The input to language acquisition: Statistical analysis of infant-directed speech in Hungarian and Italian

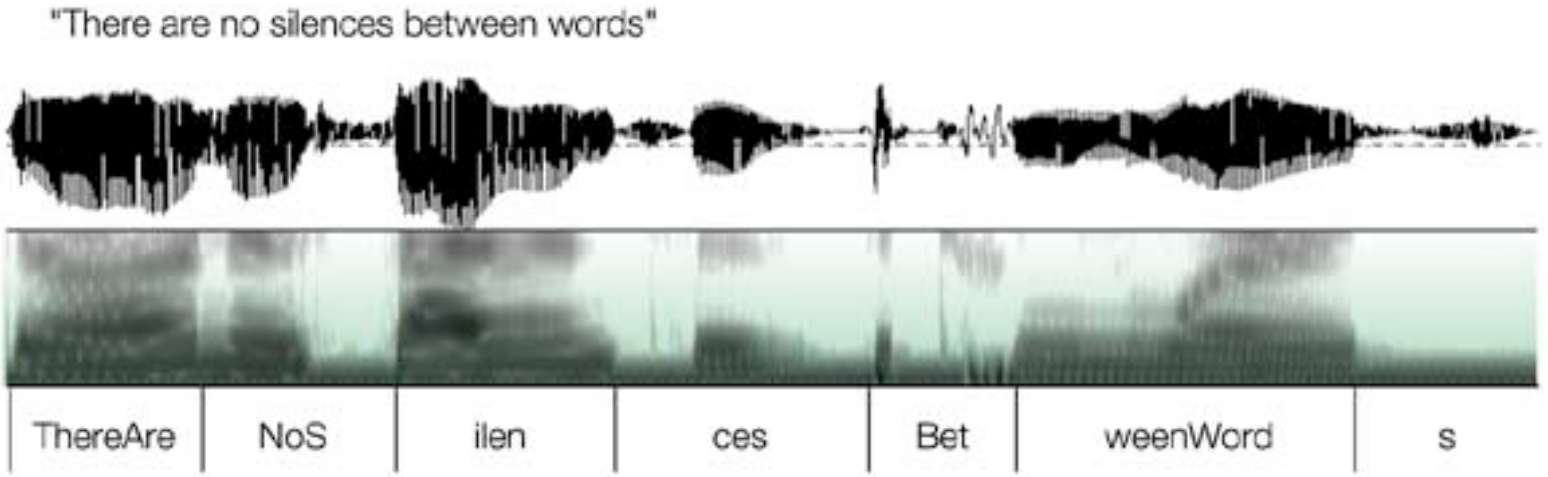
JUDIT GERVAIN
University of British Columbia
gervain@psych.ubc.ca

Statistical Structure in Language

- units in language have characteristic statistical distributions (Shannon 1948, Harris 1955):
  - probability:
    - e > *t* > *r* > *z [in English]*
    - *the* > *that* > *apartment* > *helicopter* > *rhododendron* …
  - conditional probability:
    - *q*… [*u* > e]
    - *fish and … [chips > potatoes] , a question of life and … [death > passing, termination]*
    - *He will … [V > N]*

Statistics in Language Acquisition

- speech is continuous, word boundaries are not systematically indicated by pauses



- learning problem: segmentation of word forms

Statistics in Language Acquisition

- infants, adults and some animal species are sensitive to the statistical structure of the input (Saffran et al. 96, Hauser et al. 02, Toro & Trobalon 05):

a barna maci

'the brown teddybear'

Statistics in the Input?

- is the statistical information there in the input to support segmentation?
  - early works: reasonable evidence over small corpora (Harris 55)
  - recent works: mixed evidence over larger corpora
    - moderate accuracy in English & Dutch (Brent & Cartwright 96, Swingley 05)
    - low accuracy in Spanish, Japanese, Chinese (Batchelder 02, Yang 04)

Statistics in the Input?

- does language input provide sufficient statistical information for word segmentation?
- how does statistical segmentation work in typologically different languages?

