

## Chapter 8

# Conclusion: Is Simulated Annealing Optimality Theory, thus, Better?

### 8.1 Summary

This dissertation aimed at introducing a new variant of *Optimality Theory*, namely, the *Simulated Annealing for Optimality Theory Algorithm* (SA-OT).

After having overviewed existing variants and the “philosophical” background of OT in Chapter 1, Chapter 2 motivates the use of heuristic optimisation algorithms—such as simulated annealing—and then introduces the SA-OT Algorithm (Fig. 2.8, on page 64). The main argument for simulated annealing was that it is a plausible model for the “implementation” of language in the brain: it is fast, efficient, does not require large computational power, but makes *certain* mistakes, the ratio of which increases with production rate. Therefore, SA-OT could be used as a model of some aspects of performance in phonology.

Subsequently, Chapter 3 introduced some formal approaches to OT in order to underpin the SA-OT Algorithm. Both polynomials and ordinal numbers were introduced for that purpose. The following chapter, a set of open questions more than full-fledged proposals, points to related linguistic issues—such as the role of the lexicon and learnability—that should be elaborated in the future. It also introduced a new definition for Output-Output Correspondence.

The remaining chapters present several applications. The goal of these chapters is not so much linguistic, but methodological. Even though I tried to argue for the linguistic well-foundedness of the models, more cooperation with fellow linguists might have been useful here and there to reach an analysis which might withstand linguistic criticism. It is only to be hoped that the model will arouse enough interest among general linguists to help improve these models. More importantly, I urge experimental linguists to provide quantitative experimental data so that the predicted frequencies of new models can be tested in the future against empirical results. Nevertheless, these chapters have hopefully illustrated the methodological issues arising if one decides to use SA-OT as the framework for a linguistic (performance) model.

Indeed, SA-OT is a complex model involving many parameters and many decisions to be made, such as the definition of the topology, the choice of the constraints and their indices (their association with the domains of temperature), and so on. Consequently, it offers the possibility for tuning at many different points. Some may even argue that there are *too many* such points. To this criticism my answer is threefold.

First, SA-OT's aim is to account for a complex quantitative data set (the frequencies of different forms in different conditions), hence the complexity of the model. If the model is simpler than the data to be described, then the model has explained something from the observable complexity. There is an indication that such a reduction in complexity happens when the model correctly accounts for something that has not been aimed at originally: for instance, when the model in Chapter 5 was tuned to return the correct *andante* and *allegro* forms, but it also turned out to correctly predict which word type is more likely to change in fast speech. Second, practice has demonstrated that the high number of parameters does not trivialise the task of tuning the model, and finding a correct model is far from being a sinecure. It is not the case that just “anything” can be reproduced simply using SA-OT. Third, the parameters are restricted by further guidelines. The topology and the constraints should be cross-linguistically universal and well-founded, so they cannot be defined in an *ad hoc* way.<sup>1</sup> Varying certain parameters (typically  $T_{step}$ ) can and should be interpreted as varying the run time of the algorithm (the speed of speech production), whereas varying other parameters (e.g., those related to the *a priori* probabilities) may not have such an interpretation. If the variation depends on the frequency of the word (rare content words being pronounced more carefully than frequent function words), one may tune the parameters of the cooling schedule again, while the *a priori* probabilities of the topology, I hypothesise, accounts for differences among speakers only, since a certain speaker does not alter his or her topology.

In particular, Chapter 5 works out a model accounting for Dutch stress assimilation in normal and fast speech, and thereby analyses the role of parameters  $T_{step}$ ,  $T_{max}$  and  $T_{min}$ . Varying the former is the simplest and probably the most straightforward tool for reproducing fast speech, whereas the later two also have a slight influence on the output frequencies. Besides, this chapter also analysed the role of the definition of the constraint Output-Output Correspondence, employing what had been introduced in the previous chapter.

