoretical compromise to the English Past-Tense Debate, by using a model based on an idea taken from (classical) physics; namely, on the concept of *forces* and *fields*. In what follows, I am making these physical analogies more explicit than as found in Burzio (2002) itself.

His aim is to explain why "[l]exical sectors that are morphologically irregular tend to be phonologically regular, and vice-versa". He proposes that whenever morphology is irregular and phonology is regular ("level 1" affixes in Kiparsky's Lexical Phonology), then the phonological markedness constraints dominate. For instance, the vowel of the verb *keep* is shortened in the past tense form *kept* in order to meet the limitations on syllable size. But in other cases, morphology is regular and phonology turns to be irregular ("level 2" affixes): for instance, the regular past tense form *beeped* includes a syllable that is so long that it would be otherwise prohibited. Then, the analogy in the paradigm acts as an attraction between the forms, overranking phonological well-formedness requirements. This attraction is described by *Output-Output Correspondence* (or *Faithfulness*), and is seen as some sort of gravitational force between the lexical items.

In physics, bodies with a mass create gravitational fields around themselves, bodies (or particles) with an electric charge create additionally an electric field, and so on. The fields thus created by the individual bodies are summed up to form the field in which (the same or different) bodies follow their trajectories. The movements of the bodies are driven by the forces derived (literally) from the overall field, whereas this field in each moment is a function of the location (and speed, for magnetism) of the bodies. Two additional forces can be present: friction hinders any changes of position, whereas external forces (e.g. a gravitational field) favour some positions over others.

A field can be seen either as a scalar-valued function (*energy*, or rather *potential*) or a vector-valued function (*force*) of space (and time). If you put a given body at a given point in space and time, the properties of that body (*e.g.* its mass in the case of gravitation, its electric charge for electric interaction, its charge and speed for magnetic interaction, etc.) and the field (as a function of all the bodies or particles around) will determine what the energy of that body is, and what force the field exerts on that body. Moreover, the force is the *negative gradient* of the energy: a vector pointing into the direction in which energy declines the most, and the length of the vector is proportional to the steepness of the energy function in that direction. Indeed, the idea is that the physical force influences the body to move towards the minimal energy state. In other words: it is sufficient to define the energy (potential) as a scalar function in space, for its negative gradient (a spatial derivative) in each point gives the force acting upon some particle there.

We can now summarise the picture thus far: the position and speed of a particle in the next moment is determined—besides its mass, position and speed in the previous moment—1. by friction, 2. by the external forces, as well as 3. by the aggregate force exercised by the other bodies. The latter can be calculated from the position of the bodies and their mass or charge. If x_i and m_i are the position and mass of particle i, while $F(j \rightarrow i)$ and $V(j \rightarrow i)$ are the force acting on particle i and the energy of particle i in the field created by particle j (the influence of j on i), then the differential equation describing the trajectory of particle i is by Newton's second law:

$$\begin{aligned} m_i \cdot \frac{d^2}{dt^2} x_i &= F_{friction} + F_{external} + \sum_{j \neq i} F(j \rightarrow i) = \\ &= F_{friction} + F_{external} - \frac{d}{dx} \sum_{j \neq i} V(j \rightarrow i) \end{aligned} \tag{4.1}$$

In Burzio's model, lexical items are the bodies in a multidimensional space, whose dimensions correspond to phonological, syntactic and semantic features. The distance of two words can be measured in the number of features they differ from each other. As Burzio writes (p. 11): "*[t]his model performs a simple calculation in which the input is the position at which the object is originally placed, and the output is the ultimate resting position*". Thus, friction will correspond to *Input-Output Correspondence*, the force that acts against changing position. External forces correspond to the markedness constraints: independently of the position of the different bodies, they pull each of the bodies towards some preferred positions. Finally, the force exercised by the other bodies translates to *into Output-Output Correspondence* (OOC)—we shall return to this point in the next subsection.

So for instance, most constraints used in linguistics can be seen as external factors, such as the Earth's gravitational field in which everyday objects with mass follow a certain trajectory. Similarly to gravitation, which favours some positions over other ones, markedness constraints favour certain feature combinations, that is, specific positions in the space. Remember that the linguistic features (the phonological content, the syntactic class, semantic properties) of lexical items are encoded as the dimensions of the space. If, for example, some constraint disfavours front rounded vowels, harmony improves by moving towards [-round] in the [round] dimension—just like gravitation, which prefers the butter side of slices of bread and butter to be lower in the vertical dimension.

In OT terms, the points of the multidimensional space are the candidates, while the output, the "ultimate resting position" is the winning candidate where the forces neutralise each other. In physics, such a stable resting point is a local minimum of the energy: there the spatial derivative (the gradient) of the energy is zero, so no force acts upon the body, and moving away from that point would increase the energy. Consequently, the OT Harmony function will correspond to the (negative) energy in the physical analogy, and the goal is to find the position (the candidate) that minimises energy (maximises harmony).