> statistical analysis of Hungarian & Italian infant-directed corpora

Corpora

Hungarian

- infant-directed adult speech in the infants' day care environment

- 15 231 utterances
- 54 881 word tokens
- 8234 word types

Italian

- infant-directed adult speech in the infants' institutional & home environment

- 10 473 utterances
- 51 138 word tokens
- 4525 word types

Analysis Procedure: Basic Steps

0. corpus (target): *ez a barna maci*
1. units *ez a bar na ma ci*

2. computing statistics: ez .36 a .00 bar .92 na .01 ma .21 ci

3. segmentation (e.g. threshold = .1): ez_a | bar_na | ma_ci

   ✗ ✓ ✓ ✓ ✓

4. evaluation: eza barna maci

Analysis Procedure: Basic Steps

0.	corpus (target):	ez *a barna maci*
1.	units	ez *a bar na ma ci*
2.	computing statistics:	ez .36 a .00 bar .92 na .01 ma .21 ci
3.	segmentation (e.g. threshold = .1):	ez_a | bar_na | ma_ci ✗ ✓ ✓ ✓ ✓
4.	evaluation:	eza barna maci

Analysis Procedure: Stats

- transition probabilities (TP):
  - forward:

$$FW\ TP\ (A \rightarrow B) = F(AB) / F(A)$$

  - backward:

$$BW\ TP\ (A \leftarrow B) = F(AB) / F(B)$$

Analysis Procedure: Basic Steps

0.	corpus (target):	ez *a barna maci*
1.	units	ez *a bar na ma ci*
2.	computing statistics:	ez .36 a .00 bar .92 na .01 ma .21 ci
3.	segmentation (e.g. threshold = .1):	ez_a | bar_na | ma_ci ✗ ✓ ✓ ✓ ✓
4.	evaluation:	eza barna maci

Analysis Procedure: Segmentation

- absolute threshold:
  - word boundary: value ≤ threshold
  - word internal: value > threshold

  1st, 2nd … 100th percentile of FW & BW distribution as threshold
- relative threshold:
  - word boundary: value1 > value2 < value3
  - word internal: otherwise

Analysis Procedure: Basic Steps

0.	corpus (target):	ez *a barna maci*
1.	units	ez *a bar na ma ci*
2.	computing statistics:	ez .36 a .00 bar .92 na .01 ma .21 ci
3.	segmentation (e.g. threshold = .1):	ez_a | bar_na | ma_ci
		✗ ✓ ✓ ✓ ✓
4.	evaluation:	eza barna maci

Analysis Procedure: Evaluation

- accuracy (precision):

  accuracy = hits / (hits + false alarms)

- completeness (recall):

  completeness = hits / (hits + misses)

where:

  – hit: correct identification of word boundary (WB)

  – false alarm: positing a WB where there is none

  – miss: not positing a WB where there is one

Analysis Procedure: Evaluation

- trade-off between accuracy & completeness:
  - conservative strategy ("as few WBs as possible"): high accuracy, low completeness
  - daring strategy ("as many WBs as possible"): low accuracy, high completeness

