Simulating Language Behavior
An Introduction

Çağrı Çöltekin

c.coltekin@rug.nl

Information science/Informatiekunde

2012-02-15

Tentative Plan

Week	Subject
1	Introduction & Organization
2	Computational simulation of language acquisition (mostly segmentation)
3	Simulation of language change/diffusion Simulation of learning pronoun reference
4	Simulation of segmentation
5	Simulation of segmentation
6	Simulation of language comprehension Simulation of acquisition of words, morphology or syntax
7	Project presentations

Outline

What is language behavior?

We will be dealing with questions like:

- How does a particular aspect of language comprehension?
  - Why some sentences are harder to comprehend than others?
- How do we acquire language(s)?
  - Is there a difference between learning regular or irregular aspects of language?
- How do languages change in time?
  - What are the causes of language change, and in which ways do we expect changes to occur?

What is language behavior?

We will be dealing with questions like:

- How does a particular aspect of language comprehension?
  - Why some sentences are harder to comprehend than others?
- How do we acquire language(s)?
  - Is there a difference between learning regular or irregular aspects of language?
- How do languages change in time?
  - What are the causes of language change, and in which ways do we expect changes to occur?

Note that the questions are not only related to the directly observables. It relates to *how human cognitive system works*.

What is a model?

Examples of some models in science:

- Galilean model of solar system.
- Bohr model of atom.
- Atmospheric models used in meteorology.
- Scale models of cars, bridges, buildings etc. used in engineering.
- Animal models used in medicine.

Models: why and how

- Why do we model things at all?

Models: why and how

- Why do we model things at all?
  - If the model matches the reality well, we can make predictions.
  - We learn the phenomenon better while (formally) specifying the model.
  - Sometimes cannot study the object of interest directly. Because it is too, expensive, unethical, or unpractical to do so.

Models: why and how

- Why do we model things at all?
  - If the model matches the reality well, we can make predictions.
  - We learn the phenomenon better while (formally) specifying the model.
  - Sometimes cannot study the object of interest directly. Because it is too, expensive, unethical, or unpractical to do so.
- Once we have the model, how do we get knowledge out of it?

Models: why and how

- Why do we model things at all?
  - If the model matches the reality well, we can make predictions.
  - We learn the phenomenon better while (formally) specifying the model.
  - Sometimes cannot study the object of interest directly. Because it is too, expensive, unethical, or unpractical to do so.
- Once we have the model, how do we get knowledge out of it?
  - Study the model analytically.
  - Run simulations.

Models: why and how

- Why do we model things at all?
  - If the model matches the reality well, we can make predictions.
  - We learn the phenomenon better while (formally) specifying the model.
  - Sometimes cannot study the object of interest directly. Because it is too, expensive, unethical, or unpractical to do so.
- Once we have the model, how do we get knowledge out of it?
  - Study the model analytically.
  - Run simulations.

> All models are wrong, some are useful.
> — *Box and Draper (1986, p. 424)*

Language Acquisition

The problem of language acquisition

- Human languages are complex (recursion, ambiguity).
- Children do not receive explicit instruction during language acquisition.
- Language acquisition by children is (arguably) fast and robust.
- The input to children is not enough for learning (*Poverty of Stimulus Argument*).
  - Children do not receive input critical for learning certain phenomena.
  - Human languages are not learnable from positive input (Gold, 1967). Negative input is not available to children.

The debate

Nativism Our knowledge of language is largely determined at birth (by our genes). Contribution of environmental factors are only of secondary importance.

> *[...] in certain fundamental respects we do not really learn language; rather, grammar grows in the mind. (Chomsky, 1980, p.134)*

Plato, Descartes, Chomsky, ...

Empiricism Our knowledge is primarily due to our interactions with the environment.

Aristotle, Locke, ...

Taking one of these sides is common in linguistics.

Debate resolved: we are all nativists

To say that "language is not innate" is to say that there is no difference between my granddaughter, a rock, and a rabbit. In other words, if you take a rock, a rabbit, and my granddaughter and put them in a community where people are talking English, they'll all learn English. If people believe that, then they'll believe language is not innate. If they believe that there is a difference between my granddaughter, a rabbit, and a rock, then they believe that language is innate.
— Chomsky (2000, p.50), 'The Architecture of Language' (emphasis mine.)

Debate resolved: we are all empiricist

The obvious conclusion is that the real answer to the question, Where the knowledge come from, is that it comes from the interaction between nature and nurture, or what has been called "epigenesis." Genetic constraints interact with internal and external environmental influences, and they jointly give rise to the phenotype.
— Elman et al. (1996, pp.i–ii), 'Rethinking Innateness'

Debate in linguistics

We all agree that,

- Part of our linguistic abilities comes from our experience: people are typically able to learn more different languages than they grow different physical organs.
- Part of our linguistic abilities are innate: rocks and rabbits aside, even the species closest to us cannot match with our linguistic abilities.

Debate in linguistics

We all agree that,

- Part of our linguistic abilities comes from our experience: people are typically able to learn more different languages than they grow different physical organs.
- Part of our linguistic abilities are innate: rocks and rabbits aside, even the species closest to us cannot match with our linguistic abilities.

The disagreement seems to be on *whether the innate component is language-specific knowledge or domain-general learning abilities*.

Now we know what it is, is the debate resolved?

Short answer:

Now we know what it is, is the debate resolved?

Short answer: No.

- It is difficult to know the quantity/type of innate knowledge necessary for settling the debate: The target seems to be moving: from *P&P* (Chomsky, 1981) to *recursion* (Hauser, Chomsky & Fitch, 2002) / *merge* (Berwick et al., 2011).

Now we know what it is, is the debate resolved?

Short answer: No.

- It is difficult to know the quantity/type of innate knowledge necessary for settling the debate: The target seems to be moving: from *P&P* (Chomsky, 1981) to *recursion* (Hauser, Chomsky & Fitch, 2002) / *merge* (Berwick et al., 2011).
- Empirical evidence is scarce, and interpreted differently.

Now we know what it is, is the debate resolved?

Short answer: No.

- It is difficult to know the quantity/type of innate knowledge necessary for settling the debate: The target seems to be moving: from *P&P* (Chomsky, 1981) to *recursion* (Hauser, Chomsky & Fitch, 2002) / *merge* (Berwick et al., 2011).
- Empirical evidence is scarce, and interpreted differently.
- 'Logical arguments' are either clearly false, or misunderstood in the community at large.

... but didn't Gold (1967) prove it already?

- After Gold's (1967), there have been many different results in the field, which are typically ignored.
- Modeling is useful, but while interpreting results of models we need to consider the match between the model and the real world. In language learning case:
  - Is the formal grammar a good candidate for the natural grammar?
  - Is learning method a plausible one?
  - Is the characterization of the input match with the real-world setting?

... but didn't Gold (1967) prove it already?

- After Gold's (1967), there have been many different results in the field, which are typically ignored.
- Modeling is useful, but while interpreting results of models we need to consider the match between the model and the real world. In language learning case:
  - Is the formal grammar a good candidate for the natural grammar?
  - Is learning method a plausible one?
  - Is the characterization of the input match with the real-world setting?

Computational models are useful for investigating some arguments in the debate, but the results are unlikely to be conclusive.

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers
- prime numbers

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers
- prime numbers
- odd prime numbers

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers
- prime numbers
- odd prime numbers
- just the list of randomly chosen numbers given so far

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers
- prime numbers
- odd prime numbers
- just the list of randomly chosen numbers given so far
- ...

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers
- prime numbers
- odd prime numbers
- just the list of randomly chosen numbers given so far
- ...

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers
- prime numbers
- odd prime numbers
- just the list of randomly chosen numbers given so far
- ...

We cannot know for certain, since there are infinite number of possible input sequences (input sentences), and infinite number of ways to characterize them (grammars).

An informal game to understand Gold's results

Try to guess what the sequence of the given numbers are..

