Elephants and Optimality Again

SA-OT accounts for pronoun resolution in child language

Tamás Biró

University of Groningen and University of Amsterdam

Abstract

Children display a surprising delay in correctly resolving pronouns, while they employ Chomsky's binding principles correctly in production and in resolving reflexives. We account for the mistakes as performance errors that are predicted by an Optimality Theoretical model implemented using simulated annealing. Our experiments suggest three novel explanations of the facts. Additionally, the Optimality Theory–Harmony Grammar connection is also explored: the behaviour of the HG-based performance model converges to the OT-model if the base of the exponential weights grow large.

1   Introduction

Probably *the* central question in cognitive research on child language acquisition is what children miss: the lack of what accounts for child language errors?

Many suggestions have been advanced, reflecting the theoretical preferences of their proponents. The "orthodox innativists" aim at minimising the complexity of the learning task: What the child does not know is the correct value of some (binary) parameters (Chomsky 1981) or the correct ranking of some constraints (Prince and Smolensky 1993/2004). Many scholars adopt these frameworks—*Principles and Parameters* (P&P) and *Optimality Theory* (OT)—for their technical soundness, without subscribing necessarily to their underlying philosophy; hence, they often suggest that children also have to discover the principles or the constraints themselves. A third approach again argues that children are even short of a major component of their "language organ", that is, they first acquire some mechanism defining how to use correctly, say, the parameters or the constraints. Unlike in the first two approaches, where children miss some details specific to their future mother tongue, here children (also) do not have something universal in the adult languages of the world. Finally, a fourth approach emphasises that children (also) have insufficient performance resources, such as, working memory or computing time.

The present article will not do justice to these approaches. It rather presents how a certain child language phenomenon—namely, pronoun resolution—can be explained by several of these approaches alike. Deferring the comparison to psycholinguistic research, we shall focus here on demonstrating that a particular model of linguistic competence and performance (the SA-OT Algorithm) has

the potential to reproduce observed data. Emphasis is put on the computational, rather than on the cognitive or psycholinguistic aspects.

Section 2 introduces the pronoun resolution task, also referring to some former accounts for child language data. Then, section 3 reviews how simulated annealing, either in Optimality Theory or in a symbolic Harmony Grammar, becomes a model of linguistic performance. A Simulated Annealing for Optimality Theory-model for pronoun resolution will be developed in section 4. Finally, section 5 reports on concrete experiments with this model, summarised in section 6.

2   The pronoun resolution problem in child language

Recovering the referent of a pronoun is unquestionably a challenging subtask of sentence comprehension, both in language technology (for an overview, see (Jurafsky and Martin 2009), pp. 729-754) and in cognitive modelling. See for instance Reuland (2001) for a minimalist approach. The context comprises an (often poorly defined) set of *entities*—persons, objects, etc.—and the listener has to reconstruct the element or subset to which the speaker is referring. A number of cues help the listener, such as grammatical markers (number, gender), semantic constraints (certain thematic roles must be filled only with human agents, edible substances, etc.), knowledge of the world and of the specific context, and so on.

Nine out of ten languages employ a set of *reflexive pronouns* that are different from the personal pronouns,[1] and arguably universal principles guide their use. The structure of the sentence determines the domain within which the antecedent of the pronoun must be found, if it is a reflexive, and within which the antecedent must not be found, if it is a personal pronoun. So the distinction between reflexives and personal pronouns is an additional factor diminishing the ambiguity of language—at least, for adults.

A number of experiments have, namely, established that children at the age of four to six allow personal pronouns to have reflexive meanings, as well. (See Spenader et al. (2009) for a comprehensive overview of the relevant experimental literature, as well as of former theoretical accounts of the observations. Our discussion of the phenomenon shall closely follow their article.)   For instance, take two pictures, both depicting an elephant and an alligator, and the elephant hitting the alligator on the first one and himself on the second one.[2] Then, show one of them to a child, accompanied by one of the following sentences:

(1)   a.   *Here you see an elephant and an alligator.*
              The elephant is hitting him.
        b.   *Here you see an elephant and an alligator.*
              The elephant is hitting himself.

---

[1]According to the *Typological Database Nijmegen*, only in 13 out of 113 languages are the reflexive pronouns the same items as the personal pronouns. See the *Typological Database System* (TDS) at `http://languagelink.let.uu.nl/tds/main.html`.

[2]See the pictures drawn by Robbert Prins on the website of Petra Hendriks: `http://www.let.rug.nl/hendriks/vici.htm`.

In more than 80% of the cases, children correctly interpret sentence (1b): the sentence is true only for the picture with the elephant being hit. Yet, sentence (1a) is significantly more often misunderstood: children can accept it as describing both pictures. Surprisingly, in their production the same children use both pronouns correctly (that is, as in adult speech) in more than 80% of the cases. Thus, we observe that children *produce* a structure correctly earlier than they *interpret* it correctly, which is just the reverse of what one would expect.

How can we explain that the interpretation performance is between 50% to 80% (depending on the protocol used by the experimenters), whereas the production performance is significantly above 80%?

