The sotaq program
and secondary stress in two varieties of Portuguese

Maria Bernadete Abaurre <abaurre@iel.unicamp.br>
Charlotte Galves <galvesc@obelix.unicamp.br>
Filomena Sandalo <sandalo@iel.unicamp.br>
Department of Linguistics
IEL - UNICAMP

Arnaldo Mandel <am@ime.usp.br>
Department of Computer Science
IME
Universidade de São Paulo

© AGMS, 2001
Outline

1: The problem of secondary stresses
2: Different approaches
3: Optimality Theory
4: OT for secondary stress
5: sotaq
6: Final observations
Secondary stresses — An example

- O organizador apresentou a catalogadora.
- **EP**: O *organizador* apresentou a *catalogadora*.
- **BP**: O *organizador* apresentou a *catalogadora*.
Approaches

Derivational Analysis

- Given an input sentence, apply some rules sequentially, until a (unique) final form surfaces.
Approaches

Derivational Analysis

- Given an input sentence, apply some rules sequentially, until a (unique) final form surfaces.

- Wrong and insufficient results are dealt with by ad-hoc changes for each special case.
Approaches

Optimality Theory

Given an input sentence, consider an associated collection of candidate final forms. A family of constraints yields function to be optimized on this set, so that the optimal solutions are the required final forms.
Optimization can be viewed as competition between the constraints.
- Optimization can be viewed as competition between the constraints.
- Wrong and insufficient results are dealt with by ad-hoc changes to the model parameters: constraints and ranking.
An Overview of OT

- started by Prince & Smolensky, 1993.
- General framework for linguistic theories.
Setting for OT models

Problem

- Given:
  - A set $I$ of inputs, and
  - For each $p \in I$, a nonempty set $Gen(p)$, of structures.

- Find: For each $p \in I$, its outputs in $Gen(p)$.

Assumption:

A preference (i.e. reflexive and transitive) relation is given $\preceq$ on each $Gen(p)$, and the outputs are the minimal members of $Gen(p)$.
OT models

Notation: $S = \bigcup_{p \in I} Gen(p)$.

Ingredients:

- **Constraints**: $C_1, \ldots, C_k : S \rightarrow \mathbb{N}$.
  Define also $C = (C_1, \ldots, C_k) : S \rightarrow \mathbb{N}^k$.

- **A preference** relation in $\mathbb{N}^k$ satisfying: $x \preceq y$ if $x_1 \leq y_1, \ldots, x_k \leq y_k$

Mix:

- A derived induced preference on each $Gen(p)$

$$s \preceq t \iff C(s) \preceq C(t)$$

$C(p)$ is the **violation vector** of structure $p$

(AKA tableau line)
Rank Optimality

- **Stratification**: a surjective map

\[ \varphi : \{1, \ldots, k\} \rightarrow \{1, \ldots, r\} \]

- **Recount**: \( \bar{\varphi} : \mathbb{N}^k \rightarrow \mathbb{N}^r \) given by

\[ \bar{\varphi}(x)_j = \sum_{i: \varphi(i) = j} x_i \]

- **Reordering**: for \( x, y \in \mathbb{N}^k \), \( x \preceq y \) if \( \bar{\varphi}(x) \) precedes \( \bar{\varphi}(y) \) lexicographically.
The **lexicographic order** on $\mathbb{N}^k$ is defined as follows: $x < y$ if there is an index $j$ such that $x_i = y_i$ for every $i < j$, and $x_j < y_j$. 
Rank optimality with less math

Partition the constraints into classes \( \{C_1, \ldots, C_r\} \) (Stratified Dominance Hierarchy).

Compute a totalized violation vector by adding components on each class.

For structures \( s \) and \( t \), \( s \preceq t \) is decided as follows: considering the total violation
vector for each class, in the first class in which these counts are not the same for both segmentations, $s$ has fewer violations.
Valued Optimality

Start with any function $\varphi : \mathbb{N}^k \rightarrow \mathbb{N}$, and let for structures $s$ and $t$,

$$s \preceq t \quad \text{if and only if} \quad \varphi(C(s)) \leq \varphi(C(t)).$$
An OT model for stress placement

Input: sentences

Não aguento mais catalogadoras
An OT model for stress placement

**Input**: sentences

Não aguento mais catalogadoras.

**Structure**: segmentation

Não * a * guento * mais * cata * loga * doras.

* regular/pseudo segment *
core syllable
pseudo segment
An OT model for stress placement

**Input:** sentences

Não aguento mais catalogadoras.

**Structure:** segmentation

Não * a * guento * mais * cata * loga * doras.

Just turn cores into accents!
Some actual constraints

Linguistic OT hypothesis: constraints are universal, languages only differ on the ranking

- **DEP**<sub>ST</sub> - counts one violation for segment disrespecting lexical accent

- **TROCHEE** - counts one violation for segment not starting at its core
sotaq

Program for stress placement.
Specify stratification of constraints on the command line.
Input is parsed into syllables, with additional info.
Gives all optimal solutions.
Written in Perl, based on Dijsktra’s shortest path algorithm
The graph inside sotaq

Given a sentence with $n$ syllables:

1. Enumerate from 0. Add a dummy $n$’th syllable.

2. Vertices:
   - All triples $(i, j, r), 0 \leq i \leq r < j \leq n$ (segment starting with syllable $i$, ending before syllable $j$, core $r$)
   - All pairs $(i, i+1), 0 \leq i < n$, representing pseudo-segments.
   - Special vertices $(0, 0), (n, n)$.

3. Edges:
   - A directed edge from $u$ to $v$ if the second component of $u$ equals the first component of $v$.

The internal vertices in any path from $(0, 0)$ to $(n, n)$ spell a segmentation of the sentence.

Costs on vertices and edges reflect the constraints.
Conclusions, further developments

Sotaq allowed some hypotheses to be tested, with mixed degrees of success

Forced the emergence of BINGRAD/FOOTBIN