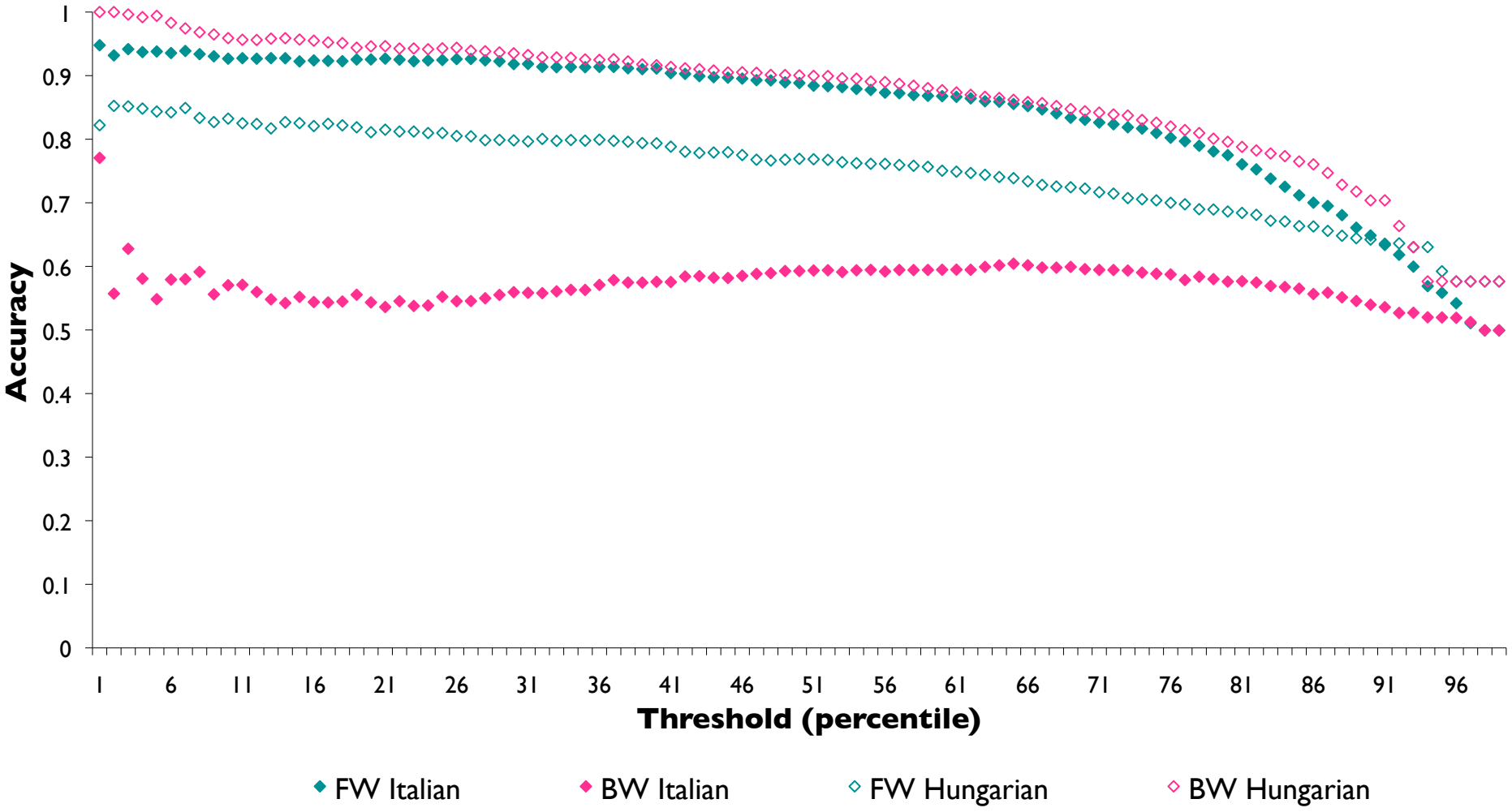
Experiments: Overview

ITALIAN HUNGARIAN (syllabified)	Absolute Threshold	Relative Threshold
FW	EXP 1	EXP 2
BW	accuracy completeness	accuracy completeness

