An statistical effects

Marc van Oostendorp

Workshop on West Germanic Phonology

Mannheim
Problem

- Morpheme Structure Constraints are often ‘soft’ constraints, violated by many words
- Although Optimality Theory formally is a theory of soft constraints, it does not provide the right type of softness
- We propose a revised theory of lexicon optimization to account for these facts in a truly evolutionary way
Example 1: Intervocalic geminates

- /ŋ/ is disallowed before a full vowel in Dutch (*tano) but not before schwa (eŋəl);
- this could be seen as a special case of intervocalic geminates being dispreferred before schwa: *dubəl is better than dubbo.
- however, the later is only a statistical preference, not an absolute generalisation
Example 1: Data

tense - schwa  1414  formule
lax - schwa    985  antenne
tense - full   481  aroma
lax - full     92   gorilla

(Pearson’s Chi-squared test, X-squared = 46.2844, df = 1, p-value = 1.023e-11)
Example 2: Final devoicing in Frisian

- It is well-known that Final Devoicing was not completed in Frisian (dialects) until somewhere in the 20th Century.
- Some of our sources show an interesting picture of the process being in progress; this gives the picture of words being affected one at a time.
Example 3: F voicing

As is well known, (the well-known West Germanic language) Italian has a process of intervocalic s voicing.

It is less well-known that Italian also has intervocalic f voicing, but again, this only functions as a statistic effect.
### Example 3: Data

<table>
<thead>
<tr>
<th>Letter</th>
<th>Count 1</th>
<th>Count 2</th>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>14728</td>
<td>99868</td>
<td>0.15</td>
</tr>
<tr>
<td>b</td>
<td>12682</td>
<td>78465</td>
<td>0.16</td>
</tr>
<tr>
<td>t</td>
<td>98928</td>
<td>352127</td>
<td>0.28</td>
</tr>
<tr>
<td>d</td>
<td>25896</td>
<td>122801</td>
<td>0.21</td>
</tr>
<tr>
<td>s</td>
<td>12379</td>
<td>314936</td>
<td>0.04</td>
</tr>
<tr>
<td>z</td>
<td>13166</td>
<td>55082</td>
<td>0.24</td>
</tr>
<tr>
<td>f</td>
<td>13024</td>
<td>74079</td>
<td>0.18</td>
</tr>
<tr>
<td>v</td>
<td>59412</td>
<td>87344</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Formalising the ravages of time

**Soft Morpheme Structure**
- Morpheme Structure and lexicon optimization
- Soft constraints which cannot be formalised in OT
- Lexicon Stratification

**Selective Lexicon Optimisation**
- An evolutionary interpretation of Lexicon Optimisation
- Grammar change

**Analysis**
Well-formedness

- In classical generative grammar, MSCs were treated as constraints on underlying forms
- OT does not have the possibility to deal with constraints on underlying representations
- This can often be seen as an advantage, since OT in this way famously solves the problem of *duplication*
Free rides

- e.g. Dutch does not have *[ə.V] within a morpheme (but [ə.CV] and [V.V] are allowed; Van Oostendorp 1995, Booij 1999)
- but schwa is deleted before full vowels: *Romə + -ein → romεin
- Within OT, morphemes get a free ride on this surface constraint
It is assumed that [ə.V] are not stored because of a principle of Lexicon Optimisation:

“Suppose that several different inputs I₁, I₂, . . . , Iₙ, when parsed by a grammar G lead to corresponding outputs O₁, O₂, . . . , Oₙ, all of which are realized as the same phonetic form Φ – these inputs are all phonetically equivalent with respect to G. Now one of these outputs must be the most harmonic, by virtue of incurring the least significant violation marks: suppose this optimal one is labelled Oₖ. Then the learner should choose, as the underlying orm for Φ, the input Iₖ.” (Prince en Smolensky 1993)

For refinements cf. Inkelas (1994, 2000), McCarthy (2004); for criticism see Hall (2007); Nevins and Vaux (2007)
LO Tableaux

<table>
<thead>
<tr>
<th>mən</th>
<th>NoHiatus</th>
<th>Faith-ə</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.  man</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>b.  mən</td>
<td>*!</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>man</th>
<th>NoHiatus</th>
<th>Faith-ə</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.  man</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b.  mən</td>
<td>*!</td>
<td>*</td>
</tr>
</tbody>
</table>

