October 20, 2005

FG-MoL 2005:
The 10th conference on Formal Grammar
and
The 9th Meeting on Mathematics of Language
Edinburgh
5–7 August 2005

Organizing Committee:
Gerhard Jaeger **Paola Monachesi**
Gerald Penn **James Rogers**
Shuly Wintner

CENTER FOR THE STUDY
OF LANGUAGE
AND INFORMATION

October 20, 2005

Contents

October 20, 2005

1

How to Define Simulated Annealing for Optimality Theory?

Tamás Bíró †

Abstract

Optimality Theory (OT) requires an algorithm optimising the *Harmony function* on the set of candidates. *Simulated annealing*, a well-known heuristic technique for combinatorial optimisation, has been argued to be an empirically adequate solution to this problem. In order to generalise simulated annealing to a non-real valued Harmony function, two representations of a violation profile are proposed: using polynomials and ordinal numbers.

Keywords Optimality Theory, heuristic combinatorial optimization, simulated annealing, ordinal numbers, polynomials

1.1 Optimality Theory and optimisation

A grammar in Optimality Theory (Prince and Smolensky (2004), aka Prince and Smolensky, 1993) consists of two modules, *Gen* and *Eval*. The input—the underlying representation *UR*—is mapped by Gen onto a *set of candidates* *Gen*(*UR*), which reflects language typology. For each language, the language-specific Eval chooses the element (or elements) appearing as the surface form.

Eval is usually perceived as a *pipeline*, in which constraints filter out sub-harmonic candidates. Each constraint assigns violation marks to the candidates in its input, and all candidates with more marks than some other ones are out of the game. Nonetheless, Eval can also be seen

†I wish to acknowledge the support of the University of Groningen's Program for High-Performance Computing; as well as thank the following people for valuable discussions: Gosse Bouma, Gertjan van Noord, Krisztina and Balázs Szendrői.

FG-MoL 2005.
Organizing Committee:, Gerhard Jaeger, Paola Monachesi, Gerald Penn, James Rogers, Shuly Wintner.

as a function assessing the candidates for *Harmony*: the most harmonic one will surface in the language.

A constraint C_i is a function mapping from the candidate set to the set of non-negative integers. The (universal) constraints are ranked into a (language-specific) *hierarchy*: $C_N \gg C_{N-1} \gg ... \gg C_0$. Eval assigns a vector (a *violation profile*, a *Harmony value*) to each candidate w:

$$H(w) = \Big(C_N(w), C_{N-1}(w), ..., C_0(w)\Big) \in \mathbb{N}_0^{N+1} \tag{1.1}$$

Eval also includes an *optimisation process* that finds the optimal candidate(s), and returns it (them) as the surface representation *SR* corresponding to underlying form *UR*:

$$SR(UR) = \mathrm{argopt}_{w \in Gen(UR)} H(w) \tag{1.2}$$

Here, optimisation is with respect to *lexicographic ordering. Lexicographic ordering* of vectors is the way words are sorted in a dictionary: first compare the first element of the vectors, then, if they are the same, compare the second one, and so on. Formally speaking:

Definition 1 $H(w_1)$ is *more optimal* (*more harmonic*) than $H(w_2)$ ($H(w_1) \succ H(w_2)$), or simply candidate w_1 is *better* than w_2 ($w_1 \succ w_2$), if and only if there exists $k \in \{N, N-1, ..., 0\}$ such that

1. $C_k(w_1) < C_k(w_2)$; and
2. for all $j \in \{N, N-1, ..., 0\}$, if $j > k$ then $C_j(w_1) = C_j(w_2)$.

Two violation profiles are *equal* ($H(w_1) = H(w_2)$, two candidates are *equivalent*: $w_1 \simeq w_2$) iff for all $j \in \{N, N-1, ..., 0\}$, $C_j(w_1) = C_j(w_2)$.