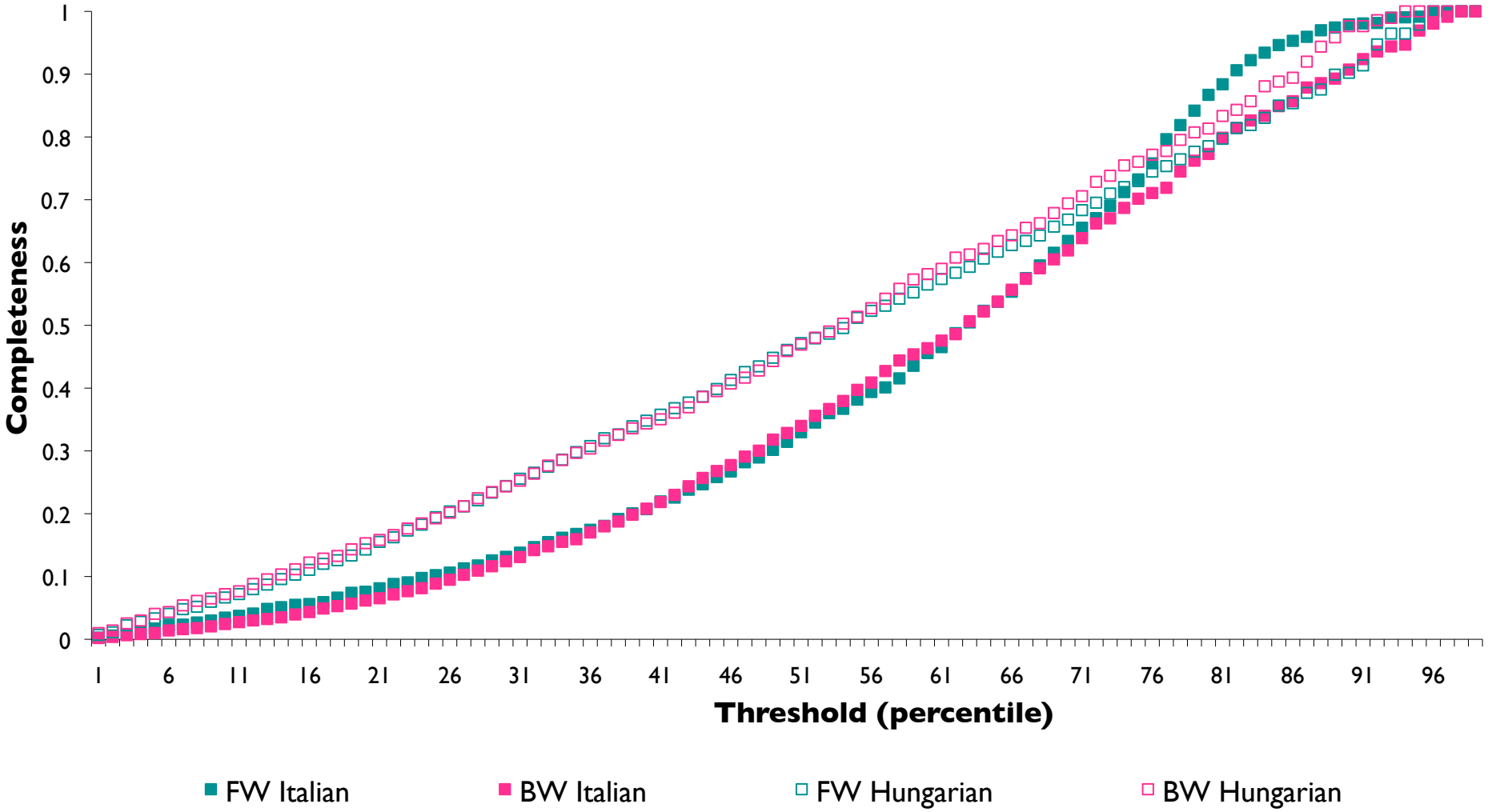
Experiments: Overview

HUNGARIAN (phonemicized)	Absolute Threshold
FW	EXP 3
BW	accuracy completeness

Experiment 1: Accuracy



Experiment 1: Completeness



Experiment 1: Qualitative Results

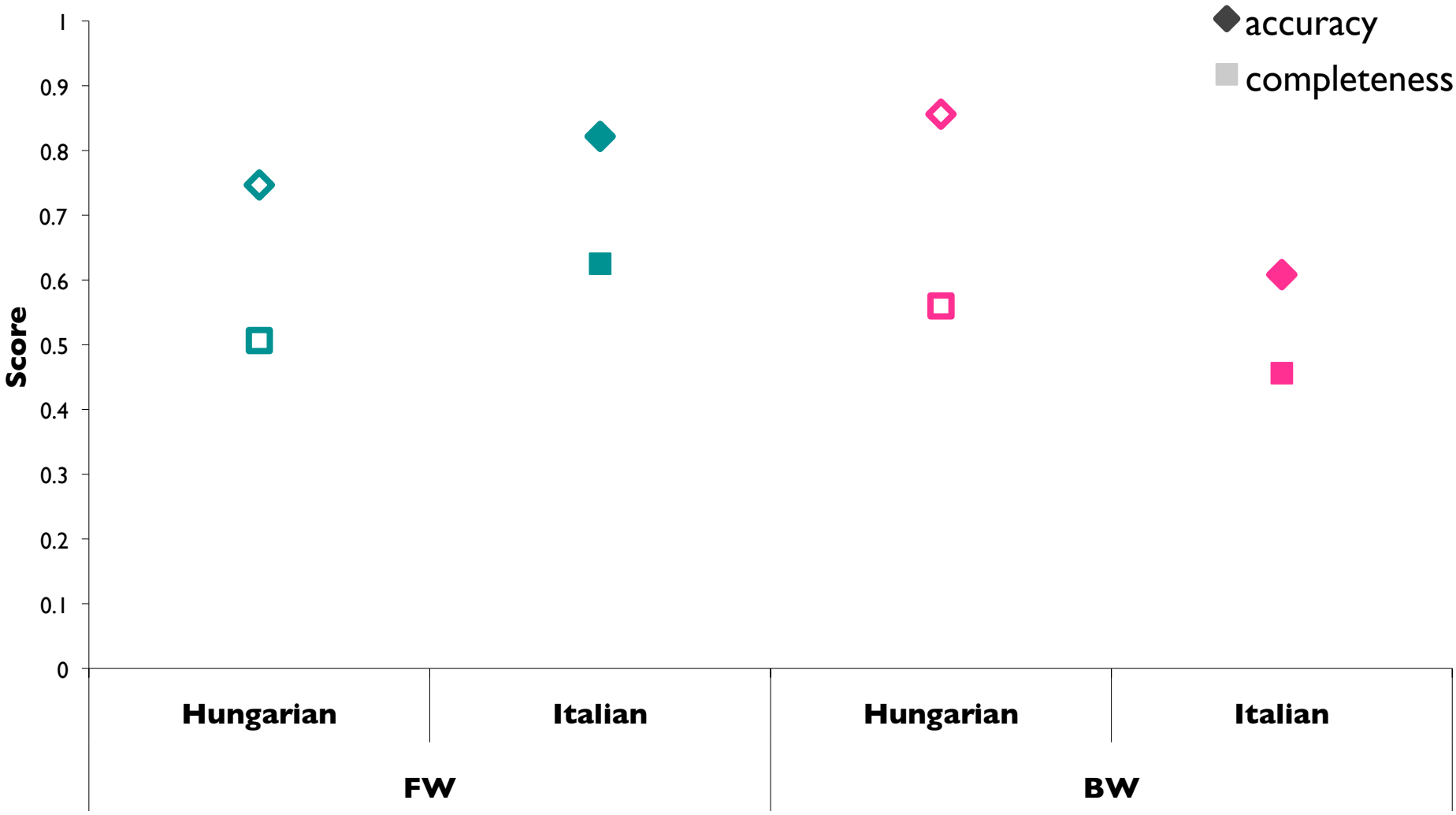
lowest threshold(s)