Here, energy includes not only the energies from the interactions with each of the other particles, but the external forces and friction are also integrated (literally) into energy. Actually, friction should rather be replaced by springs, also mentioned by Burzio. The more the spring is pulled, the larger its energy, which corresponds to a larger force pushing the particle back to the origin. In turn, candidates or lexical items are strings stretching between the input form and the output form. A candidate's energy (or harmony) is the sum of the spring's energy (Input-Output Faithfulness or Input-Output Correspondence), of the energy from the external field (markedness constraints) and of the energy from the interaction with the other lexical items (Output-Output Correspondence).

It is unclear how precisely this sum has to be calculated in Burzio's model. As he later employs an OT-model referring to strict constraint ranking, I suggest the polynomials or ordinal numbers as representations, following Chapter 3. Thereby, it will be possible both to interpret the physical analogy (involving sums and derivatives), and to save the connection to Optimality Theory.

Burzio does not elaborate either on what the "ultimate resting position" is, he simply supposes that it is the global minimum of the energy (harmony), following the principles of standard Optimality Theory. Indeed, in quantum physics, a non-global local minimum is only a metastable position, as sooner or later (this time range is called the *half-life*) the particle jumps to some lower minimum. But if the half-life is very long, as well as in classical mechanics, local minima can also be quite stable. Therefore, if the "ultimate resting position" is only required to be some local minimum (following the physical analogy), we obtain a similar picture to that used in SA-OT: possible surface forms are local optima, among which the global optimum is (usually) the most frequent one. Indeed, "local optimum" is the central concept, and the global optimum is but a special local optimum.

Additionally, the parallel between Burzio's model and the topology in SA-OT becomes even stronger if we make explicit that in Burzio's model neighbours—a concept required in the definition of local optima—are points whose distance is 1, that is, candidates that differ exactly in one feature, in one basic transformation. Alternatively, a quantum physics-like model, in which non-global local optima may be metastable if the half-life is very long, corresponds to another type of SA-OT topology: to the definition in which any two candidates are neighbours, but the *a priori* probability diminishes with distance. In this case, a candidate can be attested because it is a "metastable local optimum" in the sense that jumping to a better one is extremely improbable, because better candidates are very far away (so SA-OT will be stuck there); similarly to radioactive isotopes found in nature whose half-life is comparable or longer than the age of the universe, so that they have not decayed yet.

In brief, Burzio's search space is a special case for the search space employed in SA-OT. A special case, but a very self evident and general one. He does not specify the way he would perform the search for the "ultimate resting position". (Would he calculate step by step the trajectory of each item from the input form to the output form? Does anything guarantee that such a trajectory ends in a resting position?). And yet, the physical systems that motivated simulated annealing (including the $e^{-\Delta E/T}$ factor) are the same as those inspiring Burzio. Hence, the close connection between the two proposals, I believe, is worth further research.

4.1.3 Burzio's Output-Output Correspondence

Let us turn our attention now to the way Burzio (2002) introduces the most interesting type of force present in his model, Output-Output Correspondence, that is, the interaction between particles. This "gravitational force" between pairs of words is argued to be responsible for phenomena such as analogical effects.

To sum up what we have discussed so far, the lexicon of a language is composed of lexical items that optimise locally their "energy" (*i.e.*, their harmony function). The energy of an output form depends on its well-formedness (phonological markedness constraints), on its distance from the input form (Input-Output Correspondence), as well as on its interaction with the other output forms (OOC). Hence the Saussurian concept of the language as a complex system: altering one surface form influences all other outputs through their interaction.

Burzio introduces the notion of *representational entailments* (on p. 176), which, he argues, is cognitively plausible. The position vector of some word $A = (a_1, ..., a_n)$ can be seen as a set of entailments of the form "if position i has a_i, then position j has a_j" for all possible i's and j's. Take now a second lexical item, B, whose coordinates equal those of A in k out of the n dimensions (features), and differ in $n-k$ dimensions. Given this, B violates $k(n-k)$ entailments of A: there are k different positions i and $n-k$ different positions j, such that B has a_i in position i, and yet, not a_j in position j. Hence, Burzio's proposal—the way I interpret the August 1999 version of his paper, which is slightly more explicit (Burzio, 1999)—defines the "gravitational" potential $V(A \rightarrow B)$ exercised by word A upon word B as the number of entailments of A violated by B. This potential as a function of the non-Euclidian distance k is:

$$V(k) = k(n-k) = nk - k^2 \tag{4.2}$$

The "gravitational" force with which A attracts B is the spatial derivative of this potential:[3]

$$F(k) = \frac{\partial V(k)}{\partial k} = n - 2k \tag{4.3}$$

The direction of this force points towards word A.