If Chapter 5 focuses on parameters  $T_{step}$ ,  $T_{max}$  and  $T_{min}$ , then Chapter 6 and section 7.1 add parameters  $K_{step}$ ,  $K_{max}$  and  $K_{min}$  to the analysis. Here, unlike in traditional OT, candidates that can never win and constraints that are vacuously satisfied may significantly interfere with a model's output frequencies. In particular, section 6.5 presents a model—followed by a mathematical discussion and experiments—that relies heavily on parameter  $K_{max}$  in addition to  $T_{step}$ . Due to the infinity of the search space, a larger initial stage in the simulation enhances the “channelling effect”. A similar model appears in section 7.1,

---

<sup>1</sup>Many readers have not been convinced by my arguments for certain topologies being a natural choice in the particular case. Future work should therefore either proliferate the number of phenomena that require a certain topology, so that the choice becomes an unquestionable necessity; or some general principles should determine the topology. For instance, Gerhard Jäger has proposed to connect the neighbourhood structure to a psycholinguistic notion of similarity, whereas Adam Albright has suggested Steriade's P-Map.

which ends by remarking that the different behaviour of the two allomorphs of the Hungarian article can be tuned by varying the indices (domains of temperature) with which the relevant constraints are associated. Introducing empty domains (or inactive constraints) between these two constraints lengthens the period in which temperature is located between these constraints, thereby making the divergence in the behaviour of the two allomorphs more pronounced. This technique is related to the way the first component of the temperature is diminished in the outer loop of the SA-OT Algorithm, hence also an observation on the role of  $K_{step}$ . Finally, the same model makes parameter  $K_{min}$  dispensable by measuring the “specific heat”: the algorithm runs until the random walker has not moved for 30 consecutive iterations.

An additional morale of the first sections in Chapter 5 is that—unlike standard simulated annealing and the SA-OT models presented in earlier chapters—some SA-OT models do not converge to maximal precision if the number of iterations is increased. This remark has also opened some speculation about how to account for linguistic irregularities by using a simple grammar together with an algorithm that is not always correct but which makes predictable errors.

Finally, Chapter 7 (especially subsection 7.1.3 and section 7.2) brings another parameter to our attention, namely, the definition of the topology (the neighbourhood structure). We demonstrate how changing the parameters of the *a priori* probabilities influences the output frequencies. The issue turns more important as the candidate set becomes larger (infinite), and as candidates are assigned a larger number of neighbours.

In what follows, we return to the assessment of the OT variants in section 1.3, and ask the question: is SA-OT any better?

## 8.2 Advantages (and disadvantages) of SA-OT

### 8.2.1 SA-OT and specific linguistic phenomena

Arguments for some approach and against other ones can be of three different sorts. People often present cases where a given model is unable to account for some phenomenon. Second, one may show that an approach is *in general* unable to come to grips with some aspects of the explanandum. Finally, one might formulate “philosophical” preferences and theoretical expectations not matched by that approach. As an example, the reader is referred to Keller and Asudeh (2002)’s criticism of Boersma and Hayes (2001)’s Stochastic OT, replied to by Boersma (2004b).

Concerning the first sort of criticism, I refer to tableau (5.4) on page 128. It shows that for both Types 0 and 2 in Dutch stress assignment, all possible parses of the observed fast speech forms are harmonically bounded, so that therefore, the loser forms cannot win for any hierarchy. This fact could be a counter-argument for all approaches that wish to generate the alternative forms with constraint reranking (such as an *ad hoc* reranking, Anttila’s proposal or Boersma’s Stochastic OT). The same tableau demonstrates why Coetzee’s approach would fail: an attested alternative form violates the highest ranked constraint, hence the critical cut-off point must be above this highest ranked constraint, therefore all candidates should be attested. At the same time, I could argue that Simulated Annealing Optimality Theory does the job nicely.

Although this type of argument may support a certain approach, and can help it become more popular, it is certainly not decisive. Namely, nothing shows that using different constraints would not work within the alternative approaches (cf. e.g., Boersma, 1998a). Add a new constraint to the top of the hierarchy, and the train of thought holds no longer. Keeping in mind that the set of constraints is supposed to be universal across languages, even while it varies across linguists and papers, we have been shown only that we have not been creative enough.

Similar remarks apply to any argument demonstrating that a certain SA-OT model is not able to reproduce a certain phenomenon: in which the output frequencies do not match the experimentally observed frequencies, or in which the local optima are not exactly the attested alternating forms. An example for both was actually the case of Type 2 words—such as *perfectionist*—in Table 5.17 on page 155. But such a fact is not an argument against SA-OT in general, for nothing proves that different rankings, different constraints and different neighbourhood relations would have no chance to work either. This failure only points to the need for future work. Only the repetitive failure of a model and the lack of success might slowly motivate the linguistic community to drop it and to adopt different approaches.