- 7, 11, 13, 17
- 5, 7, 11, 13
- 13, 17, 19, 23

- ordered sequence of prime numbers
- prime numbers
- odd prime numbers
- just the list of randomly chosen numbers given so far
- ...

We cannot know for certain, since there are infinite number of possible input sequences (input sentences), and infinite number of ways to characterize them (grammars).

Natural languages are not provably unlearnable. Even if they were learnable, still we cannot arrive at an empiricist conclusion: the initial knowledge in these models are rather complex.

... if we knew the language is

not innate we clearly could not solve 'Plato's problem'. We may have substantial innate knowledge in another domain of cognition.

innate it would not necessarily solve the bigger debate either. We know more clear genetically determined factors affecting our cognition. Yet, these do not seem to have declared the victory for nativism.

To sum up...

- The nature–nurture debate is intriguing, yet an unresolved debate (which should eventually be resolved by research in neuroscience and genetics)

To sum up...

- The nature–nurture debate is intriguing, yet an unresolved debate (which should eventually be resolved by research in neuroscience and genetics)
- It has a central role in linguistics. I believe this role is not well motivated:

To sum up...

- The nature–nurture debate is intriguing, yet an unresolved debate (which should eventually be resolved by research in neuroscience and genetics)
- It has a central role in linguistics. I believe this role is not well motivated:
  - Linguistics is just any other domain that may contribute to the debate.

To sum up...

- The nature–nurture debate is intriguing, yet an unresolved debate (which should eventually be resolved by research in neuroscience and genetics)
- It has a central role in linguistics. I believe this role is not well motivated:
  - Linguistics is just any other domain that may contribute to the debate.
  - The contribution of the debate to the study of language is uncertain in many cases.

To sum up...

- The nature–nurture debate is intriguing, yet an unresolved debate (which should eventually be resolved by research in neuroscience and genetics)
- It has a central role in linguistics. I believe this role is not well motivated:
  - Linguistics is just any other domain that may contribute to the debate.
  - The contribution of the debate to the study of language is uncertain in many cases.
  - More importantly, taking a priori sides in this unresolved debate can be unfruitful and even misleading.

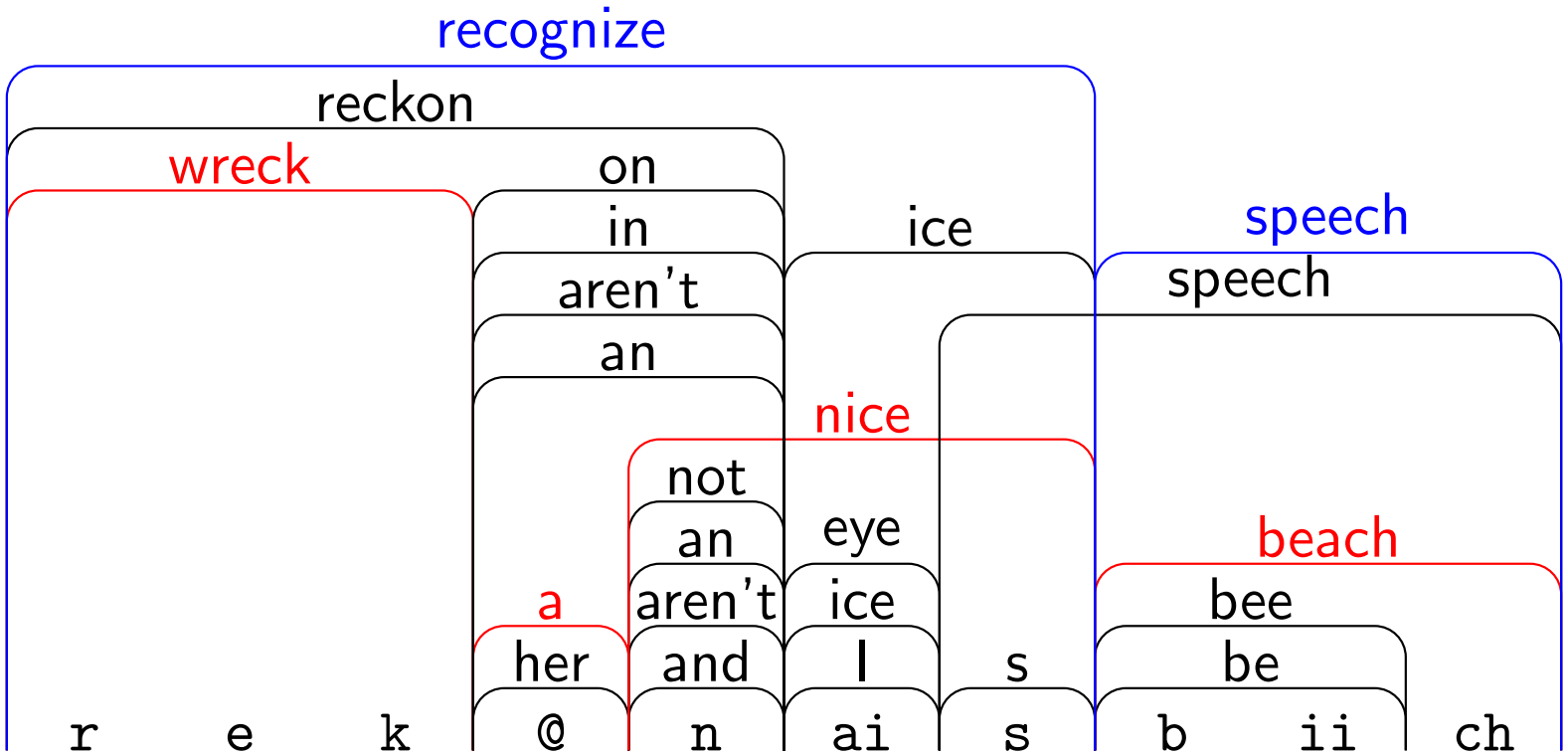
An example: segmentation

Difficulties of learning segmentation

- No clear acouistic markers in fluent speech.
- Large speaker variation in acoustic input.
- Noise in the environmet.
- Children has to start with no knwledge of words.
- Even with a comprehensive knowledge of words, segmentation is still difficult because of multiple plausible segmentations.
  For example:
  /6go/: /6go/ 'ago' or /6 go/ 'a go'?
  /Itsnoz/: /Its noz/ 'its nose' or /It snoz/ 'it snows'?

Recognize speech, or, wreck a nice beach

An automatic speech recognizers attempt to recognize the phrase 'recognize speech':



*Example reproduced from: (Shillcock, 1995)

The puzzle to solve

ljuuzuibutsjhiuljuuz

ljuuztbzjubhbjompwfljuuz

xibutuibu

ljuuz

epzpvxbounpsfnjmlipofz

ljuuzljuuzephhjf

opnjxibuepftbljuuztbz

xibuepftbljuuztbz

ephhjfeph

ephhjf

opnjxibuepftuifephhjftbz

xibuepftuifephhjftbz

mjuumfcbczcjsejf

cbczcjsejf

zpvepoumjlfuibupof

plbznpnnzublfuijtpvu

dpx

uifdpxtbztnppnpp

xibuepftuifdpxtbzopnj

The puzzle to solve

ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj

- No clear boundary markers
- No lexical knowledge

How do children segment?

Children very early in life (8-months) seem to be sensitive to statistical regularities between syllables (Saffran, Aslin, Newport 1996)

How do children segment?

Children very early in life (8-months) seem to be sensitive to statistical regularities between syllables (Saffran, Aslin, Newport 1996)

Training: bidakupadotigolabubidakugolabupadoti...

How do children segment?