Chomsky (1981) introduces the following well-known *principles* of his Binding Theory (p. 188):[3]

(2)   *Principle A*: An anaphor is bound in its governing category.
       Principle B: A pronominal is free in its governing category.
       Principle C: An R-expression is free.

Consequently, the pronominal *him* in sentence (1a) must refer to the alligator; should it refer to the elephant, Principle B would be violated. Similarly, the anaphor *himself* of sentence (1b) has to refer to the elephant, otherwise Principle A would not be satisfied. Children fail to reproduce this logic when interpreting sentence (1a), despite their correct use of the same principles when producing similar sentences or when interpreting sentence (1b).

A number of explanations can be advanced. For instance, children might not have learnt the exact meaning of the personal pronouns yet: they could believe that pronominals also have a reflexive meaning. Were this the case, however, one would expect a production pattern that is as erroneous as the comprehension pattern. Within the four types of explanations of child language phenomena mentioned in section 1, this explanation can be categorised as parameter setting: the child has not yet acquired that the parameter "personal pronoun also has reflexive meaning" must be set to NO. (As mentioned earlier, approximately one tenth of the languages of the world has this parameter set to YES.)

The second of the four explanation types in section 1 would argue that children do not have Principle B yet, which would exclude the reflexive interpretation of sentence (1a). A sentence such as *He looks like him* seems to violate Principle B, unless one realizes that the two pronouns must not be coindexed, and such apparent violations of Principle B might delay its acquisition—unlike the acquisition of Principle A, which is never violated even on the surface. A similar proposal has been suggested by Chien and Wexler (1990). Yet, they remain faithful to Chomsky's ideas, and argue (even experimentally) that Principle B is innate; what has to be learnt is a slightly different "Principle P" (but with the same effect for our purposes), which they expel to pragmatics.

Third, Hendriks and Spenader (2005/2006) argue that children lack a major

---

[3] An anaphor is a reflexive in Chomsky's terminology, a pronominal is a personal pronoun, an R-expression is an expression directly referring to an entity in the world. "Being free" is being unbound.

component of the competence mechanism. In their analysis Hendriks and Spenader reduce Principle B to Principle A combined with a bidirectional optimisation procedure in Optimality Theory. By taking the speaker's perspective also into account, the listener reasons as follows: "if the speaker meant a reflexive meaning, she would have used a reflexive pronoun; hence, the fact that she used a personal pronoun indicates that she meant a non-reflexive meaning". What children are short of, Hendriks and Spenader propose, is a Theory of Mind, that is, the ability to take the speaker's perspective while interpreting sentence (1a). In other words, children miss the bidirectional optimisation mechanism, an important constituent of the (adult) linguistic competence.

Finally, as an example for an explanation blaming performance, let us refer to Reinhart (2004), who argued that children at this age have insufficient working memory to perform the high complexity mental computations required by a sentence such as (1a), as opposed to the more easily computable sentence (1b).

The explanations to be advanced are somehow similar to Reinhart's account in that the emphasis is put on performance. But before presenting the new explanations, we need to review quickly the *Simulated Annealing for Optimality Theory Algorithm* (Bíró 2005a, 2005b, 2006).

3   Simulated annealing as performance

Both Smolensky and Legendre (2006) and Bíró (2006) view the Chomskyan distinction of linguistic competence and performance as the distinction between a grammar and its implementation. Within the Optimality Theoretical camp, a grammar is a *harmony function* $H(w)$ over possible forms (candidates). The range of this function is a totally ordered set, and the grammar defines a mapping from an underlying representation U to the corresponding surface representation as

$$SR(U) = \arg\max_{w \in Gen(U)} H(w) = \arg\min_{w \in Gen(U)} E(w) \tag{3}$$

Here $Gen(U)$ is the set of candidates corresponding to the underlying form U, generated by the *Gen* function. Instead of maximising the harmony function $H(w)$, one can also minimise its inverse, the *energy* $E(w)$ (Bíró 2006). This approach has the advantage that it directly reflects the standard idea: violation marks (the famous stars in OT tableaux) are to be minimised.

The energy (harmony) function is composed of elementary functions $C_i(w)$ (called "constraints" for historical reasons). For instance, a (symbolic) *Harmony Grammar* can be defined as the linear combination of those constraints:

$$E(w) = g_n \cdot C_n(w) + ... + g_i \cdot C_i(w) + ... + g_0 \cdot C_0(w) \tag{4}$$

From now on, we shall use an *exponential weight system* with base $q > 1$, and call the grammar thus defined a *q-HG grammar*:

$$E(w) = C_n(w) \cdot q^n + ... + C_i(w) \cdot q^i + ... + C_1(w) \cdot q + C_0(w) \tag{5}$$