Therefore, we now turn to the second type of arguments. A number of observations show that several linguistic phenomena—such as Dutch stress assignment or the behaviour of the Hungarian article—are about a gradual shift of frequencies in function of certain parameters (speech rate, sociolinguistic parameters, and so forth). This fact is a serious argument against Coetzee’s approach, who refuses to predict the frequencies quantitatively, as well as against Anttila’s proposal, which requires a very different grammar in order to approximate a slightly different frequency distribution (as mentioned by Boersma and Hayes, 2001). Nonetheless, Stochastic OT, MaxEnt OT and SA-OT make it possible to vary the output frequencies as a function of external parameters.

Our SA-OT experiments on the definite article in Hungarian showed how simply the frequency of the most harmonic candidate can be fine-tuned between 0% and 100%. For instance, in Fig. 7.3 (page 201),  $K_{max} = 20$  or 40 never returned the globally optimal form [az#E], whereas  $K_{max} = 3$  would have done it with a frequency close to 100%. On the other hand, remember that in section 6.5 the channelling effect could also enforce the alternate form using a slightly modified tableau: then the global optimum can never be produced in more than half of the cases. Hence, the framework of SA-OT does not restrict the possibilities in a purely categorical fashion. As language data do not seem to be restricted too strongly, either, a strong prediction would be an argument against a certain approach. Indeed, Stochastic Optimality Theory requires the interplay of at least three, almost equally ranked constraints, otherwise it predicts that the grammatical form must have a probability exceeding 50%. While this would seem to count against it, only time can decide which of the two approaches fits better all kinds of linguistic phenomena.

Additionally, the parameter that determines the output frequencies can often be simply interpreted in SA-OT, because it is directly related to the algorithm’s run time. Therefore, fast speech phenomena indeed emerge in a speeded up algorithm. In other cases, nonetheless, similarly to the parameters determining the output frequencies in Stochastic OT and MaxEnt OT, the connection is not so obvious: why would for instance sociolinguistic factors or word frequency

(familiarity) influence  $K_{max}$ ?<sup>2</sup>

A further argument can be brought in favour of SA-OT based on the Dutch fast speech data, which, again, is not decisive, for different constraints might do the job within Stochastic OT, as well. As mentioned, the empirical data display a huge difference in the behaviour of different word types. Output-Output Correspondence (Output-Output Faithfulness) is able to account for the different winning *forms*, but is it also able to account for the different *frequencies*? In Stochastic OT, the frequency of the alternating form was derived from the probability of reranking the constraints at evaluation time. Thus, we have to postulate that different morphological types either employ a different evaluation noise (why?); or associate Output-Output Correspondence with a slightly different rank. The latter possibility is not absurd, for OOC is an odd constraint anyway, and its rank might depend on an argument, the reference string.<sup>3</sup> But how? On the other hand, SA-OT predicted correctly which form is more likely to be mispronounced—a surprising result, since we had not created our model with this goal in mind. The reason why different word types result in different frequencies in SA-OT is that OOC alters the landscape, due to which the local optima have a catchment area (the basin from which rain flows into a particular river) of different size for different inputs. Again, the candidates not appearing on the surface heavily influence the output frequencies.

In brief, while Coetzee’s proposal explicitly rejects accounting for quantitative phenomena that SA-OT can explain, Anttila’s approach is unable to cope with them, but Boersma’s Stochastic OT, similarly to MaxEnt OT, is theoretically able to face them. Nonetheless, there are certainly cases where competitors of SA-OT could turn out to be more convincing.

For instance, it seems to be a coincidence for SA-OT that fast speech prefers the forms that are phonologically less marked, and slower speech is more faithful to morphology. In the constraint reranking approaches, however, these intuitive observations become *the* explanation of the phenomenon. SA-OT’s replies that it is exactly the neighbourhood structure and the “landscape” that explain *why* faithfulness becomes less important and markedness more significant in fast speech: if faithfulness is ranked higher than markedness, then the faithful global optimum seems to be difficult to find in fast speech, and less faithful but unmarked local optima may be returned more easily.