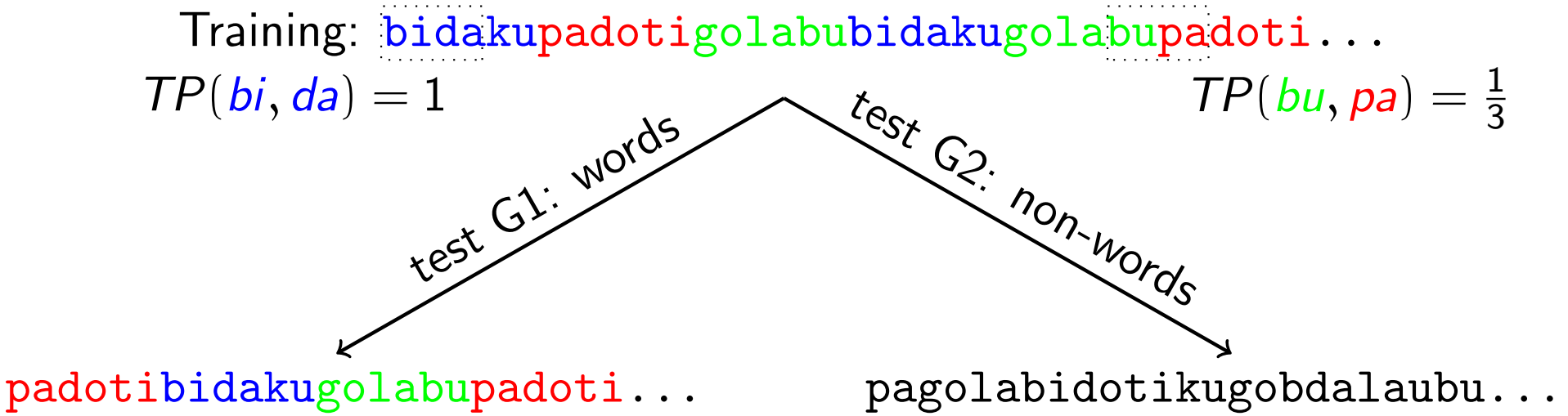
Children very early in life (8-months) seem to be sensitive to statistical regularities between syllables (Saffran, Aslin, Newport 1996)

Training: bidakupadotigolabubidakugolabupadoti...

$TP(bi, da) = 1$ $TP(bu, pa) = \frac{1}{3}$

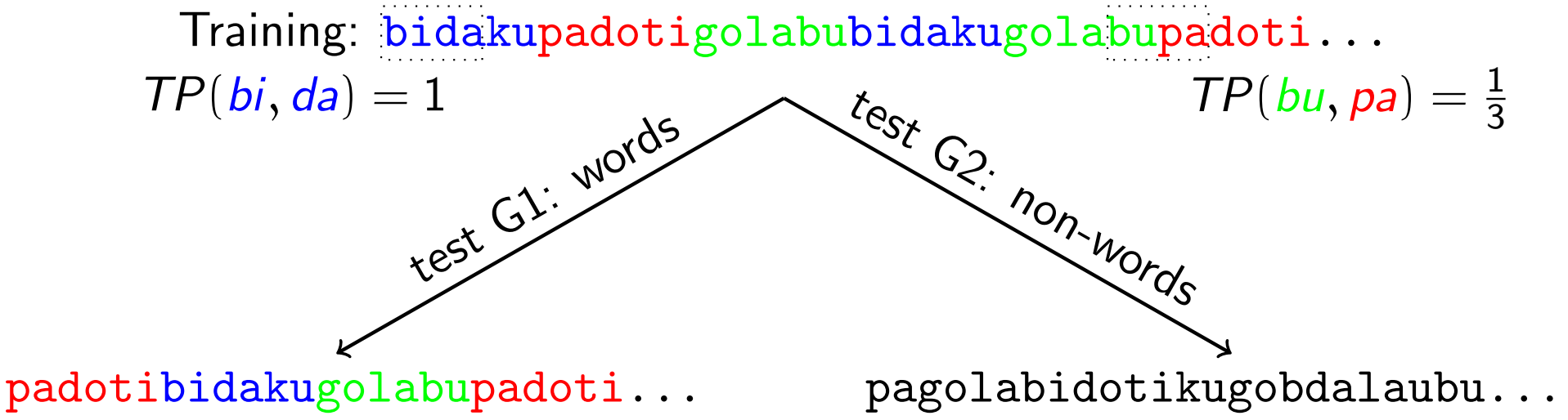
How do children segment?

Children very early in life (8-months) seem to be sensitive to statistical regularities between syllables (Saffran, Aslin, Newport 1996)



How do children segment?

Children very early in life (8-months) seem to be sensitive to statistical regularities between syllables (Saffran, Aslin, Newport 1996)



Children showed preference towards the 'words' that are used in the training phase.

Predictability

Predictability within units is high, predictability between units is low.

Predictability

Predictability within units is high, predictability between units is low.

Given a sequence `lr`, where `l` and `r` are sequences of phonemes:

- If `l` help us predict `r`, `lr` is likely to be part of a word.
- If observing `r` after `l` is surprising it is likley that there is a boundary between `l` and `r`.

Predictability

Predictability within units is high, predictability between units is low.

Given a sequence `lr`, where `l` and `r` are sequences of phonemes:

- If `l` help us predict `r`, `lr` is likely to be part of a word.
- If observing `r` after `l` is surprising it is likley that there is a boundary between `l` and `r`.

The strategy dates back to 1950s (Haris, 1955), where he used a measure called *successor variety* (SV):

The morpheme boundaries are at the locations where there is a high variety of possible phonemes that follow the initial segment.

Back to the puzzle: some cues

ljuuzuibutsjhiuljuuz

ljuuztbzjubhbjompwfljuuz

xibutuibu

ljuuz

epzpvxbounpsfnjmlipofz

ljuuzljuuzephhjf

opnjxibuepftbljuuztbz

xibuepftbljuuztbz

ephhjfeph

ephhjf

opnjxibuepftuifephhjftbz

xibuepftuifephhjftbz

mjuumfcbczcjsejf

cbczcjsejf

zpvepoumjlfuibupof

plbznpnnzublfuijtpvu

dpx

uifdpxtbztnppnpp

xibuepftuifdpxtbzopnj

Back to the puzzle: some cues

`ljuuzuibutsjhiuljuuz`
`ljuuztbzjubhbjompwfljuuz`
`xibutuibu`
`ljuuz`
`epzpvxbounpsfnjmlipofz`
`ljuuzljuuzephhjf`
`opnjxibuepftbljuuztbz`
`xibuepftbljuuztbz`
`ephhjfeph`
`ephhjf`
`opnjxibuepftuifephhjftbz`
`xibuepftuifephhjftbz`
`mjuumfcbczcjsejf`
`cbczcjsejf`
`zpvepoumjlfuibupof`
`plbznpnnzublfuijtpvu`
`dpx`
`uifdpxtbztnppnpp`
`xibuepftuifdpxtbzopnj`

Cues for the solution:

- Acoustic cues, such as *pauses*, *stress*, *coarticulation*, *allophonic alternations*, *vowel harmony*

Back to the puzzle: some cues

ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj

Cues for the solution:

- Acoustic cues, such as *pauses*, *stress*, *coarticulation*, *allophonic alternations*, *vowel harmony*
- lexical knowledge

Back to the puzzle: some cues

```
ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj
```

Cues for the solution:

- Acoustic cues, such as *pauses*, *stress*, *coarticulation*, *allophonic alternations*, *vowel harmony*
- lexical knowledge
- phonotactics

Back to the puzzle: some cues

ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj

Cues for the solution:

- Acoustic cues, such as *pauses*, *stress*, *coarticulation*, *allophonic alternations*, *vowel harmony*
- lexical knowledge
- phonotactics
- utterance boundaries

Back to the puzzle: some cues

ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj

Cues for the solution:

- Acoustic cues, such as *pauses*, *stress*, *coarticulation*, *allophonic alternations*, *vowel harmony*
- lexical knowledge
- phonotactics
- utterance boundaries
- distributional regularities

