Testing the robustness of online word segmentation: Effects of linguistic diversity and phonetic variation

Luc Boruta[1,2], **Sharon Peperkamp**[2], **Benoît Crabbé**[1] & **Emmanuel Dupoux**[2]

luc.boruta@inria.fr

[1]ALPAGE, Univ. Paris 7 & INRIA
[2]LSCP–DEC, EHESS, ENS & CNRS

CMCL — June 23, 2011

Yet another study on word segmentation...

What this work is not about

- New models of word segmentation.

What this work is about

- The acquisition of **word segmentation**;
- The acquisition of **phonological knowledge**;
- Interactions between the two.

Yet another study on word segmentation...

What this work is not about

- New models of word segmentation.

What this work is about

- The acquisition of **word segmentation**;
- The acquisition of **phonological knowledge**;
- Interactions between the two.

Yet another study on word segmentation...

What this work is not about

- New models of word segmentation.

What this work is about

- The acquisition of **word segmentation**;
- The acquisition of **phonological knowledge**;
- Interactions between the two.

Word segmentation vs. allophonic rules

French devoicing allophonic rule

$$/\mathbf{r}/ \rightarrow \begin{cases} [\mathbf{\chi}] & \text{before a voiceless consonant} \\ [\mathbf{ʁ}] & \text{otherwise} \end{cases}$$

Consequence

$$/\text{kana}\mathbf{r}/ \rightarrow \begin{cases} [\text{kana}\mathbf{\chi}\text{␣flotɑ̃}], \textit{ canard flottant} \\ [\text{kana}\mathbf{ʁ}\text{␣ʒon}], \textit{ canard jaune} \end{cases}$$

Word segmentation vs. allophonic rules

French devoicing allophonic rule

$$/\mathbf{r}/ \rightarrow \begin{cases} [\mathbf{\chi}] & \text{before a voiceless consonant} \\ [\mathbf{ʁ}] & \text{otherwise} \end{cases}$$

Consequence

$$/\text{kana}\mathbf{r}/ \rightarrow \begin{cases} [\text{kana}\mathbf{\chi}\text{␣flotɑ̃}], \textit{ canard flottant} \\ [\text{kana}\mathbf{ʁ}\text{␣ʒon}], \textit{ canard jaune} \end{cases}$$

Word segmentation

The task

- Input: /əwʊdʧʌkwʊdʧʌkwʊd/
- Output: /ə␣wʊdʧʌk␣wʊd␣ʧʌk␣wʊd/

Phonemic transcripts = idealized input

- Models are typically evaluated using **phonemic** transcripts;
- Assumption: kids know how to undo **allophony**/coarticulation.

Word segmentation

The task

- Input: /əwʊdʧʌkwʊdʧʌkwʊd/
- Output: /ə␣wʊdʧʌk␣wʊd␣ʧʌk␣wʊd/

Phonemic transcripts = idealized input

- Models are typically evaluated using **phonemic** transcripts;
- Assumption: kids know how to undo **allophony**/coarticulation.

Related work

Rytting, Brew & Fosler-Lussier (2010)

- Input unit: probability vector over a finite set of symbols;
- Symbols: limited to the phonemic inventory.

Daland & Pierrehumbert (2010)

- Input: phonemic transcripts, conversational reduction processes;
- Reduction processes: implemented by hand;
- Transcripts: adult-directed speech.

Related work

Rytting, Brew & Fosler-Lussier (2010)

- Input unit: probability vector over a finite set of symbols;
- Symbols: limited to the phonemic inventory.

Daland & Pierrehumbert (2010)

- Input: phonemic transcripts, conversational reduction processes;
- Reduction processes: implemented by hand;
- Transcripts: adult-directed speech.

Which segmentation models?

Desirable properties

[Brent, 1999; Gambell & Yang, 2004]

- Start without any knowledge specific to a particular language;
- Learn in an unsupervised manner and operate incrementally.

Which segmentation models?

- MDBP-1: Brent, 1999;
- NGS-u: Venkataraman, 2001;
- Two random baselines.

Which segmentation models?

Desirable properties

[Brent, 1999; Gambell & Yang, 2004]

- Start without any knowledge specific to a particular language;
- Learn in an unsupervised manner and operate incrementally.

Which segmentation models?

- MDBP-1: Brent, 1999;
- NGS-u: Venkataraman, 2001;
- Two random baselines.