Optimality Theory (OT) differs from Harmony Grammar by requiring *strict domination* (Prince and Smolensky 1993/2004, Smolensky and Legendre 2006, Jäger and Rosenbach 2006): the effect of a higher ranked constraint cancels the eventual (joint) effects of any lower ranked constraints. The order (comparison) of $H(w_1)$ and $H(w_2)$ in (3) depends only on the highest ranked constraint such that $C_i(w_1) \neq C_i(w_2)$, and all lower ranked constraints do not influence the comparison. After introducing the convention that constraints are indexed such that $i > j$ corresponds to $C_i \gg C_j$, strict domination can be realized by taking the $q \to +\infty$ limit in eq. (5) (or by having $q = \omega$ where ω is the first transfinite ordinal; cf. Bíró (2005a), as well as Bíró (2006) Chapter 3). In section 5 we shall compare the behaviour of an OT grammar to the behaviour of the corresponding q-HG grammars with different q values. "Corresponding" means that the same constraints are employed, and if constraint C_i is ranked higher than C_j in the OT grammar, then the weight of C_i is greater than the weight of C_j in the HG grammar.

To summarise, the *linguistic competence* of a child or of an adult will be modelled using an OT grammar or using a q-HG grammar. To mimic their *linguistic performance*, we need an implementation of that grammar: an algorithm that computes the *SR(U)* function as defined in eq. (3).

Simulated annealing, a variant of the hill climbing algorithm, has been employed for that purpose. Here I summarise the *Simulated Annealing for Optimality Theory Algorithm* (SA-OT), as introduced in my earlier work, which is a symbolic algorithm, unlike Smolensky's similar but connectionist implementation.

First of all, a *neighbourhood structure* (or *topology*) has to be introduced on the candidate set. It shall serve as the "horizontal" dimension of the landscape, in which a random walk will be launched. The goal of the walk is to find the (globally) optimal candidate, and to settle there. The output of the algorithm is namely the final position of the walker. The rules of the random walk on the candidate set are simple: in each iteration,

1. the walker randomly chooses a candidate w' that is a neighbour of its current position w,

2. and then the walker moves from w to w' with a *transition probability* that exclusively depends on the difference $E(w') - E(w)$ (the height of the step to be taken), but not on w or w'.

In the case of a classical hill climbing algorithm, the transition probability is 1 if $E(w') - E(w) \leq 0$, and it is 0 if $E(w') - E(w) > 0$. The walker always steps downhill, and never uphill. Obviously, a local minimum will trap the random walker. Similarly in (the current version of) simulated annealing, the random walker unquestionably moves to w' if it is a more harmonic candidate (that is, if its harmony is higher / its energy is lower than the one of w). But there is also a chance of moving to a less harmonic neighbour, at least in the initial phases of the simulation, in order to improve the chance of escaping from undesirable local optima and to settle finally on the globally optimal candidate.

If the neighbourhood structure is the horizontal dimension and $E(w)$ is the vertical dimension of the landscape in which the random walker searches for the deepest point, then simulated annealing can be visualised as follows: Initially, the random walker is "full of energy", and so it can move both uphill and downhill. The transition probability is (practically) 1, independently of the sign of $E(w') - E(w)$ (*first phase*). Later on, in the *second phase*, the "lazy" random walker will be less likely to take larger steps upwards: the transition probabilities for the cases $E(w') - E(w) > 0$ gradually decay, and this decay is faster for larger $E(w') - E(w)$ values. Finally, in the *last phase*, the transition probabilities become (practically) zero if $E(w') - E(w) > 0$: the random walker is now so "tired" that it will only move downhill, to the bottom of the current "valley".

In the case of traditional simulated annealing, when the target function $E(w)$ is real-valued (such as in the case of a q-HG grammar), the transition probability P is usually defined as

$$P(w \rightarrow w' \mid T) = \begin{cases} 1 & if E(w') \leq E(w) \\ e^{-\frac{E(w') - E(w)}{T}} & if E(w') > E(w) \end{cases} \tag{6}$$

where $T > 0$ is called the *temperature* (recalling the origins of the algorithm in statistical physics); hence the name of the algorithm. As the algorithm is proceeding, T is gradually decreased, according to some *cooling schedule*. In the first phase described above, T is much greater than the differences $E(w') - E(w)$ for neighbouring w and w' candidates. In the last phase, T is much less than the same differences. Figure 1 summarises this (real-valued) simulated annealing algorithm. The *stopping condition* can have the form $T < T_{\min}$, or may require that the random walker has not moved for a given amount of iterations.

In the case of Optimality Theory, the target function $E(w)$ is not real-valued. Therefore, definition (6) has to be replaced by the following transition probabilities, as argued by Bíró (2005a, 2005b, 2006):

$$P(w \rightarrow w' \mid T) = \begin{cases} 1 & if\, d \leq 0, or\, if\, d > 0\, and\, k < K \\ e^{-d/t} & if\, d > 0\, and\, k = K \\ 0 & if\, d > 0\, and\, k > K \end{cases} \tag{7}$$

where the temperature has the form $T = \langle K, t \rangle$; moreover, C_k is the highest ranked constraint by which w and w' are assigned a different number of violation marks; finally, $d = C_k(w') - C_k(w)$. The role of the difference $E(w') - E(w)$ is taken over by the pair $\langle k, d \rangle$: a difference of d violation marks for constraint C_k. Note that constraints ranked higher than C_k do not distinguish w and w', whereas constraints ranked lower than C_k are ignored, in consent with the philosophy of strict domination. Figure 2 presents the *Simulated Annealing for Optimality Theory Algorithm* (SA-OT).