In general, however, a major disadvantage of Simulated Annealing Optimality Theory—compared to its competitors—is that it is hard to understand exactly why it works in certain cases. Developing an exact analysis of SA-OT’s behaviour is difficult even for relatively simple landscapes. The interactions between SA-OT’s components (the neighbourhood structure, the constraint hierarchy and the algorithm’s parameters) are so complex that the success or

---

<sup>2</sup> $K_{max}$  is not necessarily to be interpreted as being connected to the speed of the algorithm. Its role is to determine the length of the initial phase of the simulation, in case the simulation is launched from the same one or few candidates. If the initial candidate is chosen from a wider pool, however, then the initial phase can be omitted. Therefore, different observed frequencies can be reproduced by changing the way the initial candidate is chosen, while  $K_{max}$  (hence, run time) is kept constant. In other words, the random walk in the initial phase can be viewed as not belonging to the SA-OT algorithm, but as a way to chose the initial candidate from this wider pool. Then,  $K_{max}$  is a parameter that determines the choice of the candidate from where (the most interesting part of) the algorithm is launched.

<sup>3</sup>In other words, a whole family of OOC constraints should exist, with each member being associated with a slightly different rank, and each member acting upon a different word type.

failure of a model cannot be simply predicted using paper-and-pen linguistics, without implementing the simulation on computers.<sup>4</sup> (For an extreme case, recall Table 5.10 on page 151 where the frequency of a local optimum surprisingly increased as we diminished  $T_{step}$ .) This is the reason why the reader is welcome to try out the SA-OT demo page at <http://www.let.rug.nl/~birot/sa-ot/>.

## 8.2.2 SA-OT, competence and performance

Finally, let us turn to the third level of criticism against various approaches, that is, to “more philosophical issues”. In particular, we shall ask how SA-OT, as well as its competitor models reflect the traditional dichotomy of competence and performance.

Often, both forms A and B are equally grammatical. In other cases, however, the use of the two forms is not symmetrical, and one may argue for form A to be the “grammatical form”, while B is regarded as a “performance error”—without any value judgement. For instance, in Chapter 5 on Dutch stress assignment, we identified the grammatical form with the *andante* pattern, that is, whose frequency diminishes at higher speech rate. (Otherwise, one should claim that fast speech is more grammatical, which would be odd.) Therefore, we expect a linguistic model to predict which form is grammatical (whatever that means), as well as what other forms may also emerge under certain conditions.

Here I refer to the idea sketched on Fig. 2.1 (page 43), which replaces the competence-performance dichotomy with a three-level picture. Between the competence in-the-narrow-sense (the static linguistic knowledge encoded in one’s brain) and the performance in-the-narrow-sense (including all the extralinguistic factors influencing linguistic products, speech), one also finds the dynamic language production process. Phenomena that are traditionally reckoned to belong to performance, but are determined by linguistic factors, might be analysed on this intermediate level. Consequently, let us ask the various approaches how they distinguish the static knowledge of the language from the dynamic language production, and whether they see a difference between theoretically grammatical forms and forms observable, say, in a corpus.

MaxEnt OT assigns a positive probability to all candidates generated by GEN. Consequently, there is no chance to differentiate in a principled way between forms that are so agrammatical (even absurd) that they cannot be attested for sure, and forms attested, though only rarely—unless one restricts GEN in a language specific manner, or the assigned probabilities drop drastically at a certain point. Such a model might be most welcome in cases where all candidates are attested in a corpus, and only their frequencies require explanation, such as was the case in Jäger and Rosenbach (2006)’s model for English genitive constructions.

Assigning the same violation profile to several candidates, creating a new grammar by reranking some constraints by hand, and the unranked hierarchies of Anttila form the next group of models. They all account for linguistic variation on the level of the competence model, at the core of the OT architecture. These approaches are adequate when the alternating forms are not differentiated with respect to their grammaticality, which definitely is the case in certain forms of variation.

---

<sup>4</sup>Note that Lauri Karttunen has made the same remark on Optimality Theory in general in his FSMNLP talk in Helsinki in September 2005 (cf. also Karttunen, 2006).

Coetzee's model is different. It distinguishes sharply between the form that is grammatical, by making it the optimal output, on the one hand, and alternate forms, the second, third best candidates still emerging, on the other. His model can be, thus, used for phenomena displaying alternation of an arguably grammatical and most frequently occurring form with ungrammatical ones.