Evaluation

Now-standard evaluation protocol

[Brent, 1999; Goldwater et al., 2009]

- Gold standard: orthographic segmentation;
- Precision, recall and **F-score** on the word segmentation;
- Precision, recall and **F-score** on the induced lexicon.

	Lexicon	Segmentation
ə␣wʊdʧʌk␣wʊd␣ʧʌk␣wʊd	✓	✓
ə␣wʊd␣ʧʌk␣wʊdʧʌk␣wʊd	✓	✗

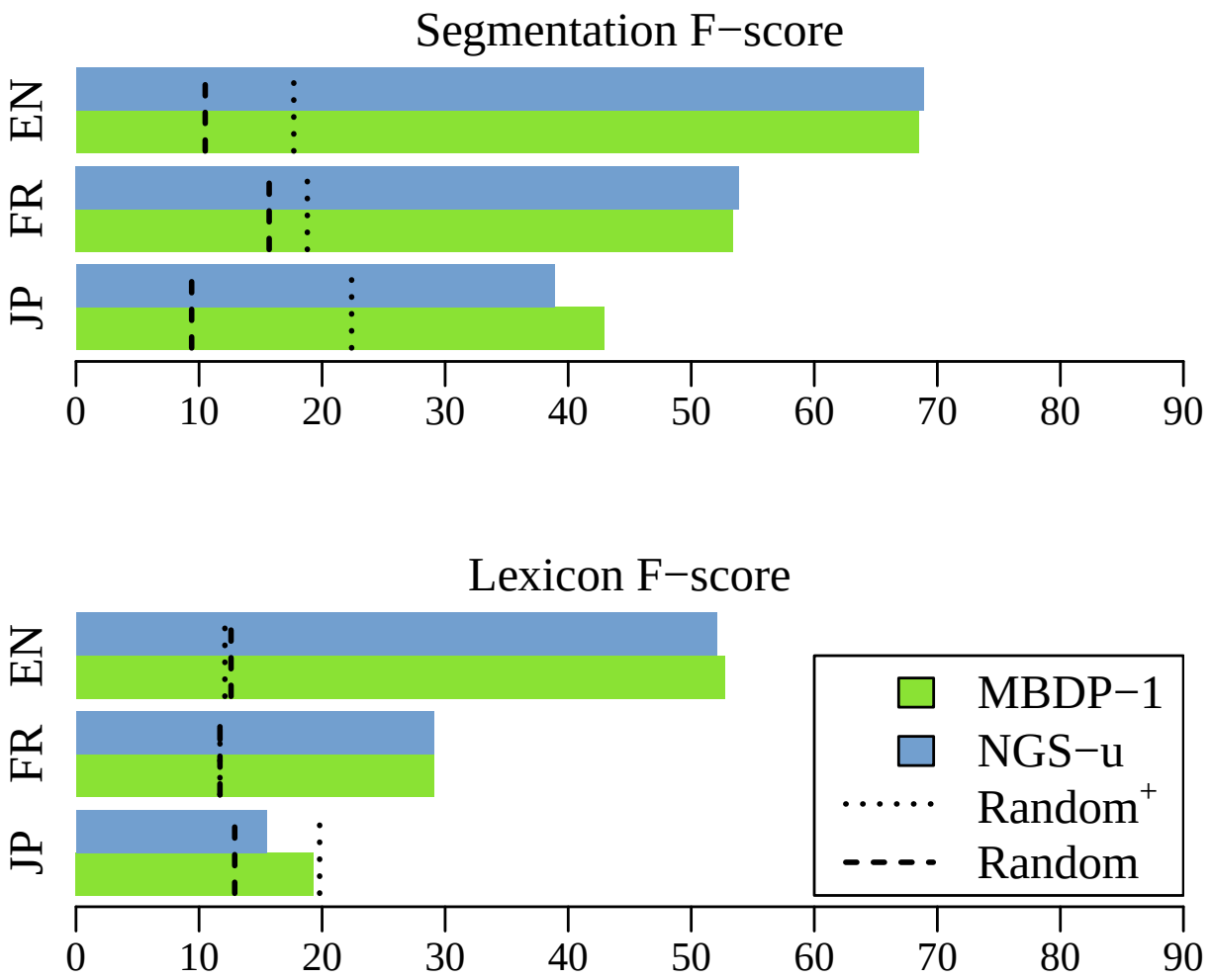
Experimental setup

CHILDES corpora of child-directed speech [MacWhinney, 2000]