Back to the puzzle: some cues

ljuuzuibutsjhiuljuuz
ljuuztbzjubhbjompwfljuuz
xibutuibu
ljuuz
epzpvxbounpsfnjmlipofz
ljuuzljuuzephhjf
opnjxibuepftbljuuztbz
xibuepftbljuuztbz
ephhjfeph
ephhjf
opnjxibuepftuifephhjftbz
xibuepftuifephhjftbz
mjuumfcbczcjsejf
cbczcjsejf
zpvepoumjlfuibupof
plbznpnnzublfuijtpvu
dpx
uifdpxtbztnppnpp
xibuepftuifdpxtbzopnj

Cues for the solution:

- Acoustic cues, such as *pauses*, *stress*, *coarticulation*, *allophonic alternations*, *vowel harmony*
- lexical knowledge
- phonotactics
- utterance boundaries
- distributional regularities
- predictability

$$\mathrm{TP}(\mathrm{ju}) = 11/27 = 0.40$$
$$\mathrm{TP}(\mathrm{zu}) = 2/23 = 0.08$$

Measures of (un)predictability

- Transitional probability

$$\mathrm{TP}(\mathtt{l},\mathtt{r}) = \frac{P(\mathtt{lr})}{P(\mathtt{l})}$$

- Successor value

$$\mathrm{SV}(\mathtt{l}) = \sum_{\mathtt{r}\in A} c(\mathtt{l},\mathtt{r})$$

- Pointwise mutual information

$$\mathrm{MI}(\mathtt{l},\mathtt{r}) = log_2 \frac{P(\mathtt{lr})}{P(\mathtt{l})P(\mathtt{r})}$$

- Boundary entropy

$$\mathrm{H}(\mathtt{l}) = -\sum_{\mathtt{r}\in A} P(\mathtt{r}|\mathtt{l})\ log_2 P\,(\mathtt{r}|\mathtt{l})$$

The assymmetric measures have their 'reverse' counterparts.

The length of the sequences `l` and `r` matters.

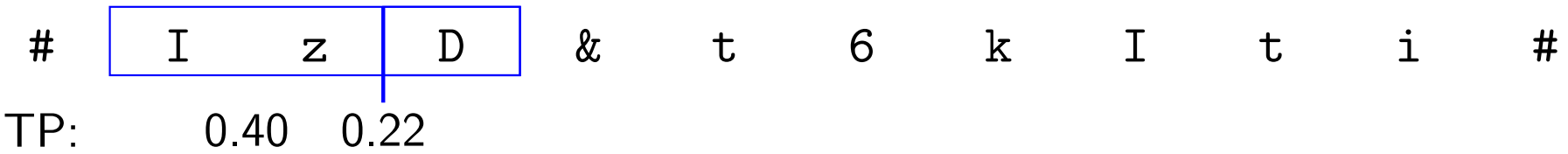
How to Calculate the Measures

I z D & t 6 k I t i

How to Calculate the Measures

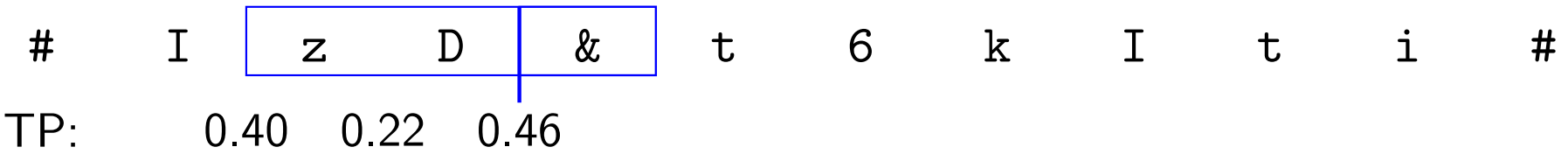
$$TP(\texttt{\#I}, \texttt{z}) = P(\texttt{z}|\texttt{\#I}) = 0.40$$

How to Calculate the Measures



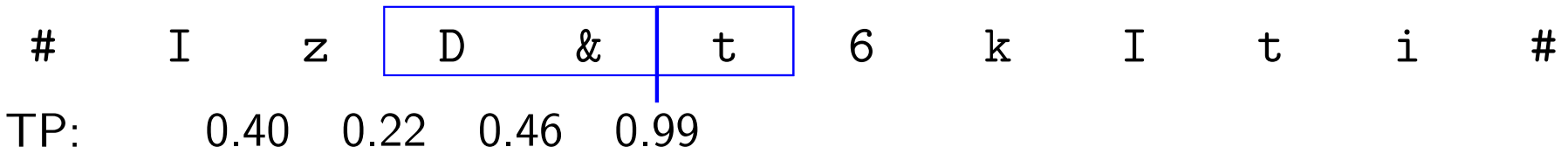
$$TP(\texttt{Iz}, \texttt{D}) = P(\texttt{D}|\texttt{Iz}) = 0.22$$

How to Calculate the Measures



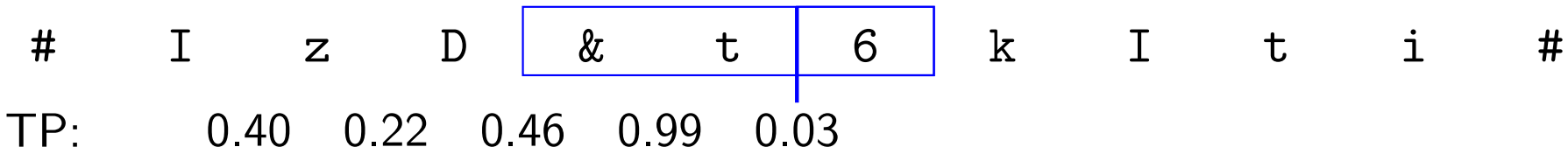
$$TP(\texttt{zD}, \texttt{\&}) = P(\texttt{\&}|\texttt{zD}) = 0.46$$

How to Calculate the Measures



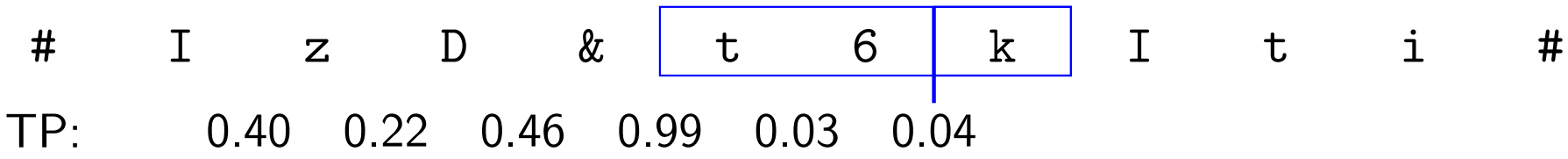
$$TP(\texttt{D\&}, \texttt{t}) = P(\texttt{t}|\texttt{D\&}) = 0.99$$

How to Calculate the Measures



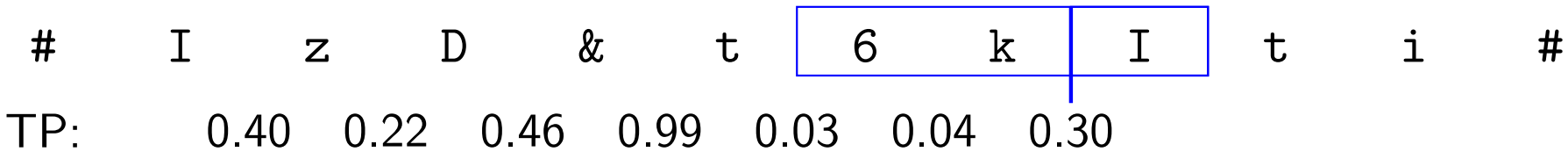
$$TP(\texttt{\&t}, 6) = P(6|\texttt{\&t}) = 0.03$$