We shall call the constraint C_k, which determines the relative ordering of $H(w_1)$ and $H(w_2)$, the *fatal constraint* (the highest ranked constraint with uncancelled marks).

This definition follows from the filtering approach: being worse on a higher ranked constraint cannot be compensated by a better behaviour on lower ranked constraints. This phenomenon is called the *categorical ranking* of the constraints (*Strict Domination Hypothesis*), and is probably a major reason why OT has become so popular.

The following properties result from definition 1 directly (Bíró, forthcoming); by them, the soundness of Eq. (1.2) follows:[1]

Theorem 1 *The set of violation profiles is a well ordered set, namely:*

- Transitivity*: if $w_1 \succ w_2$ and $w_2 \succ w_3$, then $w_1 \succ w_3$ also holds.*

[1] Importantly, the proof of the last statement requires the set of possible violation levels—the range of each constraint—form a well ordered set. This criterion is met in our case, since the violation levels are non-negative integers.

- Law of trichotomy*: for any two candidates w_1 and w_2, exactly one of the following three statements holds:*
  1.$H(w_1) \prec H(w_2)$ *(that is,* $w_1 \prec w_2$*);*
  2.$H(w_1) \succ H(w_2)$ *(that is,* $w_1 \succ w_2$*);*
  3.$H(w_1) = H(w_2)$ *(that is,* $w_1 \simeq w_2$*).*
- The existence of a most optimal subset*: Let S be a set of candidates. Then, S has a unique subset $S_0 \subseteq S$ such that*
  1.if $w_1 \in S_0$ *and* $w_2 \in S_0$*, then* $H(w_1) = H(w_2)$*;*
  2.if $w_1 \in S_0$ *and* $w_3 \in S \setminus S_0$*, then* $w_1 \succ w_3$*.*

Optimality Theory poses the following computational challenge: what algorithm realises the optimisation required by Eval? Eisner (2000) demonstrates that finding the optimal candidate is OptP-complete. In addition, numerous linguistic models use an infinite candidate set. Several solutions have been proposed, although each of them is built on certain presuppositions, and they require large computational resources. Finite state techniques (e.g. Ellison (1994), Frank and Satta (1998), Karttunen (1998), Gerdemann and van Noord (2000), Bíró (2003)) not only require Gen and the constraints to be finite state, but work only with some further restrictions. *Chart parsing* (*dynamic programming*, e.g. Tesar and Smolensky (2000), Kuhn (2000)) has assumptions met by most linguistic models, but also requires a relatively large memory. Similar applies to genetic algorithms (Turkel, 1994).

A cognitively adequate optimisation algorithm, however, does not have to be exact. Speech is full or errors, and (a part of) the performance errors could be the result of the optimisation process returning erroneous outputs. Yet, a cognitively adequate algorithm should always return *some* response within constant time, since the conversation partners are not computer users used to watch the sandglass.

This train of thought leads to heuristic optimisation techniques, defined by Reeves (1995) as "*a technique which seeks good (i.e. near-optimal) solutions at a reasonable computational cost without being able to guarantee either feasibility or optimality, or even in many cases to state how close to optimality a particular feasible solution is.*" In this paper, we implement Optimality Theory by using the simplest heuristic optimisation technique, *simulated annealing*, and introduce the *Simulated Annealing for Optimality Theory algorithm* (SA-OT).

The SA-OT algorithm will, under normal conditions, find the "correct", i.e. the grammatical output—the optimal element of the candidate set—with high probability, within constant time, using only a very restricted memory. Human speakers sometimes speed up the computational algorithm, and the price is paid in precision: we propose to see

(some) fast speech phenomena as decreased precision of Eval due to the increased speed. Similarly, by speeding up SA-OT, the chance of finding suboptimal, yet "good (i.e. near-optimal) solutions" increases. The models of fast speech phenomena thus constructed support the cognitive adequateness of SA-OT (Bíró, 2005, forthcoming).

