Abstract Phonotactic Constraints for Speech Segmentation:

Evidence from Human and Computational Learners

Frans Adriaans, Natalie Boll-Avetisyan
& René Kager
UiL-OTS, Utrecht University

4. März 2009, DGfS Meeting, Osnabrück

Phonology is abstract

- Phonotactic constraints often affect all members of a group of phonemes that share features (i.e. natural classes)
- Example:
  - OCP-Place

OCP-Place

- OCP-Place: Avoid consonant sequences that share feature [place]
  - e.g. no labial-labial {p, b, f, v, m}
- Avoidance of labial sequences in Dutch words (e.g. ?*smaf*)
- This constraint is psychologically real.
  - Well-formedness judgments
    (Hebrew: Berent & Shimron, 1997; Arabic: Frisch & Zawaydeh, 2001)
  - Lexical decision
    (Dutch: Kager & Shatzman, 2007)

Questions

1. Why do we have abstract phonotactic constraints?
2. How are such constraints acquired?

Experiments with humans to answer question 1

Computer simulations to answer question 2

Abstract phonotactics for segmentation?

- In Dutch, words cannot start with /mr/

  mr → m.r
- Dutch listeners use this knowledge to segment words from speech (McQueen, 1998)
- A role for abstract phonotactic constraints in segmentation?
- Is abstract OCP-Lab used in segmentation?

Human learners: Experiment

- Approach:
  - Artificial language learning experiment
- Artificial languages are highly reduced miniature languages. (e.g. Saffran et al., 1996)
- Construct an artificial language which contains no cues for segmentation but OCP-Lab.

(Boll-Avetisyan & Kager, 2008)

OCP-Lab for segmentation

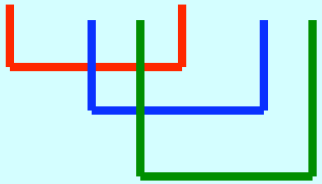
Exposed to an artificial stream of speech such as:

…P P T P P T P P T P P T P P T P P T…

P = labials {p, b, m} T = coronals {t, d, n}

Where will participants place word-boundaries?

…P P T P P T P P T P P T P P T P P T…



Prediction

	OCP-Lab
…PTP-PTP-PTP-PTP…	
…PPT-PPT-PPT-PPT…	*
…TPP-TPP-TPP-TPP…	*

- Segmentations that satisfy OCP-Lab should be preferred.

The artificial language

Position1	Position 2	Position 3	Position1	Position 2
Lab-1	**Lab-2**	**Cor**	**Lab-1**	**Lab-2**
pa	po	tu	pa	po
bi	be	do	bi	be
mo	ma	ne	mo	ma