How to Calculate the Measures



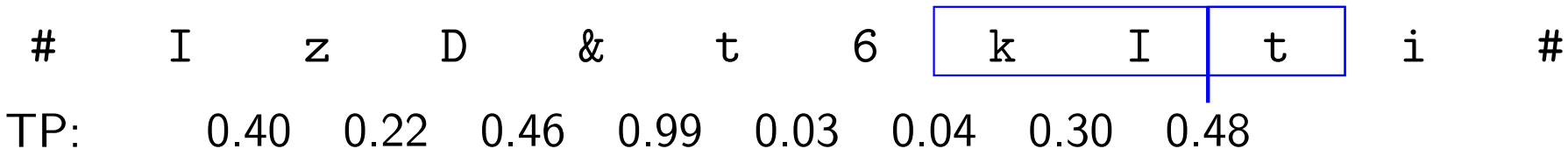
$$TP(\texttt{t6}, \texttt{k}) = P(\texttt{k}|\texttt{t6}) = 0.04$$

How to Calculate the Measures



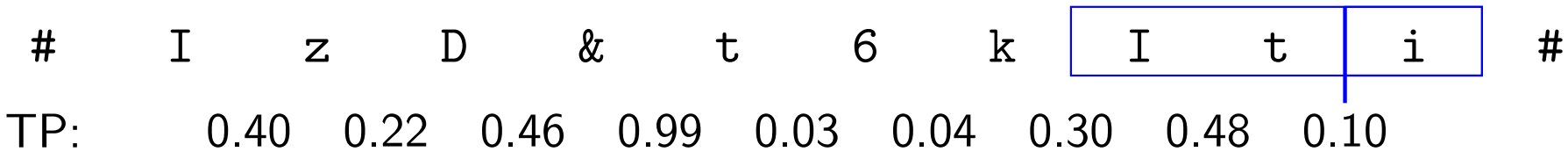
$$TP(\texttt{6k}, \texttt{I}) = P(\texttt{I}|\texttt{6k}) = 0.30$$

How to Calculate the Measures



$$TP(\texttt{kI}, \texttt{t}) = P(\texttt{t}|\texttt{kI}) = 0.48$$

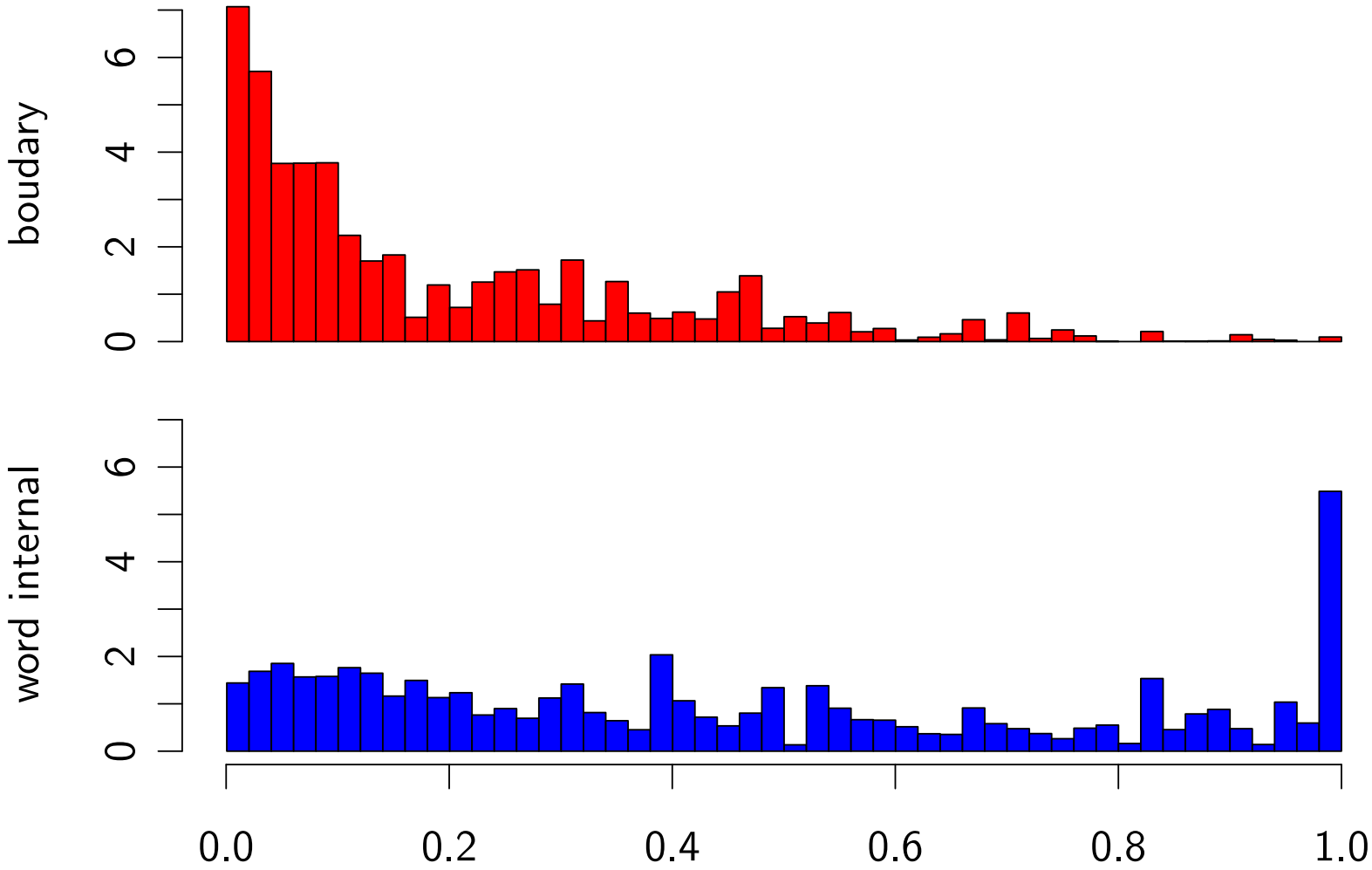
How to Calculate the Measures



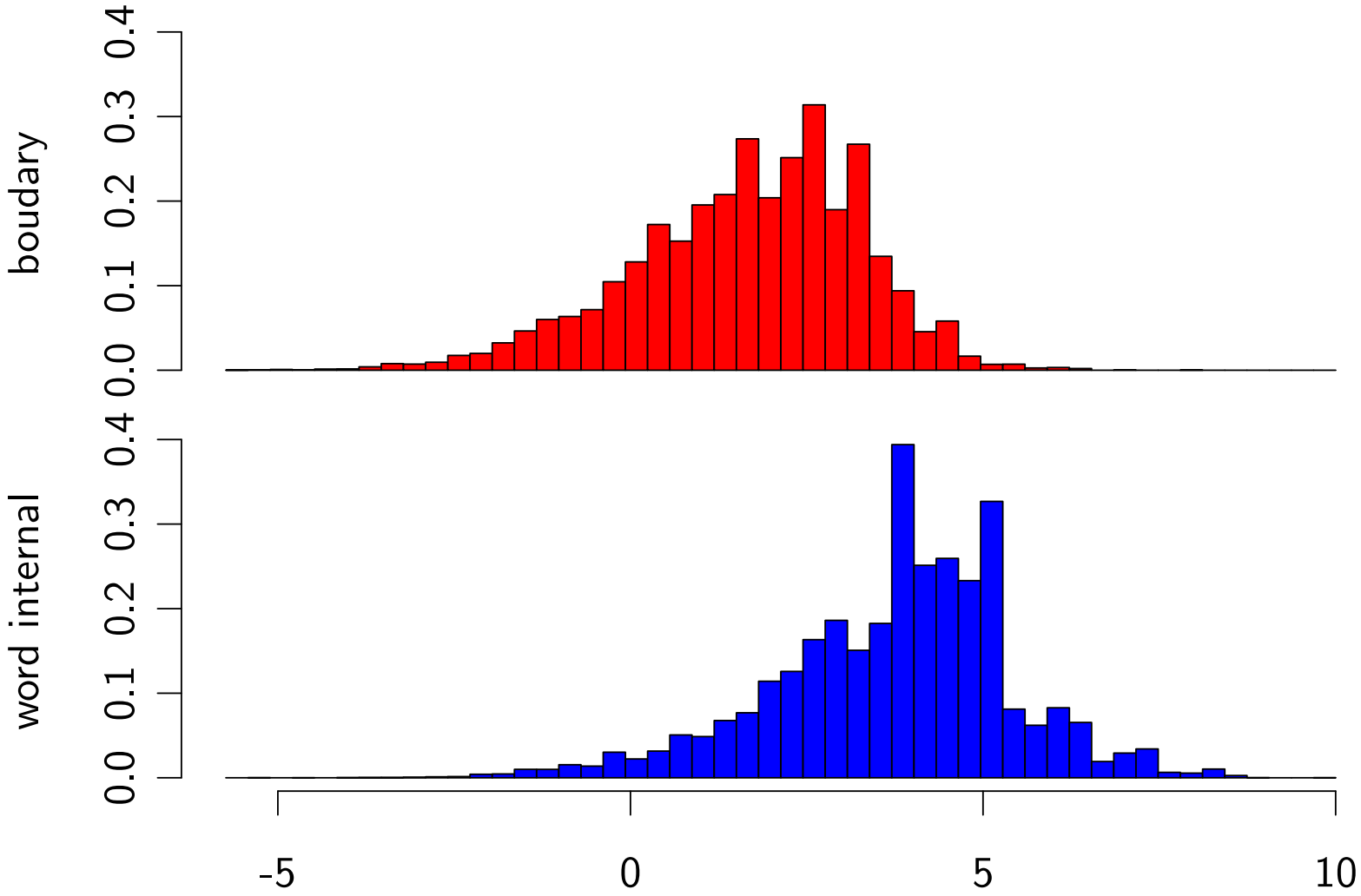
$$TP(\texttt{It}, \texttt{i}) = P(\texttt{i}|\texttt{It}) = 0.10$$