Even if it is not necessary, Boersma's Stochastic Optimality Theory may also be seen as reflecting the distinction between competence and performance. Namely, the unperturbed hierarchy can represent competence, in conformance with standard OT; whereas the noisy evaluation cycles are the model for the dynamic language production process. The form optimal for the unperturbed hierarchy is the grammatical form, whereas perturbations explain why other forms are also attested. Postulating, for instance, a larger  $\sigma$  in the evaluation noise at a higher speech rate will account for the increased frequency of the allegro form. The decoupling of the model's two levels is appealing, but it is unclear why noise should increase in fast speech.

I argue that competence and performance are most radically separated in Simulated Annealing OT. Similarly to the unperturbed hierarchy in Stochastic OT, the underlying traditional OT model accounts again for linguistic competence: the grammatical form is the (global) optimum. On top of that, however, we have introduced a separate search algorithm, which models the functioning brain during speech production. The search algorithm is computationally simple, arguably plausible: each time only one form has to be stored, which is then altered in an elementary way, supposing that this basic change does not incur too many extra violations. Not only in memory requirements is the algorithm a plausible model of the brain's functioning, but also in run time, which can be kept constant. Moreover, the precision of the algorithm (the probability of returning the grammatical form) depends also upon parameters that have a direct connection to speech rate:  $T_{step}$  can easily be argued to change in function of speech rate, since it directly influences the algorithm's run time.

SA-OT also competes with different implementations of OT whose goal is to find the optimal candidate in the candidate set. Even if SA-OT does not guarantee that one always finds the optimal candidate, it may still be more adequate in a cognitive sense than its competitors.

Indeed, I argue that simulated annealing is an adequate model for the computations in the human mind for several reasons. First of all, no severe restrictions must be made on GEN and on the constraints as in Finite State Optimality Theory (Eisner, 1997; Frank and Satta, 1998; Karttunen, 1998; Gerdemann and van Noord, 2000; Jäger, 2002; Bíró, 2003, 2005c; Karttunen, 2006). As described elsewhere (Bíró, 2005c), two approaches exist within Finite State Optimality Theory: either a new automaton must be built for each input, requiring huge computational power, or a finite state transducer maps any input to its optimal output, but this latter approach works in very restricted cases only.

Simulated annealing is an algorithm that may find the optimal candidate of a combinatorial problem in a reasonable time with a reasonable precision, even in the case of NP-complete problems, which Optimality Theory may pose (Eisner, 2000b). It does not require computational capacities as large as the ones finite-state approaches need, or even those genetic algorithms or chart-parsing may ask for. Simulated annealing produces *some* output within constant time, similarly to speech, where *something* must be produced within a specific time span—conversation partners are not computer users who are willing to watch

the hourglass! Furthermore, simulated annealing can be speeded up, just as human speech: in both cases the price to pay for shortening the time span is precision. In fast speech, indeed, we accept some ungrammatical forms for the sake of expressing ourselves more quickly. Hence, I argue again, a linguist may gain an explanation of fast speech phenomena, alternation forms or performance errors by adding some topology to the candidate set.

Paul Smolensky's implementation of Harmony Grammar (closely related to Optimality Theory) within a connectionist framework has already long included a notion of neighbourhood, whence follows the possibility of local optima existing in the system. Yet, his goal, which he successfully reaches, is to avoid allowing the system to find them.<sup>5</sup> My approach, however, differs from his not only in that I turn the errors made by the algorithm (when the system gets stuck in local optima) into my advantage; but also in that mine is not confined to neural networks. So non-connectionist scholars—linguists, but also other cognitive scientists—may employ it.

### 8.3 SA-OT as a general cognitive model

More than a decade before 1993, the year when Optimality Theory appeared in linguistics, Seymour Papert (1980) proposed a remarkable cognitive model.

Imagine a child is shown the following experiment: the water is poured from a broader vessel into a narrower one, so the level of the water will be higher than it was in the original vessel. According to observations, children below the age of six or seven will tell you that the amount of liquid has increased, whereas above this age children suddenly change their mind and give an answer in conformance with the principle requiring that the amount of liquid be conserved. How to explain both answers and the switch between them?