…pamatumomatubibetumobedomoponepabe…

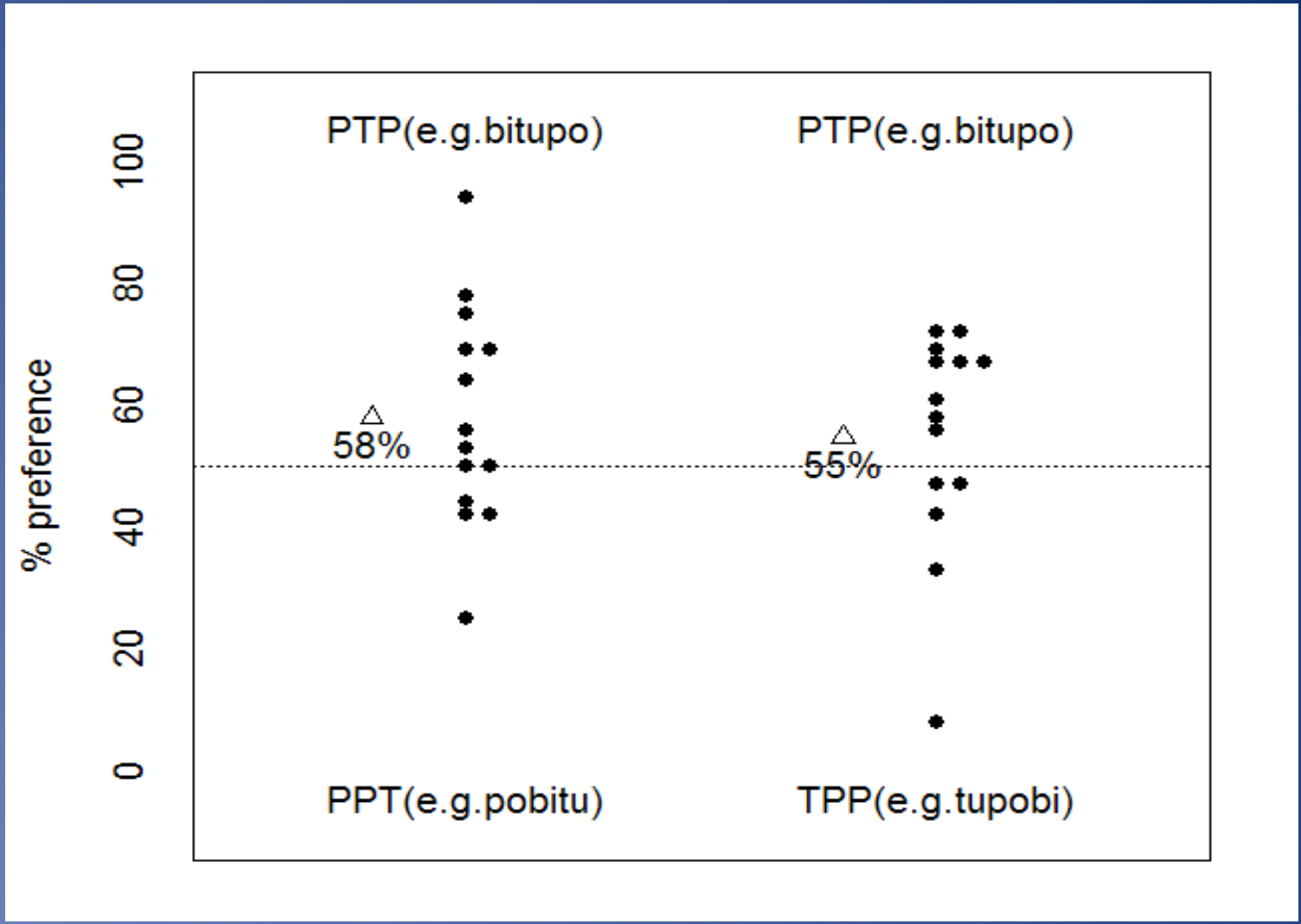
Procedure

1 language, 2 test conditions

Task: 2-Alternative Forced Choice

	Condition	Example
1.	PTP > PPT	potubi > pobitu
2.	PTP > TPP	potubi > tupobi

Results overview



PTP > PPT ** PTP > TPP *

Do the human results support *abstract* OCP-Lab?

- Does OCP-Lab do better than statistical predictors?
- Co-occurrence probabilities over $C_1C_2C_3$:
  - O/E ratio — $O/E = P(xy) / P(x)*P(y)$
  - Transitional probability — $TP = P(xy) / P(x)$
- Stepwise linear regression:

	R^2(OCP)	R^2(O/E)	OCP + O/E	O/E + OCP
	0.2757**	0.2241*	OCP**	O/E**, OCP*

	R^2(OCP)	R^2(TP)	OCP + TP	TP + OCP
	0.2757**	0.0372	OCP**	OCP*

Interim summary

- Human learners use an abstract constraint from their L1 to segment artificial speech.
- This raises questions:
  - Where did this constraint come from?
  - Did participants use OCP-Lab, or might they have used alternative constraints?

Computational learners

- Goal: To provide a computational account of the learning of abstract constraints for segmentation
- Constraint induction model:
  - **StaGe** (Adriaans, 2007; Adriaans & Kager, submitted)
- Approach:
  - Train StaGe on non-adjacent consonants in Dutch corpus
  - Segment the artificial language using induced constraint set
  - Does StaGe accurately predict human performance in the ALL experiment?