- undersegmentation
- words extracted: only a few
  - *a* 'the'
  - *ki* (i) 'who', (ii) 'out'
  - *mi* (i) 'what', (ii) us
  - *nem* 'no'
- BUT: oversegmentation starts to appear, too
  - *cso|ki* (target: *csoki* 'chocolate')

highest threshold(s)

- oversegmentation
- words extracted: many
- many words oversegmented
  - *i|de* (target *ide* 'here'; cf. *de* 'but')
  - *sé|tál* (target *sétál* 'walk'; cf. *tál* 'plate')

Experiment 2



Experiment 2: Qualitative Results

- balance between accuracy & completeness
- mostly undersegmentation errors: in frequently co-occurring words
  - *miez* (target: *mi ez* 'what (is) this')
  - *holvan* (target: *hol van* 'where is')
- oversegmentation: for frequent syllables
  - *mó|ni* (target: *móni* [name], cf. *–ni* inf. suffix)
  - *gör|be* (target: *görbe* 'crooked', cf. *be* 'in')

Discussion

- the cross-linguistic difference is related to the morphosyntactic types of the two languages
  - boundary posited if TP between A & B is low
  - when is the TP, a fraction, low?

    FW (AB) = F(AB) / F(A) BW (AB) = F(AB) / F(B)
  - FW is low if AB is infrequent, A is frequent:

    [A B]: typical in functor-initial, i.e. VO, non-agglut. languages
  - BW is low if AB is infrequent, B is frequent

    [A B]: typical in functor-final, i.e. OV, agglutinating languages

Discussion

- what is segmented out in Hungarian: multimorphemic words or single morphemes?