1.2 Heuristic Optimisation with Simulated Annealing

Simulated annealing, also referred to as *Boltzmann Machines* or as *stochastic gradient ascent*, is a wide-spread stochastic technique for combinatorial optimisation (e.g. Reeves (1995)). Only few have applied simulated annealing in linguistics, most of them for parsing (e.g. Selman and Hirst (1985), Howells (1988), Kempen and Vosse (1989), Selman and Hirst (1994)). It may also be found in the pre-history of Optimality Theory (Smolensky, 1986) and in later work on *Harmonic grammar*—including *Maximum Entropy* models of Optimality Theory (Jäger, 2003)—, though usually related to grammar learning. To our best knowledge, it has never been applied within the standard OT paradigm, especially for finding the optimal candidate.

Simulated Annealing searches for the state of a system minimising the cost function E (Energy or Evaluation) by performing a random walk in the search space. If the rule were to move always downhill (*gradient descent*), then the system would very easily be stuck in local minima. Therefore, we also allow moving upwards with some chance, which is higher in the beginning of the simulation, and which then diminishes. The control parameter T determining the likelihood of uphill moves is called "temperature", because the idea proposed independently by Kirkpatrick et al. (1983) and by Černy (1985) originates in statistical physics (Metropolis et al., 1953).

The random walk is launched from an initial state w_0. At each time step, a random neighbour state (w') of the actual state w is picked. We need, thus, to have a *topology* on the search space that defines the neighbours of a state (the *neighbourhood structure*), as well as the *a priori* probability distribution determining the choice of a neighbour in each step. Subsequently, we compare w' to w, and the random walker moves from w to w' with probability $P(w \to w' \mid T)$, where T is the "temperature" at that moment of the simulation. A random number r is generated between 0 and 1, and if $r < P(w \to w' \mid T)$, the random walker moves. If $E(w)$ is the cost function to minimise, then:

$$P(w \to w' \mid T) = \begin{cases} 1 & \text{if } E(w') \leq E(w) \\ e^{-\frac{E(w')-E(w)}{T}} & \text{if } E(w') > E(w) \end{cases} \tag{1.3}$$

Moving downhill is always possible, and moving uphill depends on the difference in E and on the temperature T. At the beginning of the simulation, T is assigned a high value, making any move very likely. The value of T is then decreased gradually, while even the smallest jump does not become highly improbable. When the temperature has reached its lowest value, the algorithm returns the state—a local minimum—into which the random walker is "frozen". Obviously, nothing guarantees finding the global minimum, but the slower the *cooling schedule* (the more iterations performed), the higher the probability to find it.

1.3 Simulated Annealing for OT: the basic idea

How to combine simulated annealing with Optimality Theory? The search space is the candidate set, as defined by standard OT. Yet, a *neighbourhood structure* (a *topology*) should be added—an unknown concept in standard OT literature—in order to determine how to pick the next candidate. We propose to consider two candidates as neighbours if they differ only minimally: if a *basic operation* transforms one into the other. What a basic operation is depends on the problem, but should be a naturally fitting choice. It is the neighbourhood structure that determines which candidates are *local* optima, which may be returned as erroneous outputs. Thus, the definition of the topology is crucial to account for speech errors.

If the topology determines the horizontal structure of the landscape in which the random walker roves, the Harmony function adds its vertical structure. Here again, standard Optimality Theory provides only the first part of the story. The transition probability $P(w \to w' \mid T) = 1$ if w' is better than w (*i.e.*, $H(w') \succ H(w)$). But how to define the transition probability to a worse candidate, in function of the temperature T? How to adopt Eq. (1.3)? What is $H(w') - H(w)$, let alone its exponent? And what should temperature look like?