In SA-OT, the temperature $T = \langle K, t \rangle$ is decreased by two, embedded loops: the outermost decreases K, whereas the inner loop decreases t. The index k of

```
ALGORITHM: Simulated Annealing
 Parameters:  w_init     # initial state (often randomly chosen)
              T_max      # initial temperature > 0
              alpha(t)   # cooling schedule
 w := w_init ;
 T := T_max  ;
 Repeat
    Randomly select w' from the set Neighbours(w);
    Delta :=  E(w') - E(w) ;
    if ( Delta < 0 )
       then  w := w' ;      # always move to a better neighbour
       else  # move to w' with transition probability P(Delta;T):
           generate random r uniformly in range (0,1) ;
           if ( r < exp(-Delta / T) )
                 then   w := w' ;
           end-if
    end-if
    T := alpha(T)  # decrease T according to cooling schedule
 Until stopping condition = true
 Return w          # w is the approximation to the optimal solution
```

Figure 1: Minimising a real-valued energy function $E(w)$ with simulated annealing.

a constraint is called its *K-value*; a higher ranked constraint must have a higher K-value. As long as the first component K of the temperature is greater than the K-value of the highest ranked constraint, the simulation is in its first phase, as described above. The simulation enters its last phase when the component K of the temperature becomes less than the K-value of the lowest ranked constraint. In the middle phase, steps increasing the violation of higher ranked constraints are prohibited, but steps increasing the violation of lower ranked ones are permitted; the first component of T tells you where the border is actually between "higher ranked" and "lower ranked" constraints.

The central question is the probability that the random walker settles in the global optimum in the final phase. This probability is the *precision* of the algorithm, that is, the likelihood that the implementation of the grammar returns the *grammatical form*, as predicted by eq. (3). Otherwise, the random walker will settle in some other local optimum, which is predicted to be a *performance error form*. A basic fact concerning simulated annealing is that the higher the number of iterations in the second phase, the closer to 1 the precision of the algorithm—at least, in the case of a real valued target function.

In the case of SA-OT, however, it has been demonstrated that there are certain grammars (constraint rankings combined with neighbourhood structures) for which the number of iterations in the second phase does not influence the precision. In the following sections, we shall analyse a model that is analogous to the one presented in Chapter 6 of Bíró (2006). Another example is presented by Bíró (2007), who also introduces the following terms for the two kinds of unwanted outputs: *fast speech forms* are local optima that are returned less frequently with more

```
ALGORITHM: Simulated Annealing for Optimality Theory
 Parameters: w_init, K_max, K_min, K_step, t_max, t_min, t_step
 w := w_init ;
 for K = K_max to K_min step K_step
     for t = t_max to t_min step t_step
         Randomly select w' from the set Neighbours(w) ;
         C    := highest ranked constraint s.t. C(w) != C(w') ;
         k(C) := K-value of constraint C ;
         d    := C(w') - C(w) ;
         if ( d < 0 )
             then  w := w' ;     # move to better neighbour
             else  w := w' with probability
                     P(C,d ; K,t) = 1            , if k(C) < K
                                  = exp(-d/t) , if k(C) = K
                                  = 0            , if k(C) > K ;
         end-if
     end-for
 end-for
 return w
```

Figure 2: The Simulated Annealing for Optimality Theory Algorithm (SA-OT).

iterations, whereas *irregular forms* persist even at a high number of iterations. In other words, if SA-OT is slowed down (the number of iterations is increased), then fast speech forms disappear, and the performance of the algorithm converges to a distribution between the grammatical form(s) and the irregular form(s).

After having introduced how OT grammars and q-HG grammars can be implemented using simulated annealing, we now return to pronoun resolution.

4   An SA-OT model for pronoun resolution

Next, we introduce an OT-style grammar for pronoun resolution. Unlike in a theoretical linguistics paper, we shall not elaborate in length on the different components of this model. Some of the choices are straightforward, and others will be justified by the success of the model in section 5. The violation levels in the cells of tableau (1) are crucial, while the exact formulation of the constraints can be debated on theoretical ground. Here I focus on creating a model whose behaviour is analogous to the one in Chapter 6 of Bíró (2006) on voice assimilation.

During the task of sentence interpretation, the input is a sentence and the candidates are possible interpretations of that sentence. Specifically, if the input is sentence (1a), "*the elephant is hitting him*", and the question is what the pronoun *him* refers to, then the possible candidates include the alligator, the elephant, both of them, none of them, or even some third or fourth referent. In short, we shall say that a candidate is a reference set, and a reference set is a subset of the set of the entities in the world $\{a, e, x, y, z, ...\}$. Here, a is a shorthand for the alligator and e stands for the elephant.