Calculations are done on a corpus of child-directed English

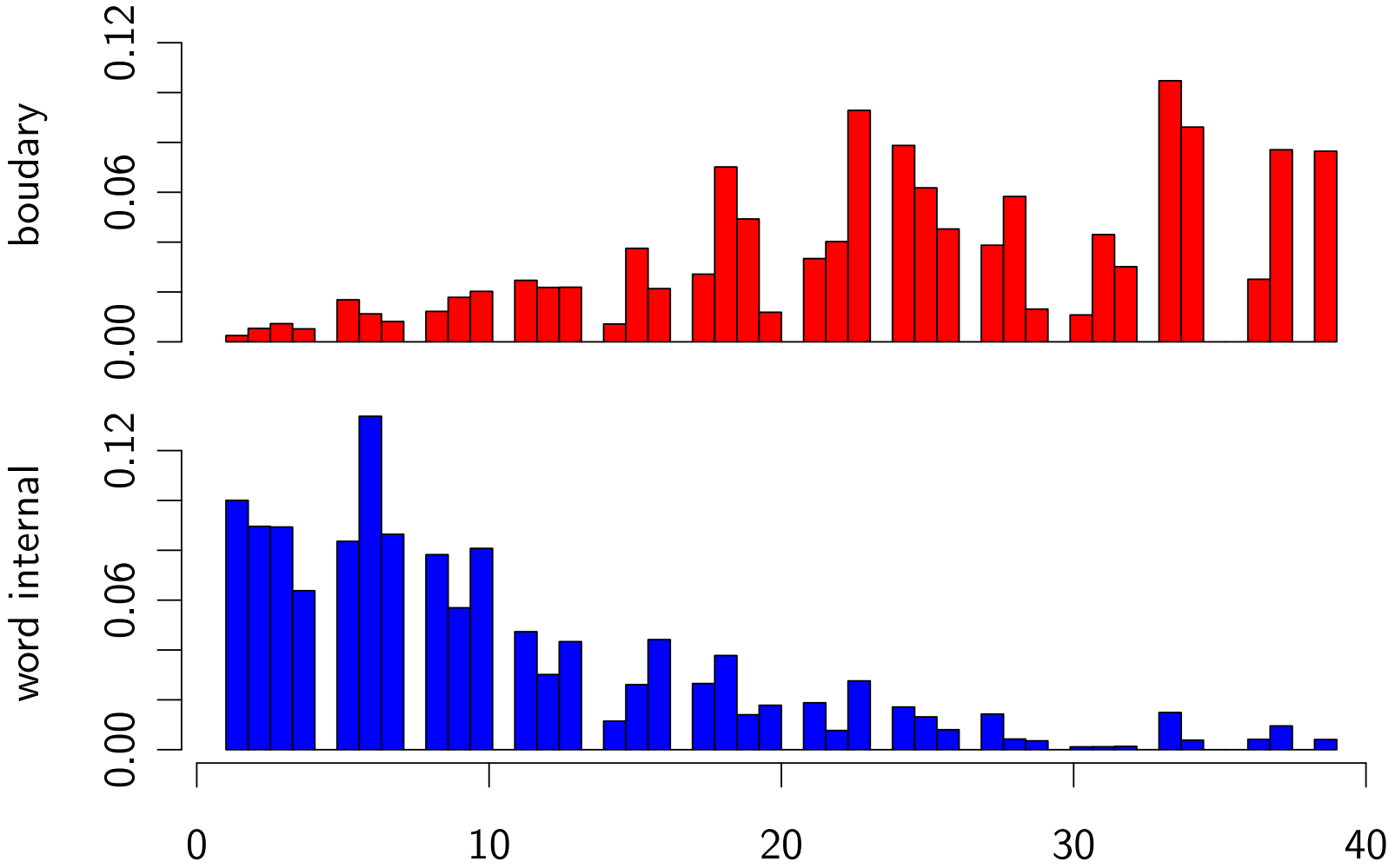
Transitional Probability



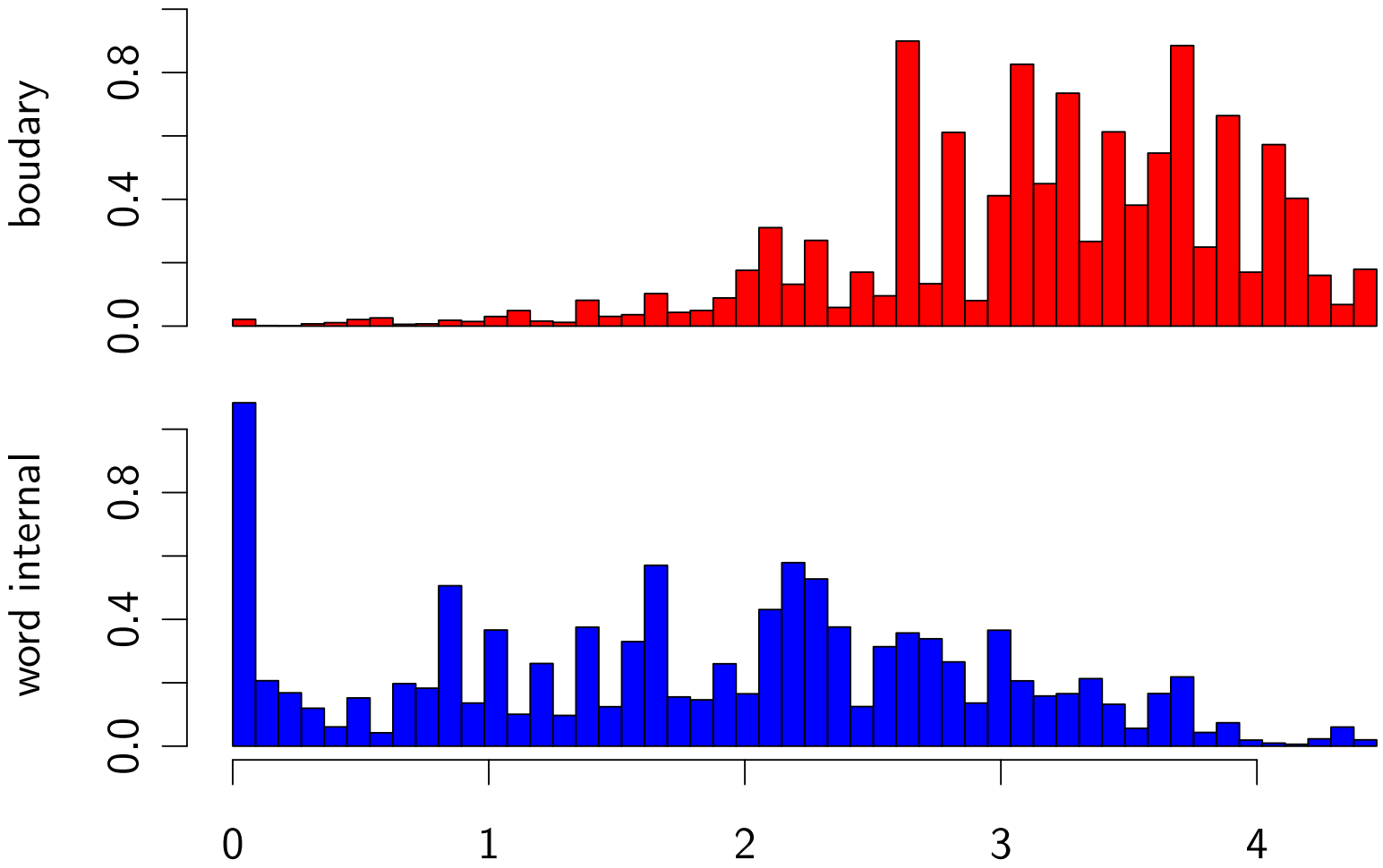
Pointwise Mutual Information



Successor Variety



Entropy



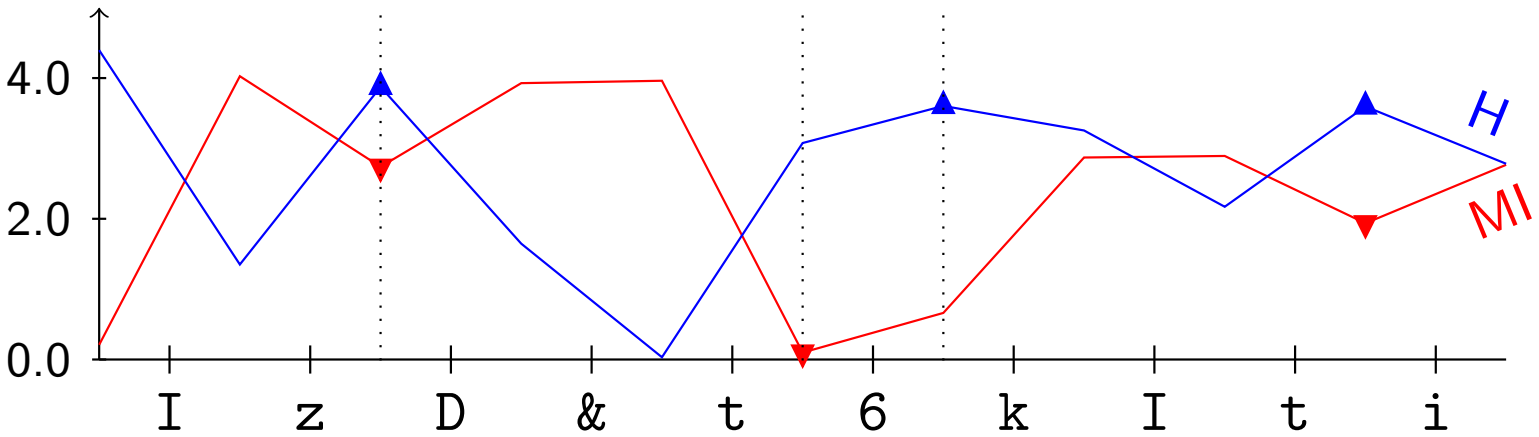
An unsupervised method

- An obvious way to segment the sequence is using a threshold value. However, the choice of threshold is difficult in an unsupervised system.

An unsupervised method

- An obvious way to segment the sequence is using a threshold value. However, the choice of threshold is difficult in an unsupervised system.