It becomes clear that the closer the two words (that is, the smaller the k), the stronger they attract each other. In that property, Burzio's inter-word force resembles vaguely gravitation and electrostatic force. If the two words are very far, attraction vanishes; even further ($k > n/2$), the force turns into repulsion ("anti-gravitation").

Subsequently, a trick often used in physics is employed by Burzio. A body composed of many particles can be replaced by its mass centre ("centre of gravity") for the purpose of calculating its gravitational attraction. This is so, because the gravitational force exerted by each particle can be decomposed into two components: when summing up the forces exerted by all the particles, the first components cancel each other, whereas the second components sum up as if all the mass were concentrated in the mass centre. This trick helps Burzio to understand the effect of a group of words on a particular word.

[3] To be more precise, the negative derivative is the repulsive force. To increase clarity, we concentrate on attraction, however. Cf. the right hand side of equation (4.1).

Suppose that a group of words share some representational entailments: in Burzio's example, *parental*, *natural*, etc. all share the entailment according to which "the ending *al* must be preceded by a noun". Other entailments that are not shared include "the ending *al* must be preceded by the string *parent*" or "the ending *al* must be preceded by the string *atur*". When summing up the entailments of these words, (hence, the force, since the derivative in Eq. (4.3) is additive), the effect of the latter entailments will neutralise each other. Yet, the group of words will yield the *macro-entailment* "the ending *al* must be preceded by a noun". This is the way Burzio hopes to explain paradigmatic effects.

Consider an arbitrary word. The gravitational force exerted on it by most of the lexicon is negligible, partially because most of the words are far enough away (having many different features), and partially because their effects neutralise each other—unless the word is located "outside" of the majority of the lexicon, such as in the case of a foreign word whose phonological features have not been assimilated into the general phonological system of the language. In the latter case, the lexicon as a whole exerts some attracting force. In the most frequent case, however, only particular words in the neighbourhoods will exert attraction: assimilation to a set of similar words, paradigmatic levelling, etc. Furthermore, the closest existing lexical items to a derived word are its root and the outputs of the previous cycles of the derivation. Through this idea, Burzio's Output-Output Correspondence is able to account for phenomena previously accounted for by cyclical derivation.

Burzio is even able to explain why the root has more influence on the derived form than vice versa. He argues that more representational entailments of the shorter root are satisfied by the derived form than vice-versa. For example, *parent* violates *parental*'s entailment "if the word's first segment is a p than its eighth segment is an l", while all entailments referring to the segments of *parent* are satisfied by *parental*. In turn, *parental* is more influenced by *parent* than vice versa. The only problem with this argument is that we become uncertain about the exact representation of a lexical item as an n dimensional vector, with always n_{segm} dimensions corresponding to phonological segments.

Finally, turning back to the Past-Tense Debate, what is Burzio's explanation of the different behaviour of Level 1 (highly irregular morphology, highly regular phonology) and Level 2 (highly regular morphology, yet often irregular phonology) word derivations? The difference is the place where Output-Output Correspondence is ranked, relative to phonological markedness constraints and to Input-Output Correspondence. Moreover and most importantly, the different ranking results from the significantly different numbers of stems belonging to a certain derivational paradigm (Burzio (2002), p. 195). Level 1 affixes take relatively few stems, and therefore gravitation's morphological levelling effect is weak: Output-Output Correspondence is ranked below phonological markedness, yielding irregular morphology and regular phonology. On the other hand, the possibly infinite number of stems to which Level 2 affixes can be applied boosts the effect of Output-Output interaction over the phonological markedness constraint—resulting in a regular morphology, and an irregular phonology.

By (literally) deriving grammatical effects (output-output constraints) from the words in the lexicon, Burzio reverses—as he himself remarks—the one-way relation from the (adult) grammar to the output dominant in the generative tradition.

4.1.4 Burzio's model and SA-OT

A very important question is still open, however. How precisely are the different forces summed up? Equation (4.3) gives the "force" with which lexical item A attracts word B. Without asking crucial details about the exact number and nature of the dimensions, and supposing that the different forces are simply summed up, it is still unclear how this effect is translated into a position of Output-Output Correspondence within the hierarchy. And this last issue seems to be the major point in assessing Burzio's proposal.

Although we are not going to come up with concrete proposals, the different ways of representing the Harmony function introduced earlier allow us to speculate about possible directions of future work.