StaGe - Background

- Induction of phonotactics from continuous speech…
- … implementing two human/infant learning mechanisms:
  - Statistical learning (e.g. Saffran, Newport & Aslin, 1996)
  - Generalization (e.g. Saffran & Thiessen, 2003)

  → pre-lexical infants learn from continuous speech input
- Previous study:
  - Feature-based abstraction over statistically learned biphone constraints improves segmentation performance

    (Adriaans & Kager, submitted)

StaGe - The model

1. Statistical learning
   - Biphone probabilities (O/E ratio) in continuous speech
2. Frequency-Driven Constraint Induction
   - Categorization of biphones using O/E ratio

Category	Constraint	Interpretation
low	*xy	'Sequence xy should not be kept intact.'
high	Contig-IO(xy)	'Sequence xy should be kept intact.'
neutral	-	-

3. Single-Feature Abstraction
   - Generalization over phonologically similar biphone constraints
   - Similarity = number of shared features
   - ⇒ Constraints on natural classes

STAGE - Examples (1)

1. Frequency-Driven Constraint Induction:
   - *tl, Contig-IO(pr), Contig-IO(bl), etc.
2. Single-Feature Abstraction:
   - Contig-IO(pl)
     Contig-IO(bl)
     Contig-IO(pr)
     Contig-IO(dr)

     $\Rightarrow$ **Contig-IO(x $\in$ {p,b,t,d}, y $\in$ {l,r})**

StaGe - Examples (2)

- Generalization affects statistically neutral biphones (e.g. /tr/)

Input: tr	*tl	Contig-IO(x ∈ {p,b,t,d}, y ∈ {l,r})
→ tr		
t.r		*

- Frequency-based constraint ranking captures exceptions to generalizations:

Input: tl	*tl	Contig-IO(x ∈ {p,b,t,d}, y ∈ {l,r})
tl	*	
→ t.l		*

The current study

- What type of L1 phonotactic knowledge did participants in the ALL experiment use?
- Three options:
  1. OCP-Lab
  2. Consonant co-occurrence probabilities (O/E ratio)
  3. StaGe (Statistically learned constraints + generalizations)

  → Does StaGe provide a better fit to human data than segment co-occurrence probabilities alone?

  → Does StaGe lead to the induction of OCP-Lab?

Simulations

- Training data:
  1. CGN (Spoken Dutch Corpus, continuous speech)
  2. CELEX (Dutch lexicon, word types)
- Test:
  - Segmentation of artificial language
- Linking computational models to human data:
  - Frequencies of test items in model's segmentation output
  - Linear regression: Item frequencies as predictor for human judgements on those items

Item scores (PTP-PPT)

ITEM	HUMAN	OCP	(CGN) O/E ratio	(CGN) StaGe	(CELEX) O/E ratio	(CELEX) StaGe
madomo	0.8095	39	39	16	39	16
ponebi	0.7381	34	21	18	25	17
ponemo	0.7381	36	20	26	20	27
podomo	0.6905	38	17	26	29	31
madobi	0.5714	32	30	4	32	12
madopa	0.5714	25	3	3	3	0
ponepa	0.5714	35	19	16	19	24
podobi	0.5476	38	17	24	29	20
potumo	0.5476	33	23	4	23	29
podopa	0.4762	40	4	8	14	0
potubi	0.4524	37	20	3	23	20
potupa	0.2381	33	14	2	14	21
mobedo	0.5476	0	0	0	0	0
pabene	0.5476	0	0	2	0	1
papone	0.5000	0	0	0	0	0
mobetu	0.4524	0	0	0	0	0
papodo	0.4524	0	0	0	0	4
pabedo	0.4048	0	0	0	0	0
pamado	0.4048	0	0	1	0	8
pamatu	0.4048	0	0	1	0	1
papotu	0.3810	0	0	0	0	0
pabetu	0.3571	0	2	0	2	0
pamane	0.3333	0	0	1	0	0
mobene	0.2619	0	0	0	0	0

Analysis 1

- STAGE adds feature-based generalization to statistical learning (O/E)
- Added value of feature-based generalization in explaining human scores?
  - CGN continuous speech: yes
  - CELEX word types: no
- Stepwise linear regression:

CORPUS	R^2(O/E)	R^2(StaGe)	O/E + StaGe	StaGe + O/E
CGN	0.3969 ***	0.5111 ***	O/E***, StaGe**	StaGe***
CELEX	0.4140 ***	0.2135 *	O/E***	StaGe**, O/E*

Analysis 2

- Does STAGE lead to the induction of OCP-Lab?
- R^2(OCP) = 0.2917 **
- Stepwise linear regression:

CORPUS	R^2(StaGe)	OCP + StaGe	StaGe + OCP
CGN	0.5111 ***	OCP**, StaGe**	StaGe***
CELEX	0.2135 *	OCP**	StaGe*

→ StaGe/CGN is the best predictor of the human data

→ OCP-Lab and StaGe/CELEX indistiguishable

Analysis 2: OCP?

- Constraints used in segmentation of the AL:

StaGe/CGN:

CONSTRAINT	RANKING
*[b]_[m]	1480.8816
*[m]_[pf]	1360.1801
*[m]_[pbfv]	1219.1565
*[C]_[pt]	376.2584
*[pbfv]_[pbtdfvsz]	337.7910
*[pf]_[C]	295.7494
*[C]_[tsS]	288.4389
*[pbfv]_[tdszSZ_]	287.5739
*[C]_[pbtd]	229.1519
*[pbfv]_[pbfv]	176.0199
*[C]_[C]	138.7298

(C = obstruents = [pbtdkgfvszSZxGh_])

StaGe/CELEX:

CONSTRAINT	RANKING
Contig-IO([m]_[n])	1206.1391
*[m]_[m]	491.4118
*[bv]_[pt]	412.0674
*[bdvz]_[pt]	395.7393
*[p]_[m]	386.4478
*[b]_[p]	323.8216
*[m]_[p]	320.2785
*[m]_[pb]	238.1173
*[pbfv]_[pt]	225.2524
*[bv]_[pbtd]	224.6637
*[pbtdfvsz]_[pt]	207.4790
*[bdvz]_[pbtd]	207.1846
*[pbfv]_[p]	195.9116
*[bv]_[pb]	194.7343
*[pbfv]_[pbfv]	133.0241
*[pbtdfvsz]_[pbtd]	108.3970
*[C]_[C]	54.9204
Contig-IO([C]_[C])	8.6359

Analysis 2: OCP?

- StaGe learns "OCP-ish" constraints
- StaGe/CGN has a preference for /p/-initial words:

Input: *bipodomo*	*C_{p,t}
→ *bi.podomo*	
bipo.domo	*
bipodo.mo	*
bipodomo	*

→ Align-{p,t}

- Unless the following consonant is /t/:

Input: *bipotubi*	*C_{p,t}	*{p,f}_C
bi.potubi	*	*
→ *bipo.tubi*	*	
bipotu.bi	**	*
bipotubi	**	*

OCP, StaGe/CELEX
→ *bi.potubi*

Analysis 2: OCP?

ITEM	HUMAN	OCP	(CGN) O/E ratio	(CGN) StaGe	(CELEX) O/E ratio	(CELEX) StaGe
madomo	0.8095	39	39	16	39	16
ponebi	**0.7381**	**34**	21	**18**	25	**17**
ponemo	**0.7381**	**36**	20	**26**	20	**27**
podomo	**0.6905**	**38**	17	**26**	29	**31**
madobi	0.5714	32	30	4	32	12
madopa	0.5714	25	3	3	3	0
ponepa	**0.5714**	**35**	19	**16**	19	**24**
podobi	**0.5476**	**38**	17	**24**	29	**20**
potumo	**0.5476**	**33**	23	**4**	23	**29**
podopa	**0.4762**	**40**	4	**8**	14	**0**
potubi	**0.4524**	**37**	20	**3**	23	**20**
potupa	**0.2381**	**33**	14	**2**	14	**21**
mobedo	0.5476	0	0	0	0	0
pabene	0.5476	0	0	2	0	1
papone	0.5000	0	0	0	0	0
mobetu	0.4524	0	0	0	0	0
papodo	0.4524	0	0	0	0	4
pabedo	0.4048	0	0	0	0	0
pamado	0.4048	0	0	1	0	8
pamatu	0.4048	0	0	1	0	1
papotu	0.3810	0	0	0	0	0
pabetu	0.3571	0	2	0	2	0
pamane	0.3333	0	0	1	0	0
mobene	0.2619	0	0	0	0	0

Conclusion (1)

- Human learners use abstract phonotactic constraints for artificial language segmentation
- Computational learners can be used to simulate the learning of such constraints
- StaGe learns OCP-like and Align-like constraints…
- … from continuous speech
- → best predictor of human data in current experiment

Conclusion (2)

- There is more to phonotactics and speech segmentation than segment co-occurrence probabilities

→ Importance of feature-based generalization in phonotactic learning and segmentation