Papert suggested the following model (Papert, 1980, p. 166f). Suppose there are (at least) three homunculi present in the brain of a person who has to compare quantities. Each of them works very simply, and the answer of the child is derived from the answers given by these homunculi.

The first of them, Papert proposes, judges the amount of anything according to its height. As objects in the world usually have more or less constant dimensional ratios, judging from the height should be reliable. After all, many important judgements in the world—for instance, the age, role, power and might of an unknown fellow human—can be made based on the other's height. According to Papert, this homunculus serves even the youngest children very well, when they have to distribute Coca Cola or hot chocolate equally among glasses.

The second homunculus relies on the horizontal dimensions. Papert writes that this homunculus is usually not as skilled as the first one, so he influences the judgements of the child much less frequently. He comes to a role in statements such as “there might be really much water in the sea”.

The third homunculus is called History, and teaches that “if two amounts were equal, then they remain equal”. It is a “folk” version of the principle of conservation of matter in physics. Even if we increased the amount of the water, this homunculus of Papert would come to this conclusion.

---

<sup>5</sup>Chapter 20, sections 3.7.4, 3.7.5 of an October 2004 print out of Smolensky and Legendre (2006), which was made available before the KNAW Masterclass “Cognitive Foundations of Interpretation” in Amsterdam.

Now, how to account for the answers of the younger children, and for those of the older ones? When the younger child is asked the question whether the amount of water has changed while pouring it from one vessel to the other, the first homunculus, the vertically minded one, will give the answer.

Concerning the older child, Papert offers three explanations. Either the first two homunculi become more “sophisticated”, so that they interfere only if everything else remains unchanged: for instance, the first homunculus learns to have an opinion only if the width of the object is the same. The second explanation—the most interesting one from the point of view of the developments in linguistics ten years after Papert’s book was published—proposes a reordering in the relative prominence (he calls it the “seniority”) of the homunculi: homunculus History suddenly jumps forward, and becomes the “dominant voice”. Does not this idea remind you of OT learning algorithms? Papert’s third possible answer introduces a fourth homunculus from the critical age onwards that combines the answers given by the first two homunculi (the geometrical ones), so this fourth homunculus will cancel their contradictory opinions.

We can summarise and reinterpret Papert’s model in the following way. In order to solve a cognitive task, the human brain invites several of its “modules” to give some answer. (Modularity of the brain was probably in the air already in 1980.<sup>6</sup>) Out of the pool of possibilities, each module picks one, and returns it as the solution. These modules work in very simple ways, and would quite often mislead the brain if they had to work alone. However, the interaction of them (unspecified by Papert, although he already alludes to some hierarchy in importance among them) results in a cognitive capacity leading to an evolutionarily successful behaviour. The brain does not necessarily return the mathematically exact solution always, and yet, even with such an “imperfect human logic”, humanity has been reasonably successful.

Indeed, this kind of Optimality Theory as a general cognitive strategy, together with simple heuristics, such as “take the higher”, “take the wider”, “quantities do not change”, form the building blocks of the *Heuristic-and-Biases Program* launched by Tversky and Kahneman (1974), and of the *ABC Research Project* (Gigerenzer et al., 1999).

Gigerenzer et al. (1999, p. 24-25) summarise the *ABC Research Project* with the following words:

The research program [...] is designed to elucidate three distinct but interconnected aspects of rationality [...]:

1. *Bounded rationality*. Decision-making agents in the real world must arrive at their inferences using realistic amounts of time, information, and computational resources. We look for inference mechanisms exhibiting bounded rationality by designing and testing computational models of fast and frugal heuristics and their psychological building blocks. The building blocks include heuristic principles for guiding search for information or alternatives, stopping the search, and making decisions.

---

<sup>6</sup>The reader who would like to argue against the strong modularity of the brain in the sense of Jerry Fodor, is welcome to replace the term “module” used here with something like “basic computational unit”, probably smaller ones than those argued for by the proponents of the modularity of the brain.