Notice that Burzio's proposal gets tangled up where it has to accommodate a traditional Optimality Theoretical framework. Supposing that representational issues—the exact form of the feature vectors—are solved, and accepting the neural plausibility of representational entailments, the potential introduced in Eq. (4.2), as well as the derived force in (4.3) are well-founded and elegant. And yet, are we sure and certain how many stars to assign to a particular candidate in any case?

This question might be avoidable in SA-OT, however. Do we really need to translate Burzio's formalism into terms of standard Optimality Theory? Observe that what we did in earlier chapters was the opposite translation: transforming constraint violations into some energy (potential, harmony) function to be optimised. As a simple real-valued function would not work for Optimality Theory in the general case, we have introduced the polynomial representation and the ordinal number representation of the Harmony function—both having the form of a sum.

Consequently, Burzio's "gravitational" potential, once well formulated, can be added directly to some formulation of the Harmony function. This new addend does not necessarily have to have the exact form of the addends obtained from the traditional constraints: we may give up on seeing the gravitational effect as a constraint. Yet, the gravitational potential should be formulated within a similar formalism, so that it can be added to the representation of the constraints. Not bad news in itself, as probably Burzio's "gravitational" potential is not really suited for a real-valued representation, for the simple reason that it requires the sum over an indeterminably large lexicon. Furthermore, although the gravitational effect in the harmony function will not have the form of a constraint, yet its magnitude within the summands probably can be estimated—and be interpreted as Output-Output Correspondence being ranked higher or lower than markedness constraints.

As a speculation, remember how Burzio explained the different ranking of Output-Output Correspondence for Level 1 and Level 2 derivations: in the first case the effect of at most a few hundred words are summed up, whereas the summation in the second case takes place on an open class of words. If the mean potential obtained from a single word in the class is $\overline{v}$, then one hundred words provide a potential of $\overline{v} \cdot 100$. Yet, in the case of fully productive morphological processes in Level 2 derivations, the open set has the cardinality of a countably infinite set ($\aleph_0$): in turn, is the summed up potential $\overline{v} \cdot \omega$? If so, the corresponding addend is of a higher magnitude in ω and we have understood why Burzio argued for promoting Output-Output Correspondence higher in the case of Level 2 morphology.

Let us now step back from speculations. Burzio's model is undoubtably attractive—at least to a person with a background in physics. However, the model is very hard to implement. In practice, the phonologists using Output-Output Correspondence define *a priori* which other output the given form must be compared to, and do not demonstrate that the interactions with *all* the other words in the lexicon are negligible, and indeed extinguish each other.

Consequently, we recommend replacing Output-Output Correspondence with correspondence constraints that refer explicitly to the process of morphological derivation. One such constraint could be Kenstowicz's BASE-IDENTITY (Kenstowicz, 1995). However, in the following section, we demonstrate that very often it is not the *base*, but the *output of the previous cycle of the derivation* that is relevant, a fact well known in the *Lexical Phonology* of Kiparsky (1982). Therefore, we recommend introducing a constraint named COMPONENT-OUTPUT CORRESPONDENCE / CONSTITUENT-OUTPUT CORRESPONDENCE (COC). If I keep the original name OOC in the next chapter while I mean in fact COC, it is because I would like to retain readability for phonologists.

4.1.5 Constituent-Output Correspondence

In this subsection, we define COC, so that it can serve us in Chapter 5 on metrical stress in Dutch fast speech.

By way of introduction, I must express my reservations with regard to the general way of defining a constraint in the OT literature. It is true that originally constraints were requirements that a linguistic form either met or did not, and therefore, introducing a constraint meant defining the condition that a form had to meet (for instance: "each syllable has an onset", "no syllable has a coda"). Nevertheless, with the advent of violable constraints, and, especially since more levels of violation could be distinguished, a constraint is rather seen as a function mapping each linguistic form to a numerical value (usually, the number of violation marks). Consequently, the definition of a constraint must tell how many violation marks are assigned to a given candidate (for instance: "the number of codas in the word", or "one star per syllable with a coda").

We particularly have to emphasise this here, because this task is especially difficult in the case of OOC and COC. The authors of most articles are lucky enough to be able to point intuitively to the fact that the optimal candidate is "clearly" better with respect to OOC than its competitors, so they can eschew giving an exact definition of OOC. Yet, Simulated Annealing Optimality Theory has to be able to compare the violation levels of any two neighbouring candidates. In turn, the number of violation marks incurred by a candidate should be defined exactly; or, at least, the difference in the violation levels ought to be given for any pairs of neighbouring candidates. The second way, undoubtedly challenging, assigns a violation difference to each of the possible basic steps.[4]

[4] For instance, in the case of metrical stress assignment to be presented in Chapter 5, moving the unobservable foot borders should not introduce any changes with respect to OOC (COC). Nonetheless, deleting and inserting a stress (a foot), as well as moving the position of a stress (changing the head syllable of a foot) may involve some changes in violating OOC (COC). One parameter will define the possible change due to deletion or insertion, and another parameter will describe the role of changing the place of a stress. These two parameters, nevertheless, correspond to the parameters used in the approach described presently.