We shall discuss two separate models. In the first one, only a and e are consid-
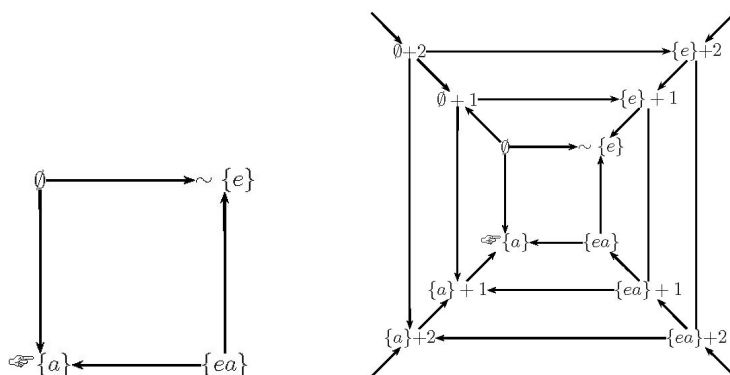
Search spaces (neighbourhood structures, topologies):
Figure 3: (Left) Candidate set 1 *without* insertion of elements not present in the context.
Figure 4: (Right) Candidate set 2 *with* insertion of additional elements.

ered, whereas the second one also allows the insertion of other entities (not present in the immediate context) into the reference set. Hence, the candidate set of the first model contains the following four candidates: the empty set $\emptyset$, the set $\{a\}$, the set $\{e\}$ and the set $\{a, e\}$. We shall refer to these simply as 0, a, e and ea. Besides them, the candidate set of the second model also contains the union of the previous sets with the subsets of $\{x, y, z, ...\}$. Then, e+1 will denote the set that contains the elephant e and some third entity of the world; candidate ea+2 is a set that contains not only the elephant and the alligator but also two further referents; and candidate 0+k contains k entities (but not the elephant or the alligator).

Having defined the candidate set(s), we next have to define a neighbourhood structure (topology) on it. We shall say that candidate w' is a neighbour of candidate w, if and only if w' can be constructed by removing exactly one element from w or by adding exactly one new element to w. Figures 3 and 4 illustrate the neighbourhood structure of the first and of the second model, respectively. (The arrows point to the more harmonic candidates, as defined by hierarchy (8) below.) Moreover, neighbours have equal chance to be chosen before each step.

Third, let us introduce the following constraints:

> PROKNOWN: Reference set must include object from context.
> AGRNUMBER: reference set cardinality must be 1.
> NO3RD: number of inserted elements.
> PRINCIPLEB: *elephant* not in reference set.

Constraint PROKNOWN (or PRO) requires that the candidate (the reference set) contain at least one element of the context (the elephant or the alligator), possibly beside other entities. You do not use a pronoun, if you refer exclusively to entities not yet present in the context or in the discourse, do you? Thus, candidates 0 and 0+k (for any $k > 0$) violate this constraint, and all other candidates satisfy it.

		PRO	AGRNUMBER	NO3RD	PRINCIPLEB
	0	1	1	0	0
~	e	0	0	0	1
	ea	0	1	0	1
☞	a	0	0	0	0
	0 +1	1	0	1	0
	e +1	0	1	1	1
	ea +1	0	1	1	1
	a +1	0	1	1	0
	...				
	0 +k	1	1	k	0
	e +k	0	1	k	1
	ea +k	0	1	k	1
	a +k	0	1	k	0
	...				

Table 1: OT tableau corresponding to hierarchy (8).

The second constraint, AGRNUMBER is the traditional agreement requirement for grammatical number. As the pronoun *him* is singular, only the candidates representing a singleton (a, e and 0+1) satisfy this constraint. These first two constraints are highly influential in languages, and so we expect them to be ranked high.

Constraint PRINCIPLEB is an OT-version of Chomsky's Principle: it simply prohibits an interpretation such that the pronoun in the object position is coindexed with the subject. For our purposes, a candidate violates this constraint if and only if the reference set contains e (candidates e, ea, as well as e+k and ea+k).

These three constraints are binary: if candidate w satisfies constraint C_i, then $C_i(w) = 0$, whereas violation means that $C_i(w) = 1$.

Constraint NO3RD is not binary. In order to discourage insertion of additional elements in the second model, it assigns one violation mark to each referent not in the context. $C_{\mathrm{No3rd}}(w)$ is the cardinality of the set $w \setminus \{a, e\}$. Such a constraint is necessary, because without it the model could not make a difference between neighbouring candidates that differ only in the amount of inserted elements.

The last ingredient to an OT grammar is the constraint hierarchy. The following ranking is the one that yields the expected results in the next section:

$$\text{PRO} \gg \text{AGRNUMBER} \gg \text{NO3RD} \gg \text{PRINCIPLEB} \tag{8}$$

In the resulting tableau 1, the hand symbol ☞ points to the globally optimal candidate, whereas ~ (recalling the symbol usually used to denote variation) stands next to further local optima. (The tableau does not reveal whether they are fast speech forms or irregular forms.) One can simply test that candidate e is more harmonic than its neighbours, 0, ea and e+1; and in general, that the arrows on Figures 3 and 4 point to the more harmonic candidates.