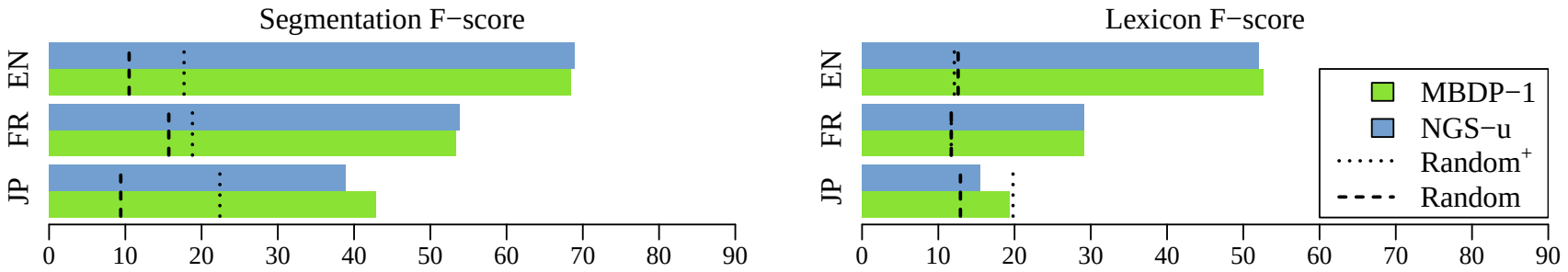
- Derived from transcribed adult-child verbal interactions;
- Phonemic transcriptions, orthographic segmentation.

	English	French	Japanese
Utterance tokens	10k	10k	10k
Word tokens	33k	51k	27k
Phoneme tokens	96k	121k	103k
Phoneme types	50	35	49

Cross-linguistic evaluation on phonemic corpora



Cross-linguistic evaluation on phonemic corpora



- Blame it on the data?
- Rich morphology (e.g. French clitics)? Hapax rate?
- Relative importance of different cues?

Effects of phonetic variation

Phonemic transcripts = idealized input

- Models are typically evaluated using **phonemic** transcripts;
- Assumption: kids know how to undo **allophony**/coarticulation.

Corpora and allophonic rules

- No phonetic transcripts of child-directed speech are available;
- How many allophones do infants have to learn?
- Where is the limit between allophony and mere coarticulation?

Effects of phonetic variation

Phonemic transcripts = idealized input

- Models are typically evaluated using **phonemic** transcripts;
- Assumption: kids know how to undo **allophony**/coarticulation.

Corpora and allophonic rules

- No phonetic transcripts of child-directed speech are available;
- How many allophones do infants have to learn?
- Where is the limit between allophony and mere coarticulation?

Experimental setup

Emulating rich phonetic transcriptions [Boruta, 2011a]

- Apply **artificial allophonic rules** to phonemic corpora;
- Benchmark models using different **allophonic complexities**;
- Control the size of the allophonic grammar.

Simplifying assumptions [LeCalvez, 2007; Boruta, 2011a]

- We only model monolateral rules: $p \rightarrow a \; / \; \_ \; c$
- No two rules introduce the same phone: $[\mathrm{r}]^{/\mathrm{t}/}$ and $[\mathrm{r}]^{/\mathrm{d}/}$

Experimental setup

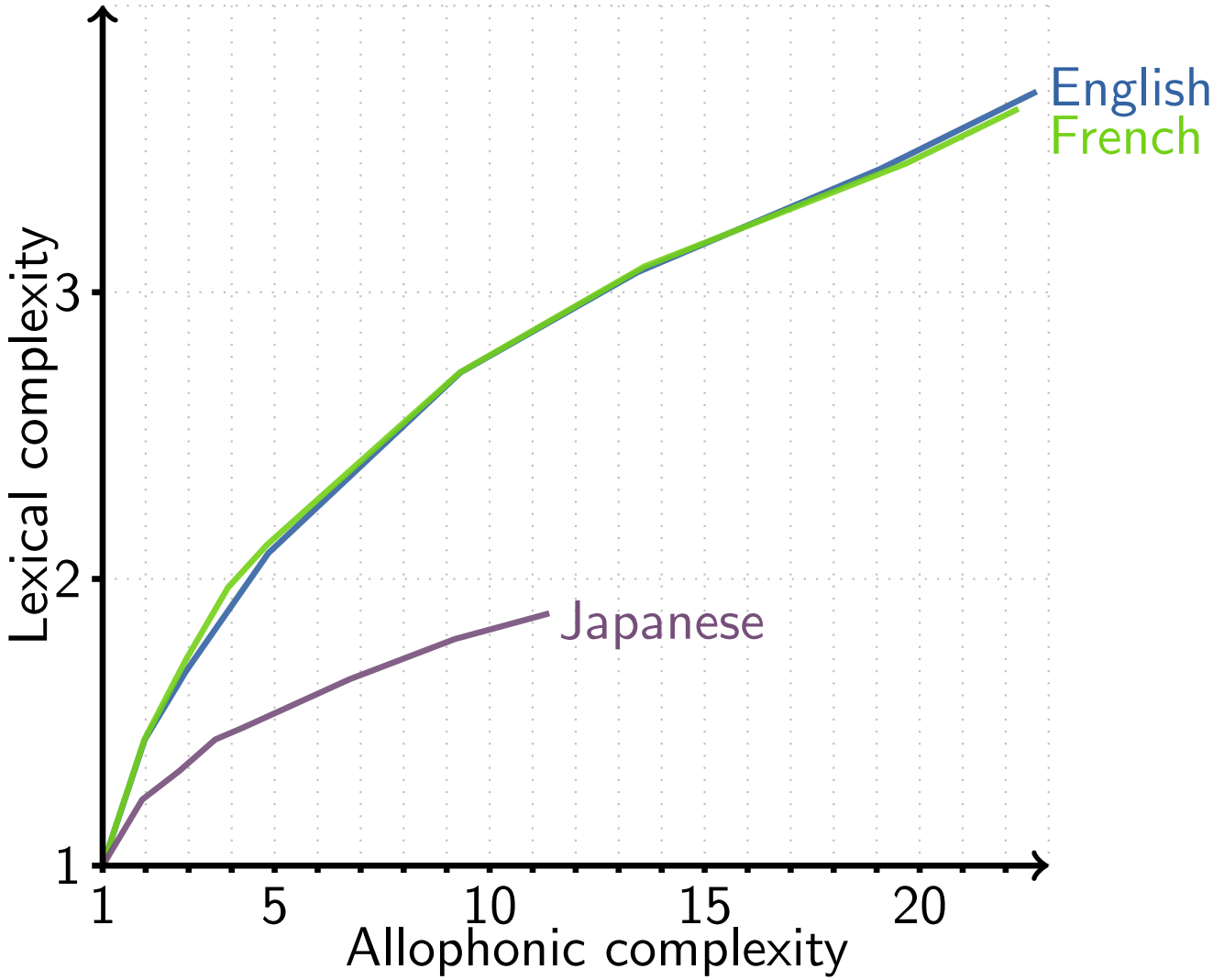
Emulating rich phonetic transcriptions [Boruta, 2011a]

- Apply **artificial allophonic rules** to phonemic corpora;
- Benchmark models using different **allophonic complexities**;
- Control the size of the allophonic grammar.

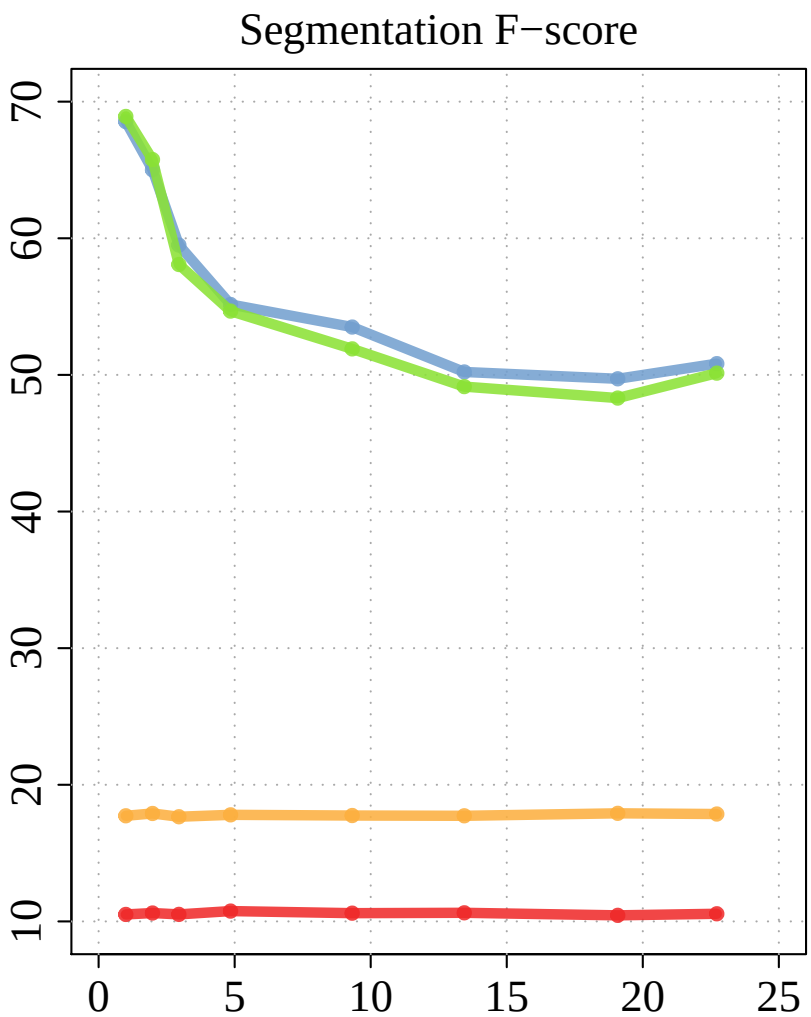
Simplifying assumptions [LeCalvez, 2007; Boruta, 2011a]

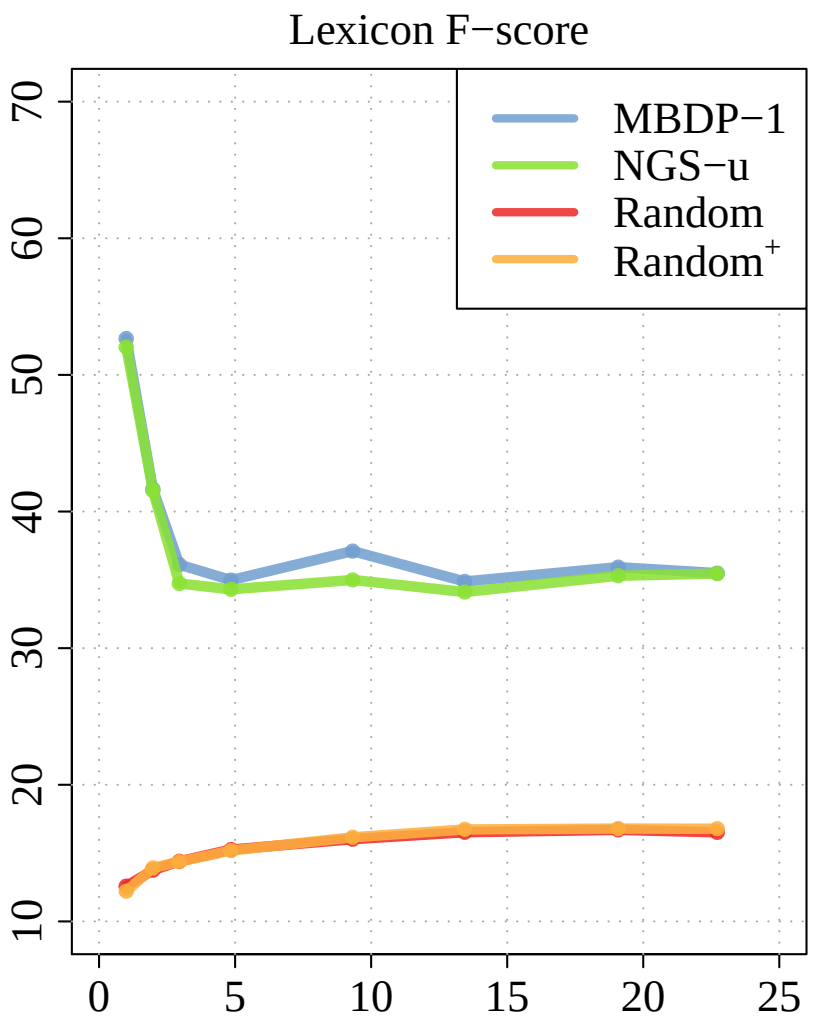
- We only model monolateral rules: $p \rightarrow a \ / \ \_ \ c$
- No two rules introduce the same phone: $[\mathrm{r}]^{/\mathrm{t}/}$ and $[\mathrm{r}]^{/\mathrm{d}/}$

Lexical complexity $\propto$ allophonic complexity

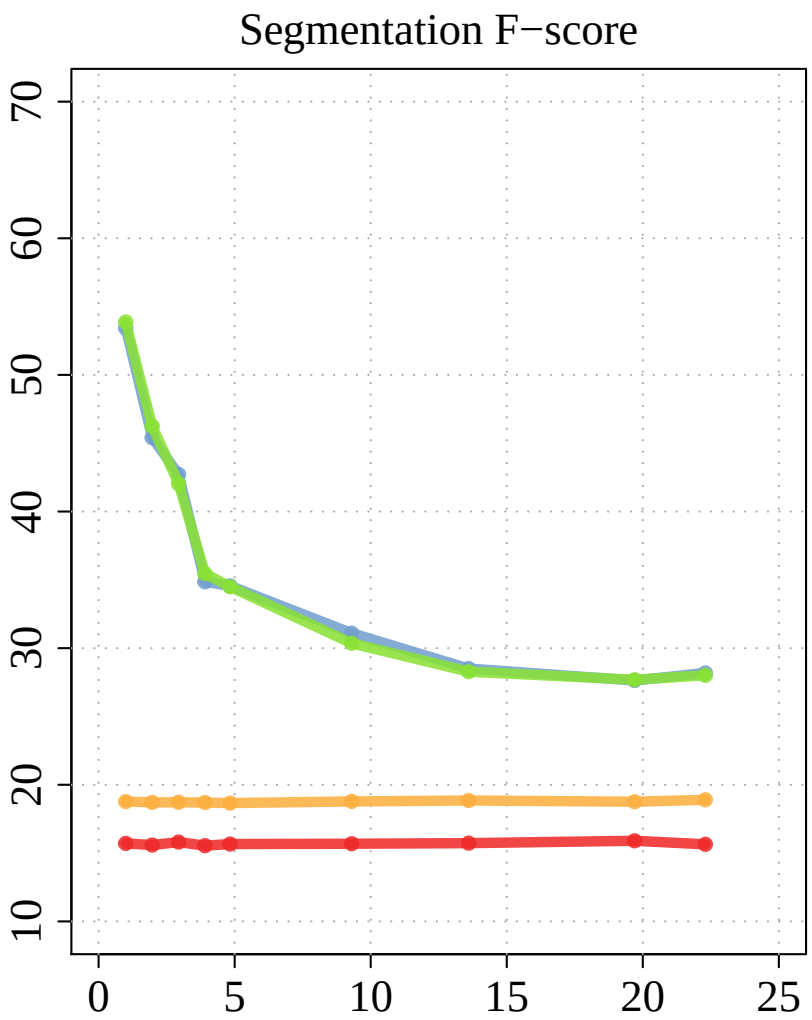


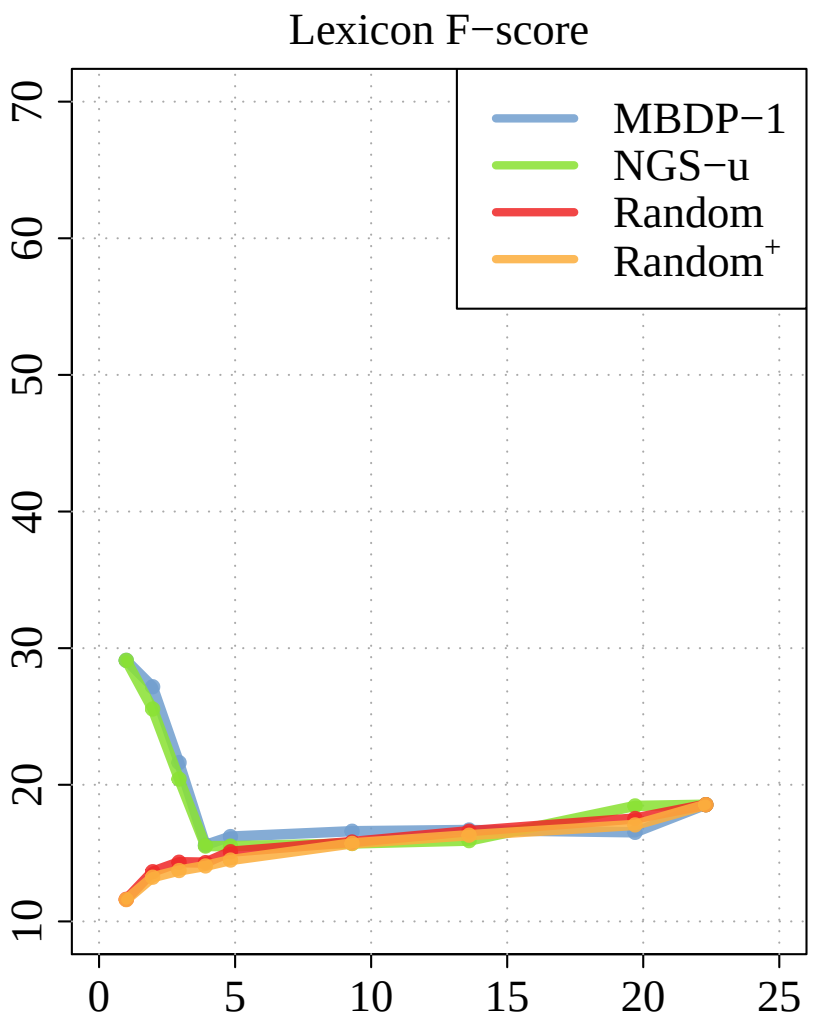
Results: English



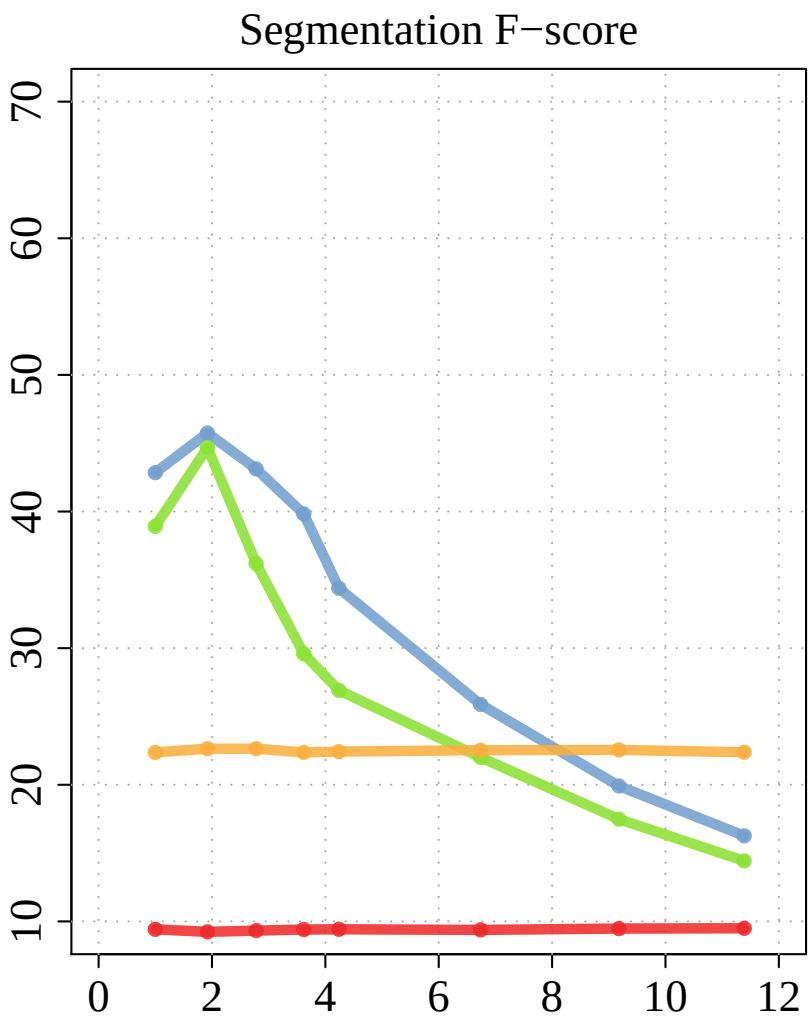