In the following, nonetheless, we follow the first way, for its being simpler and consistent with the general claim of defining each constraint as a function.

Correspondence Theory was introduced in the early years of Optimality Theory by McCarthy and Prince (1993b) (p. 67) for the sake of reduplicative phenomena. (We shall use it also in section 6.3.) In later developments, the *correspondence relation* $\mathcal{C}_w$ maps the segments of the underlying form to the "corresponding" segments of the candidate string w. Then, constraints may require that each underlying unit have a corresponding image in the candidate (constraint Max—originally called Parse with a different philosophy—prohibiting the underparsing, *i.e.* the deletion of parts of the input); each surface element be the correspondent of an underlying segment (constraint Dep—Fill in Prince and Smolensky (1993)—punishing epenthesis); and that input and output segments be the same (further types of faithfulness constraints).

Unlike in the general case, pairing the basic units of the input and the output string is easy in stress assignment, for GEN adds some structure (namely, the metrical structure) on the top of the input string without altering the latter. Hence, the input string and the output string are composed of the same number of syllables, and the nth syllable of the input string *corresponds* to the nth syllable of the output string.

Thus, we focus on the correspondence of the metric structure (stress pattern). Yet, we employ *Output-Output Correspondence*, or *Component-Output Correspondence*, and not *Input-Output Correspondence.* When assessing a candidate w with respect to *Output-Output Correspondence* (*Component-Output Correspondence*), we will compare it to a string σ of the same length. In the case of *Output-Output Correspondence*, σ has to be derived from the stress pattern of any word in the lexicon, which does not necessarily has the same length as w. Yet, as previously argued, I propose to replace *Output-Output Correspondence* with *Component-Output Correspondence*, and in this case σ is the stress pattern derived from the stress patterns of the morphological constituents of w.

In the simplest case, if w (actually, $GEN^{-1}(w)$, the corresponding underlying representation) is the concatenation of a number of morphemes, then σ is the concatenation of their stress patterns. To be more precise, the candidate (the output form-to-be) is compared to the way its components are realised as independent words (output forms) in the language—hence the name of the constraint. Affixes are not independent words of the language with some stress pattern, yet they may act as if they were: in Burzio's approach, all the words with a given affix and a given stress pattern on that affix would jointly have such an analogy effect.

Burzio's paradigmatic example is *condensation* as opposed to *compensation.* The word *còmpensátion* is derived from *cómpensàte*, and the vowel of the unstressed second syllable may be reduced to a schwa. Yet, *còndensátion* is derived from *condénse*, and the stressed second syllable in the root adds a tertiary stress to the second syllable of *condensation*, prohibiting its reduction to schwa.

A similar example has been proposed by Dicky Gilbers and Maartje Schreuder (personal communication). The six-syllable-long Dutch words *sèntimentàlitéit* ('sentimentality') and *ìndivìdualíst* ('individualistic person') have seemingly very similar syllable structure: only their third syllables differ in weight, but if the weight-to-stress principle were active, it would predict the opposite pattern. Nevertheless, their morphological derivation is different:

Cycle 1	sèn.ti.mént	ìn.di.vi.dú	
Cycle 2	sèn.ti.men.téel	ìn.di.vì.du.éel	(4.4)
Cycle 3	sèn.ti.men.tà.li.téit	ìn.di.vì.du.a.líst	

Observe that it is cycle 2 which determines the stress pattern of cycle 3. If the root were the decisive factor, *sentimentaliteit* should have a stress on its third syllable, and *individualist* on its fourth one, but the change of the stress pattern in cycle 2 causes the opposite constellation. Interestingly, the native speaker of Dutch observes that the misplacing of the stress changes the semantic field of the (non-existing) word form: *sèntimèntalitéit* is conceived of as some kind of *mèntalitéit* ('mentality'), whereas *ìndividùalíst* sounds as some sort of *dualist.*

Consequently, the stress pattern to which the different parsings of the input form *individualist* are to be compared to is sususs (s meaning stressed syllable, u referring to unstressed syllable): the stress pattern susus of *individueel* followed by the pattern s of the suffix (for the *ist* ending attracts stress). Similarly, *sentimentaliteit* is compared to the concatenation of the stress patterns suus from *sentimenteel* and of us from *-iteit.*