2. *Ecological rationality.* Decision-making mechanisms can exploit the structure of information in the environment to arrive at more adaptively useful outcomes. To understand how different heuristics can be ecologically rational, we characterize the ways information can be structured in different decision environments and how heuristics can tap that structure to be fast, frugal, accurate, and otherwise adaptive at the same time.
3. *Social rationality.* The most important aspects of an agent's environment are often created by the other agents it interacts with. [...] Social rationality is a special form of models of fast and frugal heuristics that exploit the information structure of the social environment to enable adaptive interactions with other agents. [...]

These three aspects of rationality look toward the same central goal: to understand human behavior and cognition as it is adapted to specific environments (ecological and social), and to discover the heuristics that guide adaptive behavior.

Typical examples for the “fast and frugal heuristics” used in the *ABC Research Program* include for instance: “if one of two objects is recognized and the other is not, then infer that the recognized object has the higher value” (which of the two cities mentioned is larger?, Gigerenzer et al., 1999, p. 41) or “feed your children from youngest to oldest” (Gigerenzer et al., 1999, p. 314). The claim is that modelling decision-making using such heuristics is a cognitively adequate description of the human mind, on the one hand; and that such heuristics are computationally simple, and yet efficient techniques, on the other.

The key to the success of such heuristics is the structure of the world: the structure of the information, of the society, of communication, and so on. In Todd's words: “[i]n our program, we see heuristics as the way the human mind can take advantage of the structure of information in the environment to arrive at reasonable decisions, and so we focus on the inferences” (p. 28).

Still, nothing guarantees avoiding errors. If human mind makes decision based on such heuristics, then... *errare humanum est!* But the interpretation of these errors had changed much in 25 years: what was seen by the heuristics-and-biases program (Tversky and Kahneman, 1974) as a hindrance to sound reasoning (“rendering *Homo sapiens* not so sapient”, Gigerenzer et al., 1999, p. 29), is perceived by the *ABC Research Group* rather as “enabling us to make reasonable decisions and behave adaptively in our environment—*Homo sapiens* would be lost without them” (*ibid*).

In the context of the cognitive research lines just described, Prince and Smolensky (1993)'s *Optimality Theory* can be seen as a concrete case for the specific cognitive subfield of language. OT constraints are similar heuristics, that is, simple rules to evaluate the possibilities (candidates): “take the one with the least codas”, “take the one with the most onsets”, “take the one with the least epenthetic segments”, and so forth.<sup>7</sup> Finding the most harmonic candidate (“take the best” for the ABC Research Group) is performed in OT with

---

<sup>7</sup>Here I leave open the question whether “simplicity” is also meant in computational terms, in the form of requirements that, for example, constraints must be “primitive”, finite-state friendly (Eisner, 1997; Bíró, 2003). Even constraints that do not meet these requirements are immensely simpler than grammars themselves.

respect to the *lexicographic strategy*, one of the three possibilities besides the linear model and the classification trees (Gigerenzer et al., 1999, p. 136-139). Cognitive research on decision making in general, therefore, corroborates the use of Optimality Theoretical grammars, and the emergence of OT-style linguistic systems during the evolution of general cognitive skills becomes more plausible.

A major difference still remains, however. In Optimality Theory, the candidate optimising the constraints is *by definition* the solution sought, whereas “fast and frugal heuristics” aim only at *approximating* the solution of the complex problem posed to the cognitive faculties (by seeking “good (i.e. near-optimal) solutions at a reasonable computational cost”, Reeves, 1995, p. 6). Indeed, “fast and frugal heuristics” are employed to answer problems under time pressure and when information may be incomplete. In a different study, Gary Klein reports that fireground commanders make around 80 percent of their decisions in less than one minute, sometimes within a few seconds, whereas chess players under blitz conditions make a move in average in six seconds Klein (1999, p. 4).

This difference between Optimality Theory and its cognitive background, “fast and frugal heuristics” can be explained, though. Suppose that (OT-style) language evolved from the heuristic inference system, and used its architecture to define the rules of linguistic communication. Unlike in the case of a question such as “which amount of water / which city is larger”, however, there is no *a priori* uniquely good solution to the problem how to encode a thought into a utterance. Hence, the system that had had only limited precision when solving cognitive problems, could now perfectly encode the rules of language—just because these rules<sup>8</sup> were formulated in terms of this system.