házainkban or *ház|a|i|nk|ban*

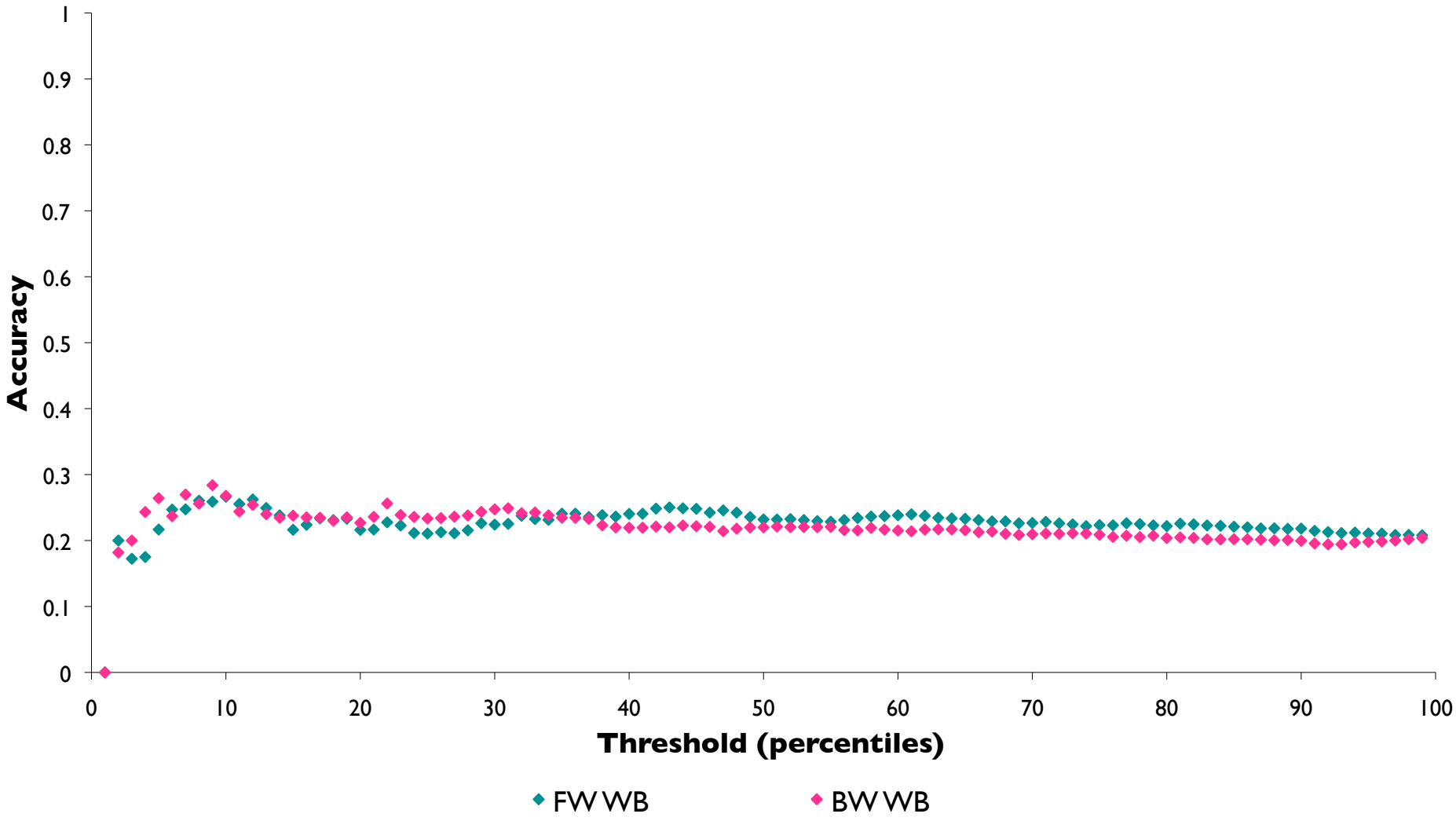
'in our houses'

- BUT resyllabification:

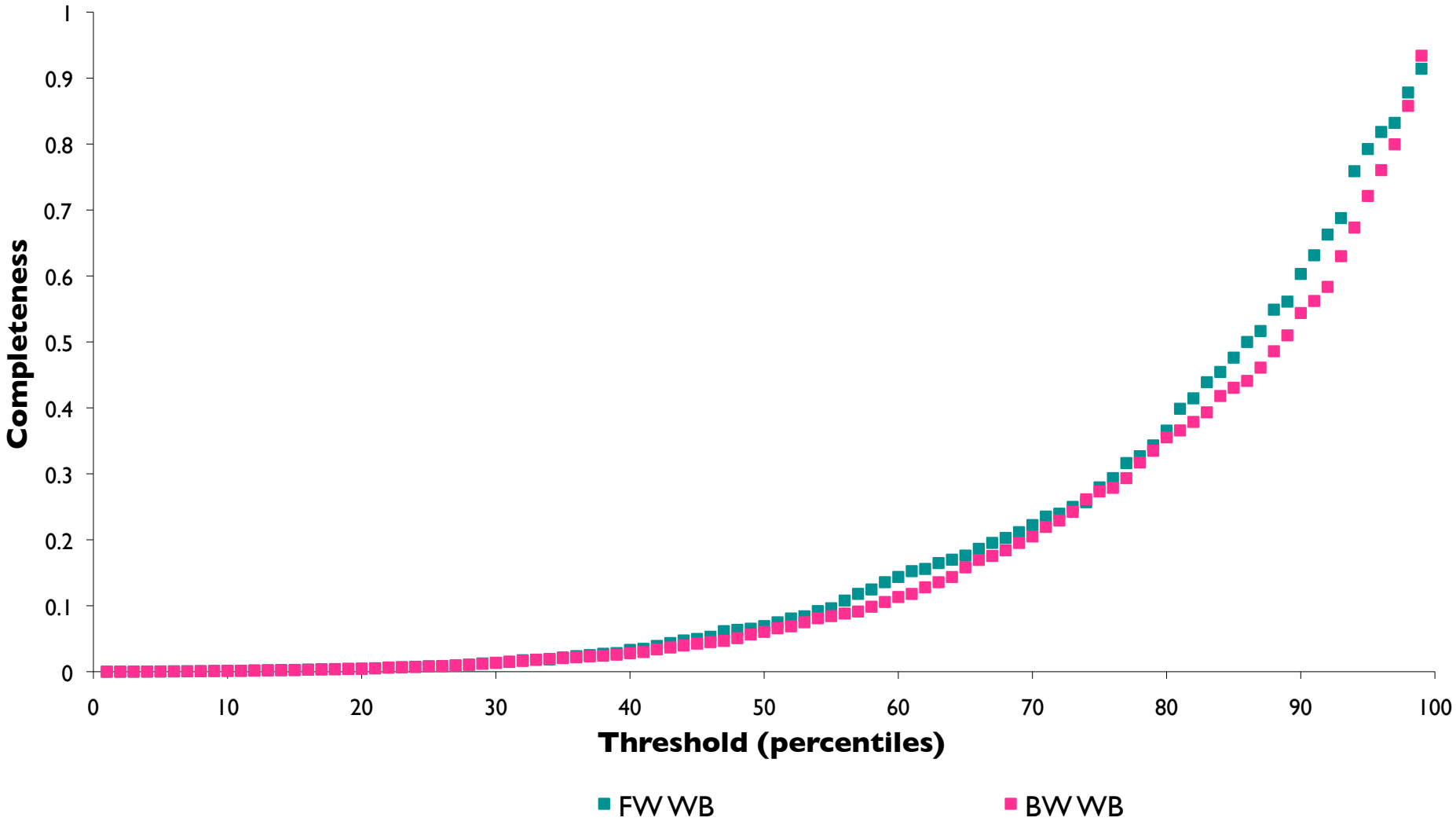
há-za-ink-ban

- let's try the phoneme as a unit!

Experiment 3: Accuracy



Experiment 3: Completeness



Discussion

- the unit of representation plays a crucial role in segmentation
- phonemes are not optimal, because they can encode very little information
  - the same phoneme pair spans word boundaries in some contexts and falls within a word in others

General Discussion

- statistical segmentation is only as good as the unit of representation chosen
- syllables result in better segmentation:
  - they encode more information (~20-40 phonemes in a language, but ~300-5000 different syllables)
  - some syllables coincide with words
  - these are typically the most frequent words, i.e. functors

General Discussion

- morphological boundaries can be segmented to the extent that they coincide with syllable boundaries
- this is insufficient, yet Hungarian infants acquire morphology relatively early, showing evidence of segmentation
- other cues are needed: vowel harmony, fixed word-level stress etc.

General Discussion

- the sentential & phrasal position of functors correlates with general morphosyntactic properties (Dryer 92)
- consequently, the preferred direction of segmentation also correlates with morphosyntactic type
  - BW: in heavily suffixing and/or OV languages
  - FW: in non-suffixing and/or VO languages

Open Questions: Acquisition

- unit of representation:
  - newborns and young infants tend to represent speech as a sequence of syllables, not phonemes (Mehler et al. 81)
- computing conditional probabilities:
  - there is evidence that infants can compute FW TPs (Saffran et al. 96)
  - it is unknown whether infants (adults and/or animals) can compute BW TPs

thankyouthankyouthankyouthan
kyouthankyouthankyouthankyou
thankyouthankyouthankyouthan
kyouthankyouthankyouthankyou
thankyouthankyouthankyouthan
kyouthankyouthankyouthankyou

thankyouthankyouthankyouthan
kyouthankyouthankyouthankyou
thankyouthankyouthankyouthan
kyouthankyouthankyouthankyou
thankyouthankyouthankyouthan
kyouthankyouthankyouthankyou