After these preparations, we are ready to define the constraint Component-Output Correspondence. The number of violation marks assigned to a candidate w is the number of mismatches with the corresponding string σ, after a pairwise comparison of the corresponding elements of the (equally long) strings:

$$\mathrm{COC}_\sigma(w) = \sum_i \Delta(w_i, \sigma_i) \tag{4.5}$$

where w_i and σ_i represent the ith element (in the present case, whether the ith syllable is stressed or not) of the candidate w and of the string σ used for the comparison; and where:

$$\Delta(w_i, \sigma_i) = \begin{cases} 1 & \text{if } w_i \neq \sigma_i \\ 0 & \text{if } w_i = \sigma_i \end{cases} \tag{4.6}$$

The definition of COC (or, OOC) is thus complete, but not satisfactory. The result is maybe not exactly what we wish. Intuitively speaking, misplacing one stress seems to be a smaller difference than missing a stress entirely, or having extra stresses. If the target string is σ =suus, then w_1 =susu seems to be closer than w_2 =suuu or w_3 =suss.[5] Yet, the above definition will assign two violation marks to w_1, because there is a mismatch in both the third and in the fourth syllable, whereas only one violation mark will be assigned to w_2 and to w_3. Candidate w_1 violates constraint COC_σ on the same level as the "totally misconceived" candidate w_4 =ssss. Is this situation that we wanted?

In turn, a modification of the constraint may assign additional violation marks to the difference in the number of stressed syllables. Let $\parallel \alpha \parallel$ denote the number of stresses in the string α:

$$\parallel \alpha \parallel = \sum_i \Delta(\alpha_i, \mathrm{s}) \tag{4.7}$$

[5] Again, from this point onwards, s refers to a stressed syllable with either a primary or a secondary stress, whereas u represents an unstressed syllable.

Subsequently, the new definition of COC is:

$$\mathrm{COC}_{z,\sigma}(w) = \sum_i \Delta(w_i, \sigma_i) + z \cdot \Big| \parallel w \parallel - \parallel \sigma \parallel \Big| \tag{4.8}$$

Notice that the present definition introduces a new parameter, namely z, which determines the relative weight of the two parts, pointwise mismatch *vs.* difference in the global number of stresses. As pointed out by several readers, here I have combined two standard OT constraints. The first addend corresponds to Ident(stress), and the second one to Max(stress). Instead of having these two constraints in a strictly dominating rank order, we have just created a weighted sum in a Harmony Grammar-style. By varying z and keeping it small, the two addends, that is, constraints Ident(stress) and Max(stress), can create different interesting landscapes, as the experiments to be described in the next chapter shall demonstrate.

Last, one would define *Component-Output Correspondence* as COC_σ (or, $\mathrm{COC}_{z,\sigma}$) with σ being always the concatenation of the immediate morphological components (in the present case, their stress pattern), and not the concatenation of deeper components. This would be how OT could account for the *bracket erasing convention* in Kiparsky (1982)'s Lexical Phonology.

On the other hand, one can make use of the above definition of COC_σ (or, $\mathrm{COC}_{z,\sigma}$) when defining Burzio's Output-Output Correspondence. Then, σ can be any element of the lexicon, and the definition should also define how to sum up the different COC_σs:

$$\mathrm{OOC}(w) = \sum_{\sigma \in \mathrm{Lexicon}} d(w, \sigma) \cdot \mathrm{COC}_\sigma(w) \tag{4.9}$$

with $d(w, \sigma)$ being some distance measure between the elements of the lexicon, which acts here as a weighting factor.

In Chapter 5, we shall make use of the *Component-Output Correspondence* constraint in the way we just have defined it, including also the z weight. Nonetheless, we shall call it Output-Output Correspondence, in order to make the discussion comprehensible to the reader familiar with past and current phonological literature, in which the term *Output-Output Correspondence* is used rather in the sense of *Component-Output Correspondence*, and not really following Burzio's original proposal.

4.2 Learning with SA OT?

The idea of learning a grammar has already been introduced roughly at the very end of section 1.1.3. The interest in learning is twofold: from the viewpoint of psycholinguistics, the question is whether a certain grammar model can reproduce language acquisition observations, such as those in child language, second language learning, post-traumatic language recovery, etc. The adequacy of a grammar model is clearly questionable if it cannot be acquired. On the other hand, natural language processing (NLP) may require machine learning algorithms (Mitchell, 1997) that can—at least partially—automate the construction of complex, high-coverage grammars.

In both cases, the goal is to find a grammar that reproduces the observed data (as well as possible). The problem is reversed compared to what we have been dealing with so far: *grammar implementation* is concerned with producing the linguistic forms for a given grammar, whereas *grammar learning* aims at creating a grammar for given linguistic forms.