- /man/ → [man] is preferred since it incurs fewer faithfulness violations
- LO thus prefers lexical forms which are identical to outputs
Problem: Soft constraints are too soft

- Native parts of the lexicon tend to obey stricter templates than non-native parts
- E.g. Dutch lexical stems contain at most one full vowel plus one schwa vowel
- This is not necessarily true for loanwords (*cadeau*, *encyclopedie*)
- but words which were borrowed from Latin in Tacitean times have presently all conformed to this requirement (*kers* \(\prec\) *cerisum* ‘cherry’, *keldər* \(\prec\) *cellaria* ‘cellar’).
- The problem is that there has never been any clear period where *all* words conformed to this requirement.
Soft constraints which cannot be formalised in OT

Is this statistically significant?

- It is true for all words borrowed from Latin in the time of Tacitus
- It is true for some words borrowed from French in Napoleonic times (krant ‘newspaper’ (< courant) but papaver)
- It is not true for English word adopted in the last century
Soft constraints which cannot be formalised in OT

**Tableaux**

<table>
<thead>
<tr>
<th></th>
<th>anseklopedi</th>
<th>Faith</th>
<th>Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>anseklopedi</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>b.</td>
<td>ans</td>
<td></td>
<td>!</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>kεris</th>
<th>Faith</th>
<th>Template</th>
</tr>
</thead>
<tbody>
<tr>
<td>a.</td>
<td>kεris</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>b.</td>
<td>kεrs</td>
<td></td>
<td>!</td>
</tr>
</tbody>
</table>

- The general idea is that the lexicon can be subdivided into a number of strata.
- The strata which belong to the native lexicon have higher ranked faithfulness.
Example stratified lexicon

A: full inventory
- only $M_1$ active

B: subinventory
- $M_1, M_2, M_3$ active

C: smallest subinventory
- $M_1, M_2, M_3, M_4, M_5$ active
The Japanese lexicon

<table>
<thead>
<tr>
<th></th>
<th>Foreign</th>
<th>Sino-Japanese A</th>
<th>Sino-Japanese B</th>
<th>Yamato</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. OCP(voi)</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>observes multiple obstruent voicing ban</td>
</tr>
<tr>
<td>b. REALIZE-M(ORPHEME)</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>here: realizes compound voicing morpheme</td>
</tr>
<tr>
<td>c. NO-NÇ</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>observes postnasal voicing requirement</td>
</tr>
</tbody>
</table>

Containment relations between the inventories:
Dutch stratification

- ‘Native’: MONO ≫ FAITH (kers)
- ‘Foreign’: FAITH ≫ MONO (papaver)
Problems with a stratification analysis of the lexicon

- It is not clear *how* a word can move from one lexical stratum to the next.
- We don’t know how many strata there are, or how the child learns about these strata.
- Or *why* words have the tendency to move towards the native stratum.
- Also, it is not clear why rerankings of faithfulness and markedness constraints, as opposed to other divisions of the constraint set, are involved.
Formalising the ravages of time

Soft Morpheme Structure
Morpheme Structure and lexicon optimization
Soft constraints which cannot be formalised in OT
Lexicon Stratification

Selective Lexicon Optimisation
An evolutionary interpretation of Lexicon Optimisation
Grammar change

Analysis
Noise in the input

- Standard Lexicon Optimisation approaches start out from the assumption that the input to the child is invariable.
- We assume instead that there is random phonetic noise in the input to LO.
- Let us assume that this random noise involves deletion and insertion of vowels.
- So, if a generation $x$ has a form /kɛris/, the child may hear [kɛris, kɛrs, kɛrizi, ...]
- We still need to define a mapping from wave forms to these quasi-phonological representations; possibly Boersma-like ‘cue constraints’
An evolutionary interpretation of Lexicon Optimisation

An evolutionary interpretation

- This gives way to a truly ‘evolutionary’ view, consisting of random variation and selection.
  - *Selective Lexicon Optimisation*: In case of conflicting evidence, choose the underlying representation with the lowest violation profile
  - “Diachrony proposes, synchrony disposes”
Tableaux for Dutch

