Finite State Transducers and Optimality Theory[1]

Giorgio Satta
University of Padua
Italy

[1]Work done in collaboration with Robert Frank, Johns Hopkins University, Baltimore, MD, and University of Delaware, Newark, DE

Introduction

In generative linguistic there has been a shift from systems of structural rewriting rules and derivation processes to sets of constraints and optimization mechanisms

Constraints state, in an abstract way, well-formedness conditions on linguistic representations

Optimization is applied in selecting the linguistic analysis with the "lowest cost"

As a consequence of optimization, it turns out that in certain contexts certain constraints can be violated, that is constraints are not absolutely inviolable

Outline

1. Optimality Theory and constraint violation
2. finite state transducers and Optimality Theory

Optimality Theory

(Prince and Smolensky 1993)

Optimality Theory (OT) is a paradigm that has been applied in much of the recent work on phonological theory and in some work in syntactic theory

The basic components of an OT system are

1. a function GEN
2. a set of ranked constraints

Two constraints are in conflict when no form can simultaneously satisfy the two constraints

Conflicts are solved by allowing lower ranked constraints to be violated in order for higher ranked constraints to be satisfied (optimization)

GEN

An **underlying representation** (UR) is an abstract representation of the linguistic content of some linguistic form

A **surface realization** (SR) associated with some UR is a potential linguistic realization of the UR, possibly enriched with structural information

The function GEN maps each UR to the (possibly infinite) set of all possible SRs of the UR, called the **candidate set**

The core of an OT system is devoted to choosing among the members of a candidate set to determine which SR corresponds to the actual linguistic form(s)

Constraint Evaluation

Each constraint can be evaluated on a SR; the result of the evaluation represents the number of violation of the constraint in that SR

Optimization Criterion:
A SR γ_1 is preferred over a SR γ_2 if for some i we have both

(i) each of the first $i - 1$ higher ranked constraints enforces the same number of violations on γ_1 and γ_2; and

(ii) the i-th constraint enforces on γ_1 a number of violations strictly smaller than the number of violations it enforces on γ_2

The set of optimal SRs in a candidate set, w.r.t. to the above criterion, is the output of the OT system

OT (cont'd)

The strong assumption is made that the constraints which are used to evaluate the members of the candidate set are a part of universal grammar

On this view, what varies from one language to another is the relative ranking of constraints: every ordering of the constraints corresponds to a possible phonological or syntactic system

Syllable Structure
(simplified)

A **syllable** consists of

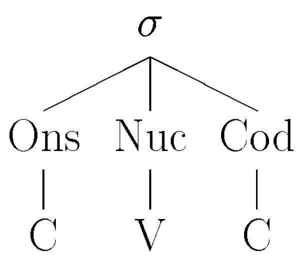
(i) an obligatory element called **nucleus**

(ii) an optional leftmost daughter called **onset**

(iii) an optional rightmost daughter called **coda**

Consonants (C) are permitted to surface only as the daughters of the onset or coda; vowels (V) may only be the daughter of the nucleus

Maximal syllabic structure is then



and is schematically represented as .CVC.

Syllable (cont'd)

Any syllabic positions may be present without being filled by elements from the UR, giving raise to an **epenthetic** segment in the SR

For instance, a SR of an UR /V/ is

$$.V\Box.$$

where a syllable is realized with an underlying nucleus and an epenthetic coda

Any C or V from the UR may be ignored for the purposes of syllabic structure, giving raise to an **unparsed** element in a SR.

For instance, a SR of an UR /CCV/ is

$$\langle C \rangle.CV.$$

where an unparsed element has been left out from the realized syllable

Optimal Syllables

Two classes of constraints govern the choice of the optimal SR of syllabic structure (Prince and Smolensky 1993)

The first class characterizes what kinds of syllabic structures are universally unmarked:

(i) ONS: Syllables must have onsets

(ii) –COD: Syllables must not have a coda

The second class of constraints, called *faithfulness constraints*, restrict the relation between the UR and SR:

(iii) PARSE: Underlying segments must be parsed into syllable structure

(iv) FILL: Syllable positions must be filled with underlying segments

Example

Assume /V/ as UR (candidate set partially specified)

Candidate	Constraints Violated	Constraints Satisfied
.V.	ONS	PARSE, FILL, –CODA
⟨V⟩	PARSE	FILL, ONS, –CODA
.□V.	FILL	PARSE, ONS, –CODA
.□V□.	FILL(×2), –CODA	PARSE, ONS
.V□.	FILL, ONS, –CODA	PARSE

Example (cont'd)

By varying ranking among the constraints we obtain different optimal SR for UR /V/ ($\gg$represents the ranking relation, $*$ represents a constraint violation)

Fill$\gg$Parse$\gg$Ons:

/V/	Fill	Parse	Ons
$\Rightarrow$ .V.			*
$\langle$V$\rangle$		*	
.□ V.	*		
.□V□.	* *		
.V□.	*		*