The basic philosophy is defined by Chomsky's *Principles and Parameters* (*P&P*) approach. Both acquisition in psycholinguistics and machine learning require a framework, otherwise the search for a grammar would be ill-defined. At this point, we cannot enter discussions about how much this framework has to be restricted, in what sense it is innate, and how poorly or amply a child is supplied with input data about her native tongue. What is usually supposed by linguists is that *some* of the grammar is *universal* (these are the *principles*), and it is already given to the learner at the beginning of the learning process. These principles reflect, as a matter of fact, features that are, arguably, characteristic of *all* languages of the world, as all human children inherit the same framework. The cross-linguistic differences are accounted for by the different values assigned to the *parameters*, and the task of the learner is to find the parameter setting reproducing the observed data.[6]

As discussed in section 1.1.3, traditional Optimality Theory postulates GEN, as well as the set of constraints to be universal. Applying *Principles and Parameters* to standard Optimality Theory means, therefore, that one searches for the constraint ranking that accounts for the input data, because it is the hierarchy that is supposed to be the only source of cross-linguistic variation, corresponding to the notion of "parameters" in *P&P*.

Grammar learning algorithms within standard Optimality Theory, *Recursive Constraint Demotion* (RCD) and *Error Driven Constraint Demotion* (EDCD), have been developed by Bruce Tesar (Tesar and Smolensky, 2000). A linguistically more informed version of RCD is *Biased Constraint Demotion* (BSD) (Prince and Tesar (2004), Tesar and Prince (2003)), used by Ota (2004) and by Pater (2005b) to learn lexically indexed faithfulness constraints.[7] Constraint Demotion, however, lacks robustness: it presupposes that the data are produced by an OT grammar, the target of the learning algorithm, and that no noise infiltrates the data set. Cases requiring *Robust Interpretive Parsing* (Tesar (1999), Tesar and Smolensky (2000)), which inevitably introduce some sort of noise, may be unlearnable. Eisner (2000b) proposes a generalisation for RCD.[8]

The most popular of the learning algorithms for variations of OT is the *Gradual Learning Algorithm* (GLA) closely connected to *Stochastic Optimality Theory* (Boersma and Hayes, 2001), and widely used in recent years.[9] Addi-

[6] Actually, many algorithms rather aim at reproducing the observed data only as well as possible. The data set may include *noise*, inconsistencies, errors, etc., and therefore finding a model that fits all the observed data perfectly is not always feasible. Furthermore, one may want to avoid *overfitting* (Mitchell, 1997): the goal, then, to be more precise, is to correctly predict the behaviour of the system on unseen data.

[7] Bruce Hayes lists the following learning algorithms with their earliest references in the manual of *OTSoft: A Constraint Ranking Software* (available at `http://www.linguistics.ucla.edu/people/hayes/otsoft/`, version of January 12, 2004): *Classical Constraint Demotion* (Tesar and Smolensky, 1993), *Gradual Learning Algorithm* (Boersma, 1997), *Low Faithfulness Constraint Demotion* (Hayes, 1999) and *Biased Constraint Demotion* (Prince and Tesar, 1999). Both of the later two are similar to Classical Constraint Demotion, but they attempt to place all faithfulness constraints as low as possible.

[8] For the application of EDCD to a heterogeneous data set, see an early manuscript at `http://www.let.rug.nl/~birot/publications/t_biro_clin2002.pdf`.

[9] For example Jäger (2003a) combines GLA with bidirectional OT (Blutner, 2000) in order to create a language evolutionary model.

tionally, stratified grammars are learned in Ota (2004) and Pater (2005b).

All these results showing that Optimality Theory is a learnable framework have significantly contributed to the success of Optimality Theory. The obvious question arising now is what *Simulated Annealing Optimality Theory* has to say about grammar learning. The question is open to further research yet, and here we can only speculate about the possibilities.

The most important contribution of SA-OT to the Optimality Theoretic paradigm is probably the *topology* (neighbourhood structure) of the candidate set. In section 2.2.2, it has been suggested that the topology should be universal and reflect the "logic" of GEN and of the inner structure of the candidates. If this is so, the structure on the candidate set does not have to be learned; rather, it is given to the learner initially. In a second approach, however, one could include a few parameters determining the details of the topology. In Chapter 7, for example, the basic operations transforming a candidate into its neighbours are supposed to be universal, and yet, the probability of applying a particular operation may vary. In such a model, a fine-tuning of the parameters is required to reproduce frequencies similar to those appearing in the learning data set. Details are postponed to further research.