<table>
<thead>
<tr>
<th></th>
<th>FAITH</th>
<th>MONO</th>
<th>*CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>kεris</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>a. 🅩 kεris</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>b. kεrs</td>
<td></td>
<td>*!</td>
<td>*</td>
</tr>
<tr>
<td>kεrs</td>
<td>FAITH</td>
<td>MONO</td>
<td>*CC</td>
</tr>
<tr>
<td>a. kεris</td>
<td>*!</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>b. 🅩 kεrs</td>
<td></td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>
## An evolutionary interpretation of Lexicon Optimisation

### Tableau des vainqueurs

<table>
<thead>
<tr>
<th></th>
<th>FAITH</th>
<th>MONO</th>
<th>*CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. κερις → κερις</td>
<td></td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>b. Κερς → κερς</td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>
Advantages

- There is only one phonology.
- It is explained *how* a word can move from one lexical stratum to the next,
- and *why* words have the tendency to move towards the native stratum (over the course of years, the chance that the right phonetic mistakes are made, becomes larger.
- Also, it is clear why faithfulness and markedness constraints, as opposed to other divisions of the constraint set, are involved, since this distinction is relevant for LO as well.
- Effects of frequency and ‘age’ can be understood without implementing them in the grammar
Tableaux for French

<table>
<thead>
<tr>
<th>sεriz</th>
<th>FAITH</th>
<th>*CC</th>
<th>MONO</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. ☞ sεriz</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>b. sεrs</td>
<td></td>
<td>*!</td>
<td>*</td>
</tr>
<tr>
<td>sεrs</td>
<td>FAITH</td>
<td>*CC</td>
<td>MONO</td>
</tr>
<tr>
<td>a. sεriz</td>
<td>*!</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>b. ☞ sεrs</td>
<td></td>
<td></td>
<td>*</td>
</tr>
</tbody>
</table>
An evolutionary interpretation of Lexicon Optimisation

Tableau des vainqueurs

<table>
<thead>
<tr>
<th></th>
<th>FAITH</th>
<th>*CC</th>
<th>MONO</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. ʓ səriz</td>
<td></td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>b. sərs</td>
<td></td>
<td>*</td>
<td></td>
</tr>
</tbody>
</table>
Change of grammar

• If a substantial part of the grammar has changed, this may result in grammar change, i.e. reranking

• Notice that the relevant cases of Lexicon Optimisation involve $F \gg M$

• However, in acquisition theory, it is usually assumed that the unmarked (initial) order is $F \gg M$

• If there are no exceptions to $M$ anymore, the child will therefore go for the default
MSCs on inflectional content

- In Modern Germanic languages, inflectional suffixes usually consist of coronal consonants and schwa, but no full vowels.
- This was not the case for older stages of Germanic: e.g. Gothic *saiwal-os* ‘souls’
Grammar change

Lexical diffusion

- Assume a markedness constraint UNMARKEDFLEX.
- Assume IDENT-[+round] \succ UNMARKEDFLEX in Early Germanic (as in Gothic)
Gothic tableau

<table>
<thead>
<tr>
<th>saiwal + os</th>
<th>IDENT-[+round]</th>
<th>UNMARKEDFLEX</th>
</tr>
</thead>
<tbody>
<tr>
<td>saiwalos</td>
<td></td>
<td>*</td>
</tr>
<tr>
<td>saiwal∅es</td>
<td>*!</td>
<td></td>
</tr>
</tbody>
</table>

> *saiwalos* wins. However, it is not perfect (it violated the markedness constraint). Hence, there will be always some attraction to positing the underlying shape -∅es (for instance for the language learner)
### Post-Gothic tableau

<table>
<thead>
<tr>
<th>saiwal + əs</th>
<th>IDENT-+[+round]</th>
<th>WORD([+round])</th>
</tr>
</thead>
<tbody>
<tr>
<td>saiwalos</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>saiwaləs</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Now the winning form is perfect.
- Notice that at some point after this, the order $\text{IDENT-}[+\text{round}] \gg \text{WORD}([+\text{round}])$ will be no longer detectable for the child.
- Who will then assume unmarked $\text{M} \gg \text{F}$ — i.e. $\text{WORD}([+\text{round}]) \gg \text{IDENT-}[+\text{round}]$: language change completed.
Formalising the ravages of time