Equation (1.3) provides the meaning of temperature: T defines the range of $E(w') - E(w)$ above which no uphill jump is practically possible ($P(w \to w' \mid T) \approx 0$, if $E(w') - E(w) \gg T$), and below which uphill moves are allowed ($P(w \to w' \mid T) \approx 1$, if $E(w') - E(w) \ll T$). In turn, we *first* have to define the difference $H(w') - H(w)$ of two violation profiles, *then* introduce temperature for OT in an analogous way. Last, we can adjust Eq. (1.3) and formulate the SA-OT algorithm.

Two approaches—two representations of the violation profile—are proposed in order to carry out this agenda. Both may have its adherents and its opponents. And yet, both approaches lead to the same algorithm.

1.4 Violation profiles as polynomials

As mentioned, a crucial feature of Optimality Theory is *strict domination*: a candidate suboptimal for a higher ranked constraint can never win, even if it satisfies the lower ranked constraints best. Prince and Smolensky (2004) present why the Harmony function $H(w)$ satisfying strict domination cannot be realised with a real-valued function.

Suppose first that an upper bound $q > 0$ exists on the number of violation marks a constraint can assign to a candidate. The possible levels of violation are $0, 1, ..., q-1$. Then, the following real-valued Energy function $E(w)$ *realises* the Harmony $H(w)$ known from (1.1):

$$E(w) = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + ... + C_1(w) \cdot q + C_0(w) \quad (1.4)$$

$E(w)$ *realising* $H(w)$ means that for all w_1 and w_2, $E(w_1) \leq E(w_2)$ if and only if $H(w_1) \succeq H(w_2)$. In other words, optimising the Harmony function is equivalent to minimising the Energy function. Observe that $E(w)$ with a lower q does not necessarily realise $H(w)$.

However, nothing in general guarantees that such an upper bound exists: Eq. (1.4) with a given q is only an approximation. Then, let us represent the violation profiles as polynomials of $q \in \mathbb{R}^+$:

$$E(w)[q] = C_N(w) \cdot q^N + C_{N-1}(w) \cdot q^{N-1} + ... + C_1(w) \cdot q + C_0(w) \quad (1.5)$$

and consider the behaviour of $E(w)[q]$ as q goes to infinity! Yet, $E(w)[q]$ also goes to infinity as q grows boundless: $\lim E(w)[q] = +\infty$.

The trick is to perform an operation first, or to check the behaviour of the energy function first, and only subsequently bring q to the infinity. By performing *continuous* operations, it makes sense to change the order of the operation and of the limit to infinity.

First, let us compare two violation profiles seen as polynomials. The following definition—comparing the limits—is meaningless: $w_1 \succ w_2$ iff $\lim_{q\to\infty} E(w_1)[q] < \lim_{q\to\infty} E(w_2)[q]$. We can, however, consider the limit of the comparison, instead of the comparison of the limits:

Definition 2 $E(w_1) \prec E(w_2)$ if and only if
either $\lim_{q\to+\infty} (E(w_2)[q] - E(w_1)[q]) > 0$,
or $\lim_{q\to+\infty} (E(w_2)[q] - E(w_1)[q]) = +\infty$.
Furthermore, $E(w_1) = E(w_2)$ iff $E(w_1)[q] = E(w_2)[q]$ for all $q \in \mathbb{R}^+$.

Energy-polynomials with this definition of $\prec$ realise the Harmony function: $E(w_1) \preceq E(w_2)$ if and only if $H(w_1) \succeq H(w_2)$. For a proof, see the Appendix and Bíró (forthcoming). Consequently, the *polynomial representation* of the Harmony function is well-founded.

Can we now use energy polynomials to apply simulated annealing to Optimality Theory? As explained, the role of temperature in simulated annealing is to define a magnitude above which counter-optimal transitions are improbable, and below which they are very probable. Thus, temperature must have the same type (dimension, form) as the function to optimise. If, in our case, the energy function takes different polynomials as values, then T should be also polynomial-like:

$$T[q] = \langle K_T, t\rangle\,[q] = t \cdot q^{K_T} \tag{1.6}$$

Temperature $T = \langle K_T, t\rangle$ looks as if it were a violation profile that has incurred t marks from a constraint—supposing that some constraint has K_T as index. But temperature can be more general: we only require $t \in \mathbb{R}^+$, whereas K_T may take any real number as value.