Concerning the hierarchy, Simulated Annealing Optimality Theory uses a traditional approach, so the learner may want to use one of Tesar's constraint demotion algorithms (EDCD or RCD). Do not forget that SA-OT deals exclusively with the way of calculating the optimal form in a standard OT model. Hence, you can also propose to build a Stochastic OT model and to learn with GLA; then, SA-OT is used to produce quickly an output at evaluation time for each hierarchy that is derived from the current ranks of the constraints by including noise. In both cases—constraint demotion and GLA—simulated annealing solves a seemingly elementary step cheaply, unimportant from the viewpoint of the learning algorithms. And yet, if generating the winner for a certain hierarchy is otherwise a costly operation, then learning algorithms calculating the optimal forms for different hierarchies many times would incur computational troubles. Consequently, a learning algorithm may be speeded up by using a heuristic technique.

SA-OT is not guaranteed to return the optimal candidate, however, and this fact introduces some noise into the learning algorithm. Does this observation disfavour less robust algorithms, such as EDCD? In fact, it most probably does not. Both EDCD and GLA generate the optimal candidate with respect to the current hierarchy in order to compare it to the piece of learning data. If the piece of learning data turns to be suboptimal, then the present hierarchy is altered in order to get closer to the target hierarchy, which would produce the observable data. Otherwise, no change is made. What happens, now, if SA-OT fails to find the optimal candidate for the current hierarchy? If the returned candidate is still better than the learning data, the detected error helps drive the learning algorithm (EDCD or GLA)—hopefully, towards the target hierarchy. Else, if the candidate returned happens to be worse than the piece of learning data, the learning algorithm mistakenly derives that the present hierarchy can account for the learning datum: in fact, the algorithm has just missed an opportunity to learn, and goes further to the next piece of data (unless this misconclusion causes the algorithm to stop). In sum, the (relatively low) noise introduced by SA-OT most probably has no other effect than to increase the number of learning steps required by the learning algorithm. Further experimentation may compare the gain in speed due to the use of a heuristic technique to the increase in the number of steps caused by this noise.

A real SA-OT learning task would be the following: the learning data are produced using an SA-OT model (with known or unknown parameter setting), and a hierarchy is sought that reproduces the same distribution of outputs. Suppose that the topology and the set of constraints are given, and the goal is to find the association of the constraints with certain indices (domains of temperature) such that the landscape created by the model has the same local optima. Either a traditional learning algorithm would work, and once the global optimum (the grammatical form) is reproduced, the other local optima (the performance errors) are given for free by the topology; or the performance errors are also informative, and they provide further information for distinguishing between hierarchies that return the same global optimum. An additional task will be then to fine-tune the frequencies.

A third direction for combining simulated annealing, learning and Optimality Theory is to use simulated annealing not for production, but for learning. SA-OT performs a random walk in the structured candidate set searching for the best candidate with respect to a certain hierarchy. The dual (inverse) problem would be to search the (structured) set of possible hierarchies in order to find the best hierarchy for a certain set of learning data. Each hierarchy is scored by the number of learning data it generates correctly, yielding an integer-valued function to be maximised. Minimal permutations of the hierarchy could be the operation defining the neighbourhood structure. In fact, already Turkel (1994) observed the duality of production and learning in Optimality Theory, and he proposed to use genetic algorithms for both problems, an optimisation technique not very far from simulated annealing (cf. also section 1.2). Nonetheless, applying simulated annealing to the two, dual problems, have only few things in common: very different type of functions have to be optimised on a very different type of search space. I think that the similarities are too few, actually, to have a guilty conscience if I also leave that to future research.

The dual problems are much more closely related in the Maximum Entropy model advanced by Goldwater and Johnson (2003).[10] Although simulated annealing and MaxEnt Optimality Theory are closely related at first sight, the two originate in very different approaches. Yet, some connection could be possible to be worked out through the polynomials used in section 3.3.

As superficially introduced in section 1.3.5, MaxEnt OT defines the probability of form o (derived from input i) as

$$p_{\{r_j\}}(o|i) = \frac{e^{-\sum_j r_j C_j(i,o)}}{Z_{\{r_j\}}(i)} \tag{4.10}$$

Here, r_j is the real-valued rank of constraint C_j, which, in turn, assigns $C_j(i,o)$ violations (not necessarily a non-negative integer) to the input-output pair (i,o). $Z(i)$ is a normalisation factor, not important for us presently.

Observe that the exponent in this expression is a sum with addends composed of two factors. The dual problems, generation and learning, interconnect at this point. In production, the ranks $\{r_j\}$ of the constraints are fixed, and we search for the output o that maximises $p_{\{r_j\}}(o|i)$ for a certain input i. The

[10] See for instance Mullen (2002) for using MaxEnt for parse selection in Dutch, and for further references in the field.

grammar learner, however, varies the ranks $\{r_j\}$, so that the observed input-output pairs have the highest probability. See Jäger and Rosenbach (2006) for an implementation of a simulated annealing-like algorithm to learning in Max-Ent OT, called there *stochastic gradient ascent*, and argued to be a modification of GLA (Jäger, 2003b).