And yet, this new system of communication did not work so perfectly—maybe due to the proliferation of meanings to be expressed, leading to the proliferation of lexical items and possible structures. Here came a second level of heuristics into play, at least according to the main claim of my thesis. Even though the best candidate sought after is well defined in terms of the constraints, still, it is not always possible to find it at production time. Now we move from the first level to the second level on Table 2.1. Once the language community has accepted a form as the grammatical (the optimal) one, locating it becomes an analogous task to finding the *a priori* correct answer to any other question faced by the cognitive system: the individual is expected to make her utmost effort to approximate the correct solution as closely as possible, within a reasonable time, and by using a limited computational capacity.

Therefore the individual will utilise heuristic techniques again. Even though simulated annealing as a “heuristic optimisation algorithm” is “heuristic” in a very different sense from the way the ABC Research Group employs the word “heuristic”, some crucial similarities still point to a possibly deeper connection. Namely, the tolerance of errors, as well as the use of the information’s structure. Errors often emerge from the trade-off between the precision required by the situation, on the one hand, and the computational resources and time available to the system, on the other. Moreover, the structure of the wor(1)d—*i.e.*, of the search space—is taken into account. That is to say, the “fast and frugal heuristics” of the *ABC Research Group* have developed during evolution so that they reflect the structure of the world, and thereby help increasing precision;

---

<sup>8</sup>The word “rules” does *not* refer here to traditional generative rewrite rules, whose dismissal was exactly OT’s main agenda. By “rules” I simply mean the laws governing language.

while in SA-OT, the neighbourhood structure mirrors what GEN does, due to which SA-OT can exploit features of the candidate set.

Additionally, it can be argued that the larger human cognitive system also employs some “heuristic optimisation algorithm” that has features remarkably reminding us of simulated annealing. Klein (1999, p. 30) observes that “[d]ecision makers usually look for the first workable option they can find, not the best option”, that “they do not have to generate a large set of options to be sure they get a good one”, and that “[t]hey generate and evaluate options one at a time and do not bother comparing the advantages and disadvantages of alternatives”.<sup>9</sup> All that because “[t]he emphasis is on being poised to act rather than being paralyzed until all the evaluations have been completed”. Later, Klein (p. 287) summarises his model for human decision making as “analytical”: “... generative, channeling the decision making from opportunity to opportunity rather than exhaustively filtering through all the permutations”. He even adds that human decision making mechanisms “trade accuracy for speed and *therefore* allow errors” [emphasis added—T. B.].

To turn to a very different domain, to believe systems, Bainbridge (2006) introduces simple connectionist networks to model agents in a society. He concludes then that “[l]ocal minima are actually very interesting, because they represent a very human quality: a reluctance to give up beliefs that function pretty well at the cost of never finding the real truth. One way of expressing the thesis of this book is to say: *Religious faith is a local minimum*” (p. 83, italics in the original text). Hence, unlike Klein’s firefighters, but like the random walker in SA-OT, Bainbridge’s believer agents are happy with being stuck in a local optimum.

In sum, I propose to view linguistic competence in its narrow sense (the first level on Table 2.1) as a by-product of the heuristics used by the human cognitive capacities. Then, on a second level, when the individual comes to produce the grammatical form defined by the first level and accepted by the language community, then these (or similar) heuristics ought to be used again, in order to solve an otherwise computationally challenging task. As a result, *errare humanum est*, even in matters of language, which is a domain created by the human mind. For is it not surprising that the system of communication developed by the human mind poses to the same mind problems whose difficulty is similar to the difficulty of the problems posed by external factors? Why is it so hard to find the right words?

---

<sup>9</sup>This last observation does not contradict the argument that Klein’s decision makers follow an algorithm similar to simulated annealing. Namely, he does not describe here *how* the decision maker comes up with particular options on a micro-level, but this process could be imagined as the random walk in gradient ascent or simulated annealing. My only point here is that neither Klein’s model, nor simulated annealing perform global comparisons or comparisons of distant options.

An important difference is indeed that Klein’s model, unlike simulated annealing, is able to decide whether a particular option (a local optimum) is in itself “good enough” or the search should be continued further. On the other hand, if the experienced speaker could somehow judge whether the local optimum returned by SA-OT is “good enough”, then the precision of SA-OT could be improved. So, similarly to Klein’s firefighter who decides to search further if the solution arrived at involves too much risk, the speaker would also run another simulation if the locally optimal output still incurs too many violation marks.