The last step is to define the transition probability of moving from candidate w to a neighbour w'. If $w' \succeq w$, the probability is 1. Otherwise, we repeat the trick: *first* perform the operations proposed by (1.3), and only *afterwards* take the $q \to +\infty$ limit:

$$P\Big(w \to w' \;\Big|\; T[q]\Big) = \lim_{q\to+\infty} e^{-\frac{E(w')[q]-E(w)[q]}{T[q]}} \tag{1.7}$$

Observe that if C_k is the fatal constraint when comparing w and w', then the dominant summand in the expression $E(w')[q] - E(w)[q]$ is $\big[C_k(w') - C_k(w)\big]q^k$. Thus, (1.7) and (1.6) yield the following

Rules of moving from w to w' at temperature $T = \langle K_T, t\rangle$:

- If w' is better than w: move! $P(w \to w'|T) = 1$
- If w' loses due to fatal constraint C_k:
  - If $k > K_T$: don't move! $P(w \to w'|T) = 0$
  - If $k < K_T$: move! $P(w \to w'|T) = 1$
  - If $k = K_T$: move with probability $P = e^{-(C_k(w')-C_k(w))/t}$.

Note that the last expression requires $t > 0$, as in thermodynamics. Gradually dropping T can be done by diminishing K_T in a loop with an embedded loop that reduces t. Thus, the height of the allowed counter-optimal jumps also diminish—similarly to usual simulated annealing.

1.5 Violation profiles as ordinal numbers

Instead of considering the limit $q \to +\infty$ of real-valued weights in polynomials, why not take *infinite weights*? In set theory, the well ordered set $\{0, 1, 2, ..., q-1\}$ defines the integer q. When the possible levels of violation formed this set, we could use weight q. In the general case, the possible levels of violation of the constraints form the set $\{0, 1, 2, ...\}$: this well ordered set is called ω, the first limit ordinal (Suppes, 1972).

Arithmetic can be defined on ordinal numbers, including comparison, addition and multiplication. These latter operations are associative, but not commutative. Therefore, we can introduce a new representation of the Harmony function $H(w)$:

$$E(w) = \omega^N C_N(w) + ... + \omega C_1(w) + C_0(w) = \sum_{i=N}^{0} \omega^i C_i(w) \qquad (1.8)$$

Because ω is the upper limit of the natural numbers, $\omega^i n < \omega^{i+1}$ for any finite n. Thus, the definition of $E(w)$ in (1.8) with the usual relation $<$ from ordinal arithmetic also *realises* the Harmony function: $E(w_1) \leq E(w_2)$ if and only if $H(w_1) \succeq H(w_2)$. The very definition of limit ordinals excludes ganging up effects.

We need now the difference of two E values as defined in (1.8). Instead of subtraction, we introduce an operation $\Delta(a, b)$ on the ordinal numbers of form $\sum_{i=N}^{0} \omega^i a_i$, such that $a = b + \Delta(a, b)$:

Definition 3 If $a = \sum_{i=N}^{0} \omega^i a_i$ and $b = \sum_{i=N}^{0} \omega^i b_i$ and $a > b$, let be $\Delta(a, b) = \sum_{i=N}^{0} \omega^i \delta_i$, where $\delta_i = \begin{cases} a_i - b_i & \text{if } \forall j.(i < j \leq N)\text{: } a_j = b_j \\ a_i & \text{otherwise} \end{cases}$

For candidates w and w', the co-efficient of the highest non-zero term in $\Delta\big(E(w'), E(w)\big)$ is the difference of the violation levels of the fatal constraint. The lower summands vanish compared to the highest term, so we can neglect them in a new definition:

Definition 4 If $a = \sum_{i=N}^{0} \omega^i a_i$ and $b = \sum_{i=N}^{0} \omega^i b_i$, and $a > b$, let be $\Delta'(a, b) = \sum_{i=N}^{0} \omega^i \delta'_i$, where $\delta'_i = \begin{cases} a_i - b_i & \text{if } \forall j.(i < j \leq N) : a_j = b_j \\ 0 & \text{otherwise} \end{cases}$

Observe that for candidates w and w', if C_k is the fatal constraint why $w \succ w'$, then $\Delta'\big(E(w'), E(w)\big) = \omega^k \big[C_k(w') - C_k(w)\big]$.

Next, we introduce the following conventions, where a, b, i and j are positive integers, and x, y and z are ordinal numbers (remember that ω means "infinity"):

$$e^{-\frac{\omega^i a}{\omega^j b}} := e^{-\omega^{i-j}\frac{a}{b}} := \begin{cases} 1 & \text{if } i < j \\ e^{-\frac{a}{b}} & \text{if } i = j \\ 0 & \text{if } i > j \end{cases} \qquad (1.9)$$

Temperature has to have the same form as the difference of two violation profiles $\Delta'(E(w'), E(w))$, so we propose

$$T = \langle K_T, t \rangle = \omega^{K_T} t \qquad (1.10)$$

```
ALGORITHM: Simulated Annealing for Optimality Theory (SA-OT)
Paramters: 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
           choose random w' in Neighbourhood(w)
           w := w' with probability P( w --> w' | T=<K,t> )
                  as defined in the ''Rules of moving''
        end-for
   end-for
return w
```

FIGURE 1 The algorithm of *Simulated Annealing Optimality Theory*.

Now, all tools are ready to define probability $P(w \to w' \mid T)$, closely following Eq. (1.3). If $E(w) \geq E(w')$ then $P(w \to w' \mid T) = 1$, else

$$P(w \to w' \mid T) = e^{-\frac{\Delta'(E(w'),E(w))}{T}} \tag{1.11}$$

Some readers may prefer the way leading to Eq. (1.7), while others the one to (1.11). Yet, the interpretation of both of them yields the same *Rules of moving*, those in section 1.4. Both trains of thought introduce temperature as a pair $T = \langle K_T, t \rangle$. Diminishing it requires a double loop: the inner one reduces t, and the outer one K_T.

1.6 Conclusion: SA-OT

The pseudo-code of the *Optimality Theory Simulated Annealing* algorithm (OT-SA) can be finally presented (Figure 1).

Out of the parameters of the algorithm, K_{max} is usually higher than the index of the highest ranked constraint, in order to introduce an initial phase when the random walker may rove unhindered in the search space. Similarly, K_{min} defines the length of the final phase of the simulation, giving enough time to "relax", to reach the closest local optimum. Otherwise, SA-OT would return any candidate, not only local optima, resulting in an uninteresting model. Typically, $K_{step} = 1$.

Parameters t_{max}, t_{min} and t_{step} drive t in the inner loop, affecting only the exponential in the last case ($k = K_T$) of the *Rules of moving*. As w and w' differ only minimally, in a *basic operation*, their violation profiles are also similar: $|C_k(w') - C_k(w)| \leq 2$ usually, motivating $t_{max} = 3$ and $t_{min} = 0$. Parameter t_{step} is the most interesting one, and can vary along more orders of magnitude: by being inversely proportional to the number of iterations performed, it directly controls the speed of the simulation, that is, its precision.

In practice, the algorithm is surprisingly successful in modelling, besides other, fast speech phenomena in Dutch metrical stress assignment (Bíró, 2005). In SA-OT, the frequency of the different forms can be fine-tuned by varying the parameters (especially t_{step}). It can also predict different frequencies for the same phenomenon in different inputs (Bíró, forthcoming), if the search space has a different structure: indeed, the *topology* of the search space is an important novel concept in SA-OT.