Soft Morpheme Structure
Morpheme Structure and lexicon optimization
Soft constraints which cannot be formalised in OT
Lexicon Stratification

Selective Lexicon Optimisation
An evolutionary interpretation of Lexicon Optimisation
Grammar change

Analysis
Example 1: Intervocalic geminates

- /ŋ/ is disallowed before a full vowel in Dutch (*tano) but not before schwa (eŋəl);
- this could be seen as a special case of intervocalic geminates being dispreferred before schwa: dubəl is better than dubbo.
- however, the later is only a statistical preference, not an absolute generalisation
Example 1: Data

tense - schwa 1414  formule
lax - schwa 985  antenne
tense - full 481  aroma
lax - full 92  gorilla

(Pearson’s Chi-squared test, $X^2 = 46.2844$, df = 1, p-value = 1.023e-11)
Analysis

- Why are geminates (and /ŋ/) dispreferred intervocalically before a full vowel, but not before a schwa?
- An obvious idea would be that this has something to do with stress; however
- *We controlled* for stress (we only looked at V₁CV₂ sequences where stress was on V₁)
- However, there is evidence that the ‘foot’ VCə has a different status than the ‘foot’ VCV (Van der Hulst and Moortgat 1981 therefore called the latter a ‘superfoot’)
Superfoot

```
SFoot
  \---
    Foot
       \---
         V ø

SFoot
  \---
    Foot
       \---
         V V
```
Evidence for the superfoot

- Reduction patterns: /fonolo̞yi/ → [fonọlọyi, fonọlẹyi, *fonọlẹyi]
- Stress is (virtually) always on the penultimate syllable in VCə, but not necessarily in VCV words
- Etc.
Crossing the superfoot boundaries

- How can the different behaviour of geminates be related to this structure?
- Crossing prosodic boundaries with a geminate is dispreferred
- The stronger the boundary, the larger the dispreference
- As far as I was able to figure out (yesterday, in the train), there are no words with the pattern VG’V (this would cross even the boundaries of the superfoot)
CrispEdge

- CRISPEDGE: Association lines (prosodic lines between segment and syllable) should not cross the boundaries of the Foot / Superfoot
Conclusions

- A classical problem of language change in OT (McMahon a.o) is that it is not clear where the change would start, which is cause and effect.
- A view of Selective Lexicon Optimisation solves this problem at least in part.
- Random phonetic variation and phonological selection may result in changes in the lexicon.
- Which in turn may result in reranking.
Inkelas (1994) has argued that we need a revised version of LO to deal with alternations:

Given a grammar $G$ and a set $S = \{S_1, S_2, \ldots S_i\}$ of surface phonetic forms for a morpheme $M$, suppose that there is a set of inputs $I = \{l_1, l_2, \ldots l_j\}$, each of whose members has a set of surface realizations equivalent to $S$. There is some $l_i \in I$ such that the mapping between $l_i$ and the members of $S$ is the most harmonic with respect to $G$, i.e. incurs the fewest marks for the highest ranked constraints. The learner should choose $l_i$ as the underlying representation for $M$. 
Turkish devoicing

- Alternating root-final plosive:
  kanat ‘wing’ kanad-ı ‘wing-Acc’
  kanat-lar ‘wing-pl’ kanad-ım ‘wing-1sg.poss’

- Nonalternating voiceless plosive:
  sanat ‘art’ sanat-ı ‘art-Acc’
  sanat-lar ‘art-pl’ sanat-ım ‘art-1sg.poss’
ALO of Turkish devoicing

<table>
<thead>
<tr>
<th>L.O.</th>
<th>FINDEV</th>
<th>IDENT-[+voice]</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. /kanad/</td>
<td>[kanat]</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>[kanad]</td>
<td>!</td>
</tr>
<tr>
<td></td>
<td>[kanadı]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[kanatı]</td>
<td>!</td>
</tr>
<tr>
<td>b. /kanaD/</td>
<td>[kanat]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>[kanad]</td>
<td>!</td>
</tr>
<tr>
<td></td>
<td>[kanadı]</td>
<td></td>
</tr>
</tbody>
</table>