A simple unsupervised method: segment at peaks/valleys.



Combining multiple measures: a simple method

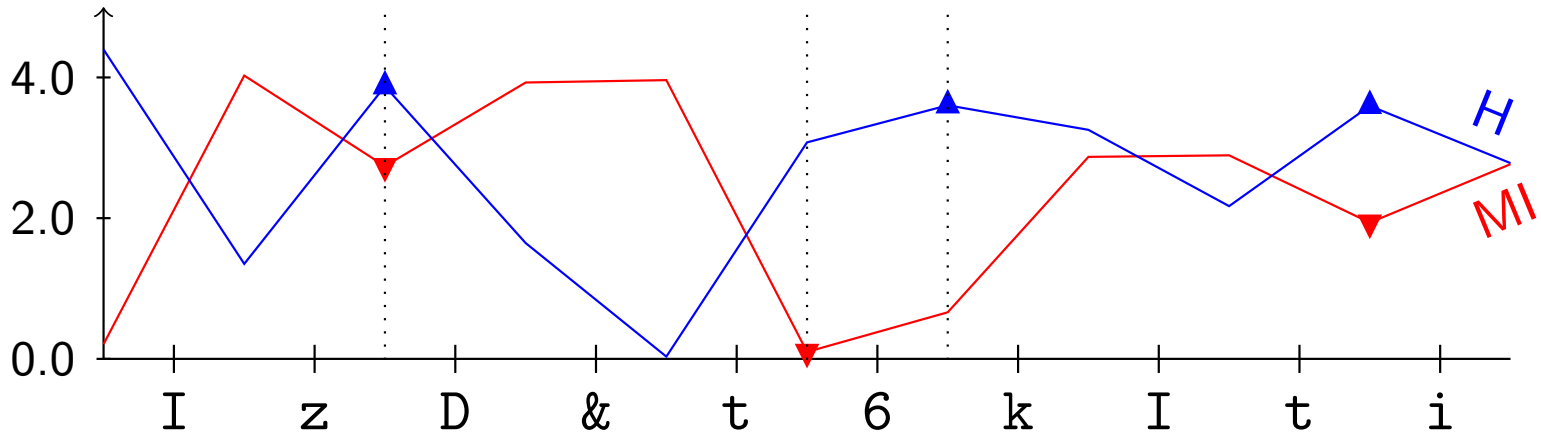
Majority voting (algorithm) works if

1. Votes are cast (relatively) independently.
2. Decisions of the voters are better than random.

Combining multiple measures: a simple method

Majority voting (algorithm) works if

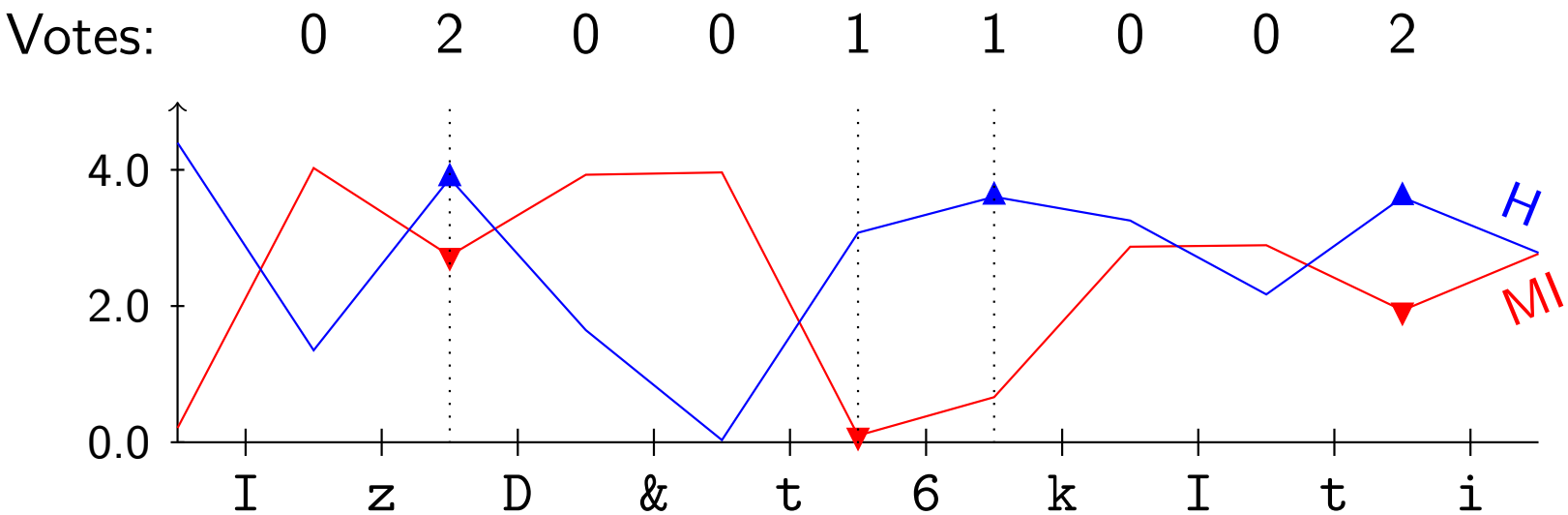
1. Votes are cast (relatively) independently.
2. Decisions of the voters are better than random.