To sum up, *Simulated Annealing for Optimality Theory* (SA-OT) is a promising algorithm to find the optimal element of the candidate set. In the present paper, we have argued that it is both cognitively plausible and mathematically well-founded, whereas further work has shown that it can account for real phenomena.

1.7 Appendix: Energy-polynomials realise $H(w)$

Here, we sketch how to prove that energy-polynomials—with Definition 2 of $\prec$ in section 1.4—realise the Harmony function: $E(w_1) \preceq E(w_2)$ if and only if $H(w_1) \succeq H(w_2)$. For a more detailed proof, see Bíró (forthcoming). First, we have to demonstrate:

Theorem 2 Law of trichotomy for energy polynomials: *for any w_1 and $w_2 \in$ GEN(UR), exactly one of the following statements holds: either $E(w_1) \prec E(w_2)$, or $E(w_1) \succ E(w_2)$, or $E(w_1) = E(w_2)$.*

For a proof, note that the polynomial $P[q] = E(w_1)[q] - E(w_2)[q]$ may have maximally N roots, the greatest of which be q_N. Unless $E(w_1)[q] = E(w_2)[q]$ for all q's, $P[q]$ is either constantly positive or constantly negative for $q > q_N$. Subsequently, we need:

Lemma 3 *If $H(w_1) \succ H(w_2)$, then $E(w_1) \prec E(w_2)$.*

Proof. Let C_k be the fatal constraint due to which $H(w_1) \succ H(w_2)$ (Def. 1). If $k = 0$ then $E(w_2)[q] - E(w_1)[q] = C_0(w_2) - C_0(w_1) > 0$ for all q. By definition, then, $E(w_1) \prec E(w_2)$.

If, however, $k > 0$, then let c be such that $c > C_i(w_1)$ and $c > C_i(w_2)$ for all $i < k$. Further, let $q_0 = \max(\frac{2c}{C_k(w_2) - C_k(w_1)}, 2)$. For all $q > q_0$:

$$E(w_2)[q] - E(w_1)[q] = \sum_{i=0}^{N} [C_i(w_2) - C_i(w_1)]q^i =$$
$$= [C_k(w_2) - C_k(w_1)]q^k + \sum_{i=0}^{k-1} [C_i(w_2) - C_i(w_1)]q^i \qquad (1.12)$$

because C_k is the fatal constraint. As $q > q_0 \geq \frac{2c}{C_k(w_2) - C_k(w_1)}$, in the first summand we use $C_k(w_2) - C_k(w_1) > 2c/q$. For the second

component, we employ the fact that $C_i(w_2) - C_i(w_1) > -c$ for all $i < k$, as well as the sum of a geometrical series. Consequently,

$$E(w_2)[q] - E(w_1)[q] > \frac{2c}{q}q^k - c\frac{q^k - 1}{q - 1} = c\frac{q^k - 2q^{k-1} + 1}{q - 1} > 0 \quad (1.13)$$

because $q > q_0 \geq 2$. In sum, either $k = 0$ or $k > 0$, we have shown that there exists a q_0 such that for all $q > q_0$: $E(w_2)[q] - E(w_1)[q] > 0$. Because this difference is a polynomial, we obtain one of the two cases required by Definition 2 of $E(w_1) \prec E(w_2)$. □

Finally, the following four statements can be simply demonstrated by using the definitions, the previous lemma and the laws of trichotomy:

Theorem 4 *Energy-polynomials realise the Harmony function:*

- $E(w_1) = E(w_2)$*, if and only if* $H(w_1) = H(w_2)$*;*
- $E(w_1) \prec E(w_2)$*, if and only if* $H(w_1) \succ H(w_2)$*.*

References

Bíró, Tamás. 2003. Quadratic alignment constraints and Finite State Optimality Theory. In *Proc. FSMNLP within EACL 2003*, pages 119–126. Budapest; also ROA-600[2].