5   Running the simulation

In the previous section, we have introduced a model in which the grammatical form {*a*} is the globally optimal candidate, and there is another local optimum: {*e*}. We shall now run the performance algorithms so that we can observe which conditions will return the global optimum only (corresponding to adult performance), and which conditions also return the second local optimum (corresponding to child performance).

For the OT grammar defined by hierarchy (8), we employed the SA-OT Algorithm on Fig. 2. The four candidates of the first model (Fig. 3) became the initial point of the random walk with equal frequency. The standard values $K_{\text{step}} = 1$, $t_{\max} = 3$ and $t_{\min} = 0$ were used as further parameters of the algorithm. Instead of employing some very small $K_{\min}$, we stopped the algorithm as soon as it had not moved for 50 iterations (corresponding to a vanishing probability of stopping the algorithm while it has not reached a local optimum yet). Two further parameters, $K_{\max}$ and t_{step}, remain to be set: we shall play with them, besides comparing the two candidate sets and comparing OT grammars to q-HG grammars.

Conform to hierarchy (8), we have assigned the K-value of 3 to the highest ranked constraint PROKNOWN, 2 to AGRNUMBER, 1 to NO3RD, and finally 0 to PRINCIPLEB. The same values will be used as the indices (exponents) i of the constraints in the q-HG grammar (5), and consequently real-valued simulated annealing (Fig. 1) can implement q-HG grammars.

In order to be able to compare OT grammars to q-HG grammars, we need comparable cooling schedules. As explained earlier, a cooling schedule can be divided into three phases: the second one begins when the transition probability to some neighbours is not (practically) 1 anymore, whereas the third phase begins when moving to any less harmonic neighbour becomes (practically) impossible. Two cooling schedules are comparable if both their first and their second phases include a comparable number of iterations. (The length of the third phase only depends on the time to reach to bottom of a valley.) In order to operationalise this idea, we define the phases of the cooling schedule as follows: the transition probability from w to w' is $1/e$ at the end of the first (second) phase, if w and w' only differ in that w' incurs one more violation of the highest (lowest) ranked constraint than w does.

Let us suppose that w' is worse than w for constraint C_i by one violation ($C_i(w') = C_i(w)+1$), but otherwise they are the same ($C_j(w') = C_j(w)$ for $j \neq i$). Then the transition probability (6) in q-HG will be $1/e$ if $T = q^i$. Similarly, in SA-OT, with transition probabilities (7), the same condition requires temperature T to be $\langle i, 1 \rangle$.

Next, let us denote the K-value of the highest ranked constraint (PROKNOWN in our case) with κ (presently $\kappa = 3$). The K-value of the lowest ranked constraint is standardly 0, and $K_{\text{step}} = 1$. Consequently, the first phase of SA-OT involves $\phi_1 = (K_{\max} - \kappa) \cdot \frac{t_{\max}-t_{\min}}{t_{\text{step}}} + \frac{t_{\max}-1}{t_{\text{step}}}$ iterations, whereas the second phase $\phi_2 = \kappa \cdot \frac{t_{\max}-t_{\min}}{t_{\text{step}}}$ iterations.

For a q-HG grammar, large differences in $E(w)$ necessitate temperature to de-

q	precision
OT	0.503 ± 0.004
30	0.498 ± 0.003
20	0.499 ± 0.005
10	0.506 ± 0.007
5	0.515 ± 0.005
3	0.554 ± 0.005
2.5	0.583 ± 0.004
2.0	0.632 ± 0.006
1.8	0.669 ± 0.004
1.7	0.690 ± 0.003
1.6	0.715 ± 0.003
1.5	0.755 ± 0.003

q	precision
1.40	0.794 ± 0.002
1.30	0.844 ± 0.003
1.20	0.910 ± 0.004
1.15	0.948 ± 0.002
1.10	0.978 ± 0.001
1.08	0.988 ± 0.001
1.06	0.994 ± 0.001
1.05	0.9964 ± 0.0007
1.04	0.9978 ± 0.0005
1.03	0.9991 ± 0.0001
1.02	0.9998 ± 0.0001
1.01	0.99992 ± 0.00008

Table 2: Precision of different types of grammar (candidate set 1, $K_{\max} = 5$, $t_{\text{step}} = 0.1$).

cay exponentially. Let r denote the ratio by which T is decreased in each iteration: $\alpha(T) = r \cdot T$ in Fig. 1. At the end of the first phase, $T = q^{\kappa}$; at the end of the second one, $T = q^0 = 1$. For the second phase to have ϕ_2 iterations, r must be

$$r = q^{\frac{-\kappa}{\phi_2}} = q^{\frac{-t_{\text{step}}}{t_{\max} - t_{\min}}} \tag{9}$$

Moreover, in order to have ϕ_1 iterations in the first phase, the initial value of the temperature in the q-HG simulated annealing should be

$$T_{\max} = q^{\kappa} \cdot r^{-\phi_1} = q^{K_{\max} + \frac{t_{\max} - 1}{t_{\max} - t_{\min}}} \tag{10}$$

Below we only mention the parameters of the OT implementation. The cooling schedule for q-HG grammars is matched according to these last two equations.