Ons$\gg$Parse$\gg$Fill:

/V/	Ons	Parse	Fill
.V.	*		
$\langle$V$\rangle$		*	
$\Rightarrow$.□V.			*
.□V□.			**
.V□.	*		*

Example

Consider the UR /CVC/

/CVC/		Fill	Parse	Ons	–Cod
⇒	.CVC.				*
	.CV⟨C⟩.		*		
	.CV.C□.	*			

/CVC/		–Cod	Parse	Ons	Fill
	.CVC.	*			
	.CV⟨C⟩.		*		
⇒	.CV.C□.				*

Selection of a single syllable or a two syllables form depends upon the constraint ranking. Both forms are apparently attested among the world's languages

Result

The generative complexity of an OT system depends upon the character of GEN and the constraints

We assume that

1. GEN is a rational transduction
2. constraints can be expressed through regular expressions

Under the above hypotheses, we show that an OT system always induces a rational transduction

The above assumptions considerably limits the expressiveness of OT. Nonetheless, this formalization still captures constraint violability which is the core aspect of OT

Model of OT

Definition An **optimality system** (OS) is a triple

$$G = (\Sigma, \textsc{gen}, C)$$

where

(i) Σ is a finite alphabet

(ii) $\textsc{gen}$ is a rational transduction over Σ

(iii) $C = \langle c_1, \ldots, c_p \rangle$, $p \geq 1$, is an ordered sequence of regular expressions, where $L(c_i)$ denotes the set of strings that satisfy constraint c_i

Model of OT (cont'd)

Let $\mathrm{OT}^0(w) = \textsc{gen}(w)$

For $1 \leq i \leq p$ let

$$\mathrm{OT}^i(w) \;=\; \begin{cases} \mathrm{OT}^{i-1}(w) & \text{if } \mathrm{OT}^{i-1}(w) \cap L(c_i) = \emptyset; \\ \mathrm{OT}^{i-1}(w) \cap L(c_i) & \text{if } \mathrm{OT}^{i-1}(w) \cap L(c_i) \neq \emptyset. \end{cases}$$

Observe that $\mathrm{OT}^i(w)$ performs a "conditional intersection", depending on whether $\mathrm{OT}^{i-1}(w) \cap L(c_i)$ is empty or not

OT^p is the **optimality mapping** associated with G

Preliminaries

Let R be a rational relation

The **image** of string w under R is the set

$$wR = \{u \mid (w,u) \in R\}$$

and is a regular language

The **left projection** of R is the language

$$\mathrm{Left}(R) = \{u \mid (u,w) \in R\}$$

The **right projection** of R is the language

$$\mathrm{Right}(R) = \{w \mid (u,w) \in R\}$$

Left(R) and Right(R) are both regular languages

The **inverse relation**

$$R^{-1} = \{(w,u) \mid (u,w) \in R\}$$

is a rational relation

Left/Right Restrictions

Let R be a rational relation and L be a regular language

The **left restriction** of R to L is the relation

$$\mathrm{Lrst}(R, L) = \{(u, w) \mid (u, w) \in R,\ u \in L\}$$

The **right restriction** of R to L is the relation

$$\mathrm{Rrst}(R, L) = \{(u, w) \mid (u, w) \in R,\ w \in L\}$$

$\mathrm{Lrst}(R, L)$ and $\mathrm{Rrst}(R, L)$ are both rational relations

Proof outline:

- left restriction: simultaneously simulate FST for R and FSA for L
- right restriction: use closure under relation inversion

OT^p

The optimality mapping OT^p is a rational relation

Proof outline:

- Assume OT^{i-1} is a rational relation R and show OT^i is rational as well (induction)
- For input w we have to compute:
  1. $wR \cap L(c_i)$
     if some form in wR satisfies c_i, that is $wR \cap L(c_i) \neq \emptyset$
  2. wR
     if no form in wR satisfies c_i, that is $wR \cap L(c_i) = \emptyset$
- Guess case 1 or 2, compute appropriate translation and verify the guess is correct

Verify guess

Language

$$L_v = \mathrm{Left}(\mathrm{Rrst}(\boldsymbol{R}, L(\boldsymbol{c}_i)))$$

is a regular language

We have

$w \in L_v$
if and only if $w\mathrm{Rrst}(\boldsymbol{R}, L(\boldsymbol{c}_i)) \neq \emptyset$
if and only if $w\boldsymbol{R} \cap L(\boldsymbol{c}_i) \neq \emptyset$

Then our guess of 1 or 2 can be verified by checking whether $w \in L_v$

Discussion

OT^p cannot be a rational function, since in general there is more than one optimal SR for a given UR

Constraint c is a **regular gradient** constraint if, for any $i \geq 0$, the set of strings in Σ^* with i violations of c is a regular language

It is an open problem whether an OS with regular gradient constraints induces a rational relation