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

p	14728	99868	0.15
b	12682	78465	0.16
t	98928	352127	0.28
d	25896	122801	0.21
s	12379	314936	0.04
z	13166	55082	0.24
f	13024	74079	0.18
v	59412	87344	0.68

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* → romɛin
- Within OT, morphemes get a free ride on this surface constraint

Lexicon Optimisation

- It is assumed that [ə.V] are not stored because of a principle of Lexicon Optimisation:
- "Suppose that several different inputs I_1, I_2 . . . , I_n, when parsed by a grammar *G* lead to corresponding outputs O_1 , O_2, . . . , O_n, 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 O_k . Then the learner should choose, as the underlying orm for Φ, the input I_k." (Prince en Smolensky 1993)
- For refinements cf. Inkelas (1994, 2000), McCarthy (2004); for criticism see Hall (2007); Nevins and Vaux (2007)

LO Tableaux

	mənɑn	NoHiatus	Faith-ə
a. ☞	mɑn		*
b.	mənɑn	*!	

	mɑn	NoHiatus	Faith-ə
a. ☞	mɑn		
b.	mənɑn	*!	*

- /mɑn/→[mɑn] 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

Is this statisticaly significant?

- It is true for all words borrowed from in 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

Tableaux

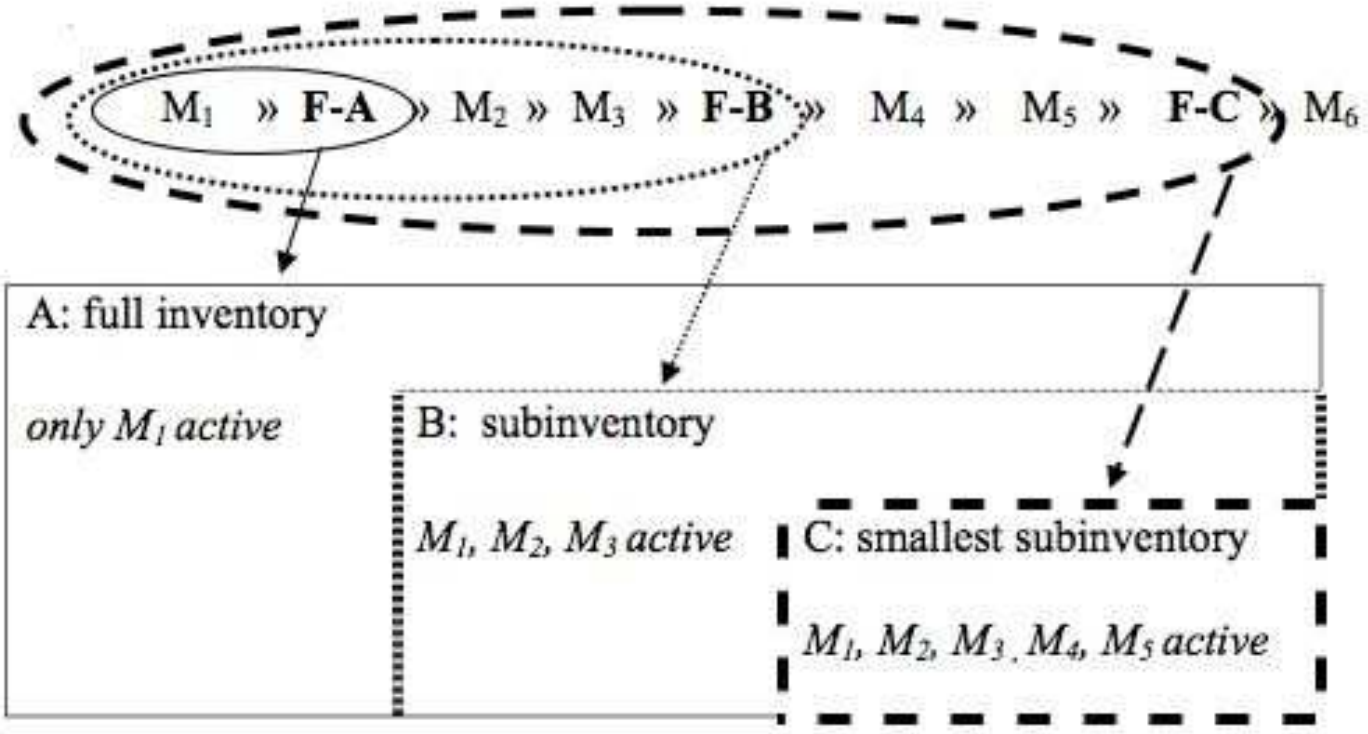
	ɑnseklopedi	Faith	Template
a. ☞	ɑnseklopedi		*
b.	ɑns	*!	

	kɛris	Faith	Template
a. ☞	kɛris		*
b.	kɛrs	*!	