Finally, we have everything in place to run performance model experiments. For each parameter combination, we report the mean $\pm$ standard deviation of the precision (the relative frequency of returning the grammatical $\{a\}$), as measured on five samples of 10 000 outputs each.

The first experiment (Table 2) compares the OT grammar to q-HG grammars with various q values, using the first candidate set (Fig. 3) and $K_{\max} = 5$. The behaviour of the OT grammar is surprising at first glance: each of the two local optima is returned in half of the cases, and this behaviour does not even depend on the number of iterations. The reason is that even though they are not equally deep valleys, escaping them has always the same probability; thus, they behave symmetrically, as explained in sections 2.3 and 6.4 of Bíró (2006).

Unlike OT grammars, q-HG grammars are traditional, real-valued optimisation problems. Further experiments confirm that their precision converges to 1, as the number of iterations grows infinite (as t_{step} diminishes). Table 2 indicates a less obvious fact: employing the same t_{step} value but diminishing q also increases the

$K_{\max}$	precision
4	0.577 ± 0.004
5	0.599 ± 0.006
6	0.617 ± 0.006
8	0.654 ± 0.006
10	0.671 ± 0.004
20	0.748 ± 0.003
30	0.788 ± 0.004

$K_{\max}$	precision
50	0.835 ± 0.003
100	0.881 ± 0.003
200	0.913 ± 0.003
300	0.930 ± 0.004
500	0.944 ± 0.002
1000	0.962 ± 0.001
2000	0.972 ± 0.002

Table 3: Precision of the OT grammar, while tuning $K_{\max}$ (candidate set 2, $t_{\text{step}} = 0.1$).

q	precision
OT	0.598 ± 0.005
30	0.594 ± 0.004
20	0.599 ± 0.005
10	0.589 ± 0.004
5	0.589 ± 0.001
3	0.594 ± 0.004
2.5	0.608 ± 0.002
2.0	0.630 ± 0.002

q	precision
1.80	0.660 ± 0.008
1.70	0.673 ± 0.004
1.60	0.698 ± 0.004
1.50	0.725 ± 0.002
1.40	0.759 ± 0.005
1.30	0.808 ± 0.005
1.20	0.875 ± 0.002
1.15	0.907 ± 0.003

q	precision
1.10	0.948 ± 0.002
1.08	0.964 ± 0.002
1.06	0.979 ± 0.002
1.05	0.9841 ± 0.0011
1.04	0.9894 ± 0.0012
1.03	0.9922 ± 0.0012
1.02	0.9966 ± 0.0006
1.01	0.9983 ± 0.0004

Table 4: Precision of different types of grammar (candidate set 2, $K_{\max} = 5$, $t_{\text{step}} = 0.1$).

precision. Put differently, a higher q corresponds to a "stricter domination", and so increasing q causes the q-HG grammar to behave more like an OT grammar.

In the next experiment, we switch to the second candidate set (Fig. 4) with the OT grammar. Fixing $t_{\text{step}} = 0.1$, we increase the initial temperature (that is, $K_{\max}$). Table 3 repeats the results obtained (and discussed in length) in section 6.5 of Bíró (2006): the longer the first phase of the cooling schedule, the higher the precision. The explanation is that among the candidates with $k > 1$ extra referents, the arrows on Fig. 4 bring you to a+k, whereas from a+k you end up in the global optimum a. A longer first phase allows the random walker to get farther away from the centre of the search space; subsequently, in the third phase, it has more chance to be trapped by a+k and thus not end up in the performance error form e.

The second candidate set yields a behaviour of SA-OT that is very different from the one observed with the first candidate set. Seemingly, including (infinitely many) candidates that never appear on the scene (for no constraint ranking) is totally superfluous. And yet, they crucially influence the behaviour of the system. The phenomenon has been, therefore, called the *Godot-effect*.

The last experiment concerns the role of q with the second candidate set. Table 4 presents again the case $K_{\max} = 5$ and $t_{\text{step}} = 0.1$. And again, a very small q value corresponds to (almost) faultless performance, whereas q-HG grammars with high q values behave like the OT grammar.

To sum up, we have seen a number of performance models that display a be-

haviour similar to those of the children with certain parameter values, and similar to those of adults (also prone to some errors) with different parameter values—as far as the interpretation of sentence (1a) is concerned.

Interpreting sentence (1b) needs a new constraint. Either we call it PRINCI-PLEA, or we simply refer to the "meaning" of the reflexives, it will disfavour candidates not containing the elephant. Intuition tells that this constraint is ranked high, and therefore there is no local optimum next to the global one. Consequently, the model correctly predicts no performance errors at all.