Combining multiple measures: a simple method

Majority voting (algorithm) works if

1. Votes are cast (relatively) independently.
2. Decisions of the voters are better than random.



Putting it all together: a simple algorithm

```
1 foreach utterance do
2     foreach phoneme position in the utterance do
3         Get the majority vote of all measures calculated using
          context sizes one to four;
4         if majority vote is positive then
5             insert a boundary;
6     output the segmented utterance ;
```

Evaluation: boundary, word and lexicon scores

We use standard evaluation scores *precision*, *recall* and *f-score* for evaluation.
In case of segmentation these values can be calculated over,

- *boundaries* (BP, BR, BF),
- word *tokens* (WP, WR, WF),
- word *types* or the lexicon, (LP, LR, LF).

The results

method	BP	BR	BF	WP	WR	WF	LP	LR	LF
Random	27.4	27.4	27.4	12.7	12.7	12.7	6.4	46.4	11.3
Predictability	92.7	76.0	83.5	77.2	67.4	72.0	28.4	65.1	39.5
Baseline	84.2	82.7	83.5	72.1	71.2	71.6	50.7	61.1	55.4

Results are obtained using Algorithm 1 on phonemic transcriptions of child directed speech from Berstain-Ratner corpus.

Segmentation puzzle: a solution

```
ljuuz uibut sjhiu ljuuz
ljuuz tbz ju bhbjo mpwf ljuuz
xibut uibu
ljuuz
ep zpv xbou npsf njml ipofz
ljuuz ljuuz ephhjf
opnj xibu epft b ljuuz tbz
xibu epft b ljuuz tbz
ephhjf eph
ephhjf
opnj xibu epft uif ephhjf tbz
xibu epft uif ephhjf tbz
mjuumf cbcz cjsejf
cbcz cjsejf
zpv epou mjlf uibu pof
plbz npnnz ublf uijt pvu
dpx
uif dpx tbzt npp npp
xibu epft uif dpx tbz opnj
```

Segmentation puzzle: a solution

ljuuz uibut sjhiu ljuuz	ljuuz uibu tsjhiuljuuz
ljuuz tbz ju bhbjo mpwf ljuuz	ljuuz tbz jubhbjompwfljuuz
xibut uibu	xibu tuibu
ljuuz	ljuuz
ep zpv xbou npsf njml ipofz	ep zpvxbounpsfnjmli pof z
ljuuz ljuuz ephhjf	ljuuz ljuuz ephhjf
opnj xibu epft b ljuuz tbz	opnj xibu ep ftb ljuuz tbz
xibu epft b ljuuz tbz	xibu ep ftb ljuuz tbz
ephhjf eph	ephhjf eph
ephhjf	ephhjf
opnj xibu epft uif ephhjf tbz	opnj xibu epft uif ephhjf tbz
xibu epft uif ephhjf tbz	xibu ep ft uif ephhjf tbz
mjuumf cbcz cjsejf	mjuumfcbczcjsejf
cbcz cjsejf	cbczcjsejf
zpv epou mjlf uibu pof	zpv epoumj lf uibu pof
plbz npnnz ublf uijt pvu	plbznpnnzublfui jtpvu
dpx	dpx
uif dpx tbzt npp npp	uif dpx tbz tnppnpp
xibu epft uif dpx tbz opnj	xibu epft uif dpx tbz opnj

Segmentation puzzle: a solution

kitty thats right kitty	kitty that srightkitty
kitty say it again love kitty	kitty say itagainlovekitty
whats that	what sthat
kitty	kitty
do you want more milk honey	do youwantmoremilkh one y
kitty kitty doggie	kitty kitty doggie
nomi what does a kitty say	nomi what do esa kitty say
what does a kitty say	what do esa kitty say
doggie dog	doggie dog
doggie	doggie
nomi what does the doggie say	nomi what does the doggie say
what does the doggie say	what do es the doggie say
little baby birdie	littlebabybirdie
baby birdie	babybirdie
you dont like that one	you dontli ke that one
okay mommy take this out	okaymommytaketh isout
cow	cow
the cow says moo moo	the cow say smoomoo
what does the cow say nomi	what does the cow say nomi

Summary

The segmentation procedure we have just reviewed

- is in line with the psycholinguistic research,
- is completely unsupervised,
- is incremental,
- performs competitive with an alternative state of the art segmentation method.

This is only the part of the solution, we can

- use information from utterance boundaries,
- keep an explicit lexicon and use it for further segmentation,
- make use of acoustic cues,
- use a better algorithm for boundary guessing.

Wrapping up

Simulation, human behaviour and nature

- Does the simulation study help us understand and predict (interesting) human behavior?

Simulation, human behaviour and nature

- Does the simulation study help us understand and predict (interesting) human behavior?
- Would it contribute to nature–nurture debate?

Simulation, human behaviour and nature

- Does the simulation study help us understand and predict (interesting) human behavior?
- Would it contribute to nature–nurture debate?
- If we know one of the positions in the debate is correct, would it help us create a better model?

Simulation, human behaviour and nature

- Does the simulation study help us understand and predict (interesting) human behavior?
- Would it contribute to nature–nurture debate?
- If we know one of the positions in the debate is correct, would it help us create a better model?
  - Clearly we assume some initial knowledge, e.g., phonemes. But, these could be learned in an earlier stage.

Simulation, human behaviour and nature

- Does the simulation study help us understand and predict (interesting) human behavior?
- Would it contribute to nature–nurture debate?
- If we know one of the positions in the debate is correct, would it help us create a better model?
  - Clearly we assume some initial knowledge, e.g., phonemes. But, these could be learned in an earlier stage.
  - If I knew for certain that phonemes are innate, it could have helped.

Simulation, human behaviour and nature

- Does the simulation study help us understand and predict (interesting) human behavior?
- Would it contribute to nature–nurture debate?
- If we know one of the positions in the debate is correct, would it help us create a better model?
  - Clearly we assume some initial knowledge, e.g., phonemes. But, these could be learned in an earlier stage.
  - If I knew for certain that phonemes are innate, it could have helped.
  - If I knew for certain that phonemes weren't innate, it may motivate me to study how it is learned.

Simulation, human behaviour and nature

- Does the simulation study help us understand and predict (interesting) human behavior?
- Would it contribute to nature–nurture debate?
- If we know one of the positions in the debate is correct, would it help us create a better model?
  - Clearly we assume some initial knowledge, e.g., phonemes. But, these could be learned in an earlier stage.
  - If I knew for certain that phonemes are innate, it could have helped.
  - If I knew for certain that phonemes weren't innate, it may motivate me to study how it is learned.
  - But does it matter if this knowledge is language specific or not?