Bíró, Tamás. 2005. When the hothead speaks: Simulated Annealing Optimality Theory for Dutch fast speech. In C. Cremers, H. Reckman, M. Poss, and T. van der Wouden, eds., *Proc. 15th Meeting of Computational Linguistics in the Netherlands (CLIN 2004)*. Leiden.

Bíró, Tamás. forthcoming. *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing*. Ph.D. thesis, Univ. of Groningen.

Černy, V. 1985. Thermodynamical approach to the Travelling Salesman Problem: an efficient simulation algorithm. *Journal of Optimization Theory and Applications* 45:41–55.

Eisner, Jason. 2000. Easy and hard constraint ranking in Optimality Theory: Algorithms and complexity. In J. E. et al., ed., *Finite-State Phonology: Proc. SIGPHON-5*, pages 57–67. Luxembourg.

Ellison, T. Mark. 1994. Phonological derivation in Optimality Theory. In *COLING-94, Kyoto*, pages 1007–1013. also: ROA-75.

Frank, Robert and Giorgio Satta. 1998. Optimality Theory and the generative complexity of constraint violability. *Comp. Ling.* 24(2):307–315.

[2]ROA stands for Rutgers Optimality Archive at http://roa.rutgers.edu

Gerdemann, Dale and Gertjan van Noord. 2000. Approximation and exactness in Finite State Optimality Theory. In *Jason Eisner, Lauri Karttunen, Alain Thriault (eds): SIGPHON 2000, Finite State Phonology*.

Howells, T. 1988. VITAL: a connectionist parser. In *Proc. 10th Annual Meeting of the Cognitive Science Society*, pages 18–25. Lawrence Erlbaum.

Jäger, Gerhard. 2003. Maximum entropy models and Stochastic Optimality Theory. m.s., ROA-625.

Karttunen, Lauri. 1998. The proper treatment of Optimality Theory in computational phonology. In *Proc. FSMNLP*, pages 1–12. Ankara.

Kempen, Gerard and Theo Vosse. 1989. Incremental syntactic tree formation in human sentence processing: a cognitive architecture based on activation decay and simulated annealing. *Connection Science* 1:273–290.

Kirkpatrick, S., C. D. Gelatt Jr., and M. P. Vecchi. 1983. Optimization by Simulated Annealing. *Science* 220(4598):671–680.

Kuhn, Jonas. 2000. Processing Optimality-theoretic syntax by interleaved chart parsing and generation. In *Proc. ACL-38*, pages 360–367. Hongkong.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. Equation of state calculation by fast computing machines. *Journal of Chemical Physics* 21(6):1087–1092.

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

Reeves, Colin R., ed. 1995. *Modern Heuristic Techniques for Combinatorial Problems*. London, etc.: McGraw-Hill.

Selman, Bart and Graeme Hirst. 1985. A rule-based connectionist parsing system. In *Proc. of the Seventh Annual Meeting of the Cognitive Science Society, Irvine*, pages 212–221. Hillsdale, NJ: Lawrence Erlbaum.

Selman, Bart and Graeme Hirst. 1994. Parsing as an energy minimization problem. In G. Adriaens and U. Hahn, eds., *Parallel Natural Language Processing*, pages 238–254. Norwood, NJ: Ablex Publishing.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of Harmony Theory. In *Rumelhart et al.: Parallel Distributed Processing*, vol. 1, pages 194–281. Cambridge, MA–London: MIT Press.

Suppes, Patrick. 1972. *Axiomatic Set Theory*. New York: Dover.

Tesar, Bruce and Paul Smolensky. 2000. *Learnability in Optimality Theory*. Cambridge, MA - London, England: The MIT Press.

Turkel, William. 1994. The acquisition of Optimality Theoretic systems. m.s., ROA-11.