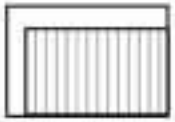
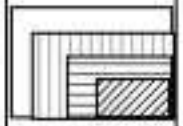
Itô and Mester (1995, 2001, 2002)

- In a series of papers, Itô and Mester (1995, 2001, 2002, 2006) have proposed an analysis of this phenomenon in terms of a stratified lexicon.
- 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



The Japanese lexicon

	Foreign	*Sino-Japanese A*	*Sino-Japanese B*	*Yamato*	
a. OCP(VOI)	no	yes	yes	yes	observes multiple obstruent voicing ban
b. REALIZE-M(ORPHEME)	no	no	yes	yes	here: realizes compound voicing morpheme
c. NO-NÇ	no	no	no	yes	observes postnasal voicing requirement
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

- 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

kɛris	FAITH	MONO	*CC
a. ☞ kɛris		*	
b. kɛrs	*!		*

kɛrs	FAITH	MONO	*CC
a. kɛris	*!	*	
b. ☞ kɛrs			*

Tableau des vainqueurs

	FAITH	MONO	*CC
a. kɛris→kɛris		*	
b. ☞ kɛrs→kɛrs			*

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

sɛriz	FAITH	*CC	MONO
a. ☞ sɛriz			*
b. sɛrs	*!	*	

sɛrs	FAITH	*CC	MONO
a. sɛriz	*!		*
b. ☞ sɛrs		*	

Tableau des vainqueurs

sɛriz	FAITH	*CC	MONO
a. ☞ sɛriz			*
b. sɛrs		*	

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≫M
- However, in acquisition theory, it is usually assumed that the unmarked (initial) order is F≫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’

Lexical diffusion

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

Gothic tableau

saiwal + os	IDENT-[+round]	UNMARKEDFLEX
☞saiwalos		*
saiwaləs	*!	

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

Post-Gothic tableau

saiwal + əs	IDENT-[+round]	WORD([+round])
saiwalos	*!	*
☞saiwaləs		

- Now the winning form is perfect.
- Notice that at some point after this, the order IDENT-[+round] ≫WORD([+round]) will be no longer detectable for the child
- Who will then assume unmarked M≫F — i.e. WORD([+round])≫IDENT-[+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

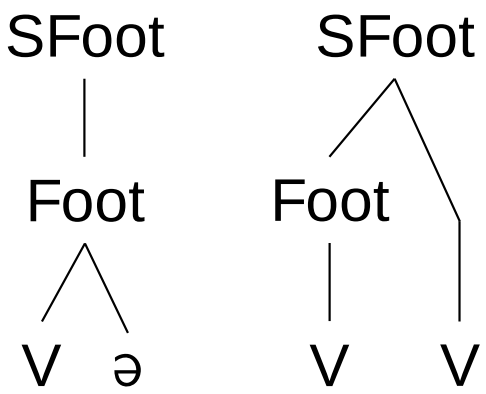
tense - schwa	1414	formule
lax - schwa	985	antenne
tense - full	481	aroma
lax - full	92	gorilla

(Pearson's Chi-squared test, X-squared = 46.2844, df = 1, p-value = 1.023e-11)

Analysis

- Why are geminates (and /ŋ/) disprefered 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 V_1CV_2 sequences where stress was on V_1)
- 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



Evidence for the superfoot

- Reduction patterns: /fonoloɣi/ → [fonəloɣi, fonələɣi, *fonoləɣi]
- 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.

Alternant Optimization

- 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 $I = \{I_1, I_2, \ldots I_j\}$, each of whose members has a set of surface realizations equivalent to S. There is some $I_i \in I$ such that the mapping between I_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 I_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

-

L.O.			FINDEV	IDENT-[+voice]
	a. /kanad/	☞[kanat]		*
		[kanad]	*!	
		☞[kanadı]		
		[kanatı]		*!
	b. /kanaD/	☞[kanat]		
		[kanad]	*!	
		☞[kanadı]		