Finally, in a production task, inputs and outputs are reversed. Candidates are, among others, the forms *him*, *himself*, *them* and *themselves*. A natural way to define the topology is to say that two neighbours differ only in a single feature, such as [number] or [reflexive suffix]. The reference set being given as the input, the highest ranked constraint PROKNOWN in hierarchy (8) is satisfied or violated independently of the choice of the candidates. Yet, the second highest ranked constraint is AGRNUMBER, which correlates with one of the features defining the topology. Consequently, there will be no local optimum if an OT grammar is applied to the candidate set formed by the four words just mentioned, and hence children are again predicted to display adult-like high performance.

6   Discussion and conclusion

OT grammars and q-HG grammars are related, though "philosophically" different grammar architectures. Since the early nineties, such grammars have been constructed for a number of linguistic phenomena, such as for the pronoun resolution task discussed in the present paper. The OT grammar expressed by hierarchy (8), as well as the corresponding q-HG grammars, correctly predict that the pronoun *him* in the sentence (1a) *The elephant is hitting him* must be resolved as referring to the alligator. Hence, hierarchy (8) is a correct model of linguistic competence.

What has received much less attention is the modelling of linguistic performance. In this paper we have shown how simulated annealing can implement both OT grammars and q-HG grammars in a comparable way. An important observation has been that with growing q the behaviour of q-HG grammars converge to the behaviour of OT grammars. Not surprisingly, since growing q also means that q-HG grammars get closer to the "strict domination" ideal of OT.

How do we explain then the performance errors made by children in pronoun resolution? One explanation is provided by the SA-OT *Godot-effect*: considering only the two entities in the context, the elephant and the alligator, necessarily leads to 50% error rate. It might be due to the general development of social cognition skills that the child learns to also include additional entities. Even though including them seems to be waste of resources, the performance nevertheless increases.

A second possible explanation blames computational power. Table 3 suggests that high precision requires a high $K_{\max}$ value (hence, more computing time), which younger children may not afford yet. This approach is supported by the experiments of van Rij et al. (2009): slowing down speech rate has a beneficial effect on pronoun comprehension, but only for children showing a delay of Principle B.

A third explanation is based on Tables 2 and 4. In phonology, quick computations are needed (since many such decisions must be made within a single sentence), and so OT grammars (or q-HG grammars with large q) are used. (Their implementations, namely, run much faster.) Yet, syntactic and pragmatic phenomena can be calculated using slower and more precise algorithms: using q-HG grammars with small q. Maybe maturation means then that the child learns that for certain tasks it is better to give up strict domination and to utilise a smaller q.

Acknowledgement

References

Bíró, Tamás (2005a), How to define Simulated Annealing for Optimality Theory?, *Proc. 10th FG and 9th MoL*, Edinburgh. Also ROA-897[4].

Bíró, Tamás (2005b), When the hothead speaks: Simulated Annealing Optimality Theory for Dutch fast speech, *in* Cremers, C. and et al., editors, *Proc. of the 15th CLIN*, Leiden, pp. 13–28. Also ROA-898.

Bíró, Tamás (2006), *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing*, PhD thesis, University of Groningen. ROA-896.

Biro, Tamás (2007), The benefits of errors: Learning an OT grammar with a structured candidate set, *Proc. Workshop on Cognitive Aspects of Computational Language Acquisition*, ACL, Prague, pp. 81–88.

Chien, Yu-Chin and Kenneth Wexler (1990), Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics, *Language Acquisition* **1**, pp. 225–295.

Chomsky, Noam (1981), *Lectures on Government and Binding*, Foris.

Hendriks, Petra and Jennifer Spenader (2005/2006), When production precedes comprehension: An optimization approach to the acquisition of pronouns, *Language Acquisition* **13**, pp. 319–348.

Jäger, Gerhard and Anette Rosenbach (2006), The winner takes it all – almost: Cumulativity in grammatical variation, *Linguistics* **44**, pp. 937–971.

Jurafsky, Daniel and James H. Martin (2009), *Speech and Language Processing*, 2nd ed., Pearson.

Prince, Alan and Paul Smolensky (1993/2004), *Optimality Theory: Constraint Interaction in Generative Grammar*, Blackwell, Malden, MA, etc.

Reinhart, Tanya (2004), The processing cost of reference set computation: Acquisition of stress shift and focus, *Language Acquisition* **12**, pp. 109–155.

Reuland, Eric (2001), Primitives of binding, *Linguistic Inquiry* **32**, pp. 439–491.

---

[4]ROA: *Rutgers Optimality Archive* at http://roa.rutgers.edu

Smolensky, Paul and Géraldine Legendre, editors (2006), *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar*, MIT Press.

Spenader, Jennifer, Erik-Jan Smits, and Petra Hendriks (2009), Coherent discourse solves the pronoun interpretation problem, *Journal of Child Language* **36**, pp. 23–52.

van Rij, Jacolien, Petra Hendriks, Jennifer Spenader, and Hedderik van Rijn (2009), Modeling the selective effects of slowed-down speech in pronoun comprehension, *Proceedings of GALANA 3*, Cascadilla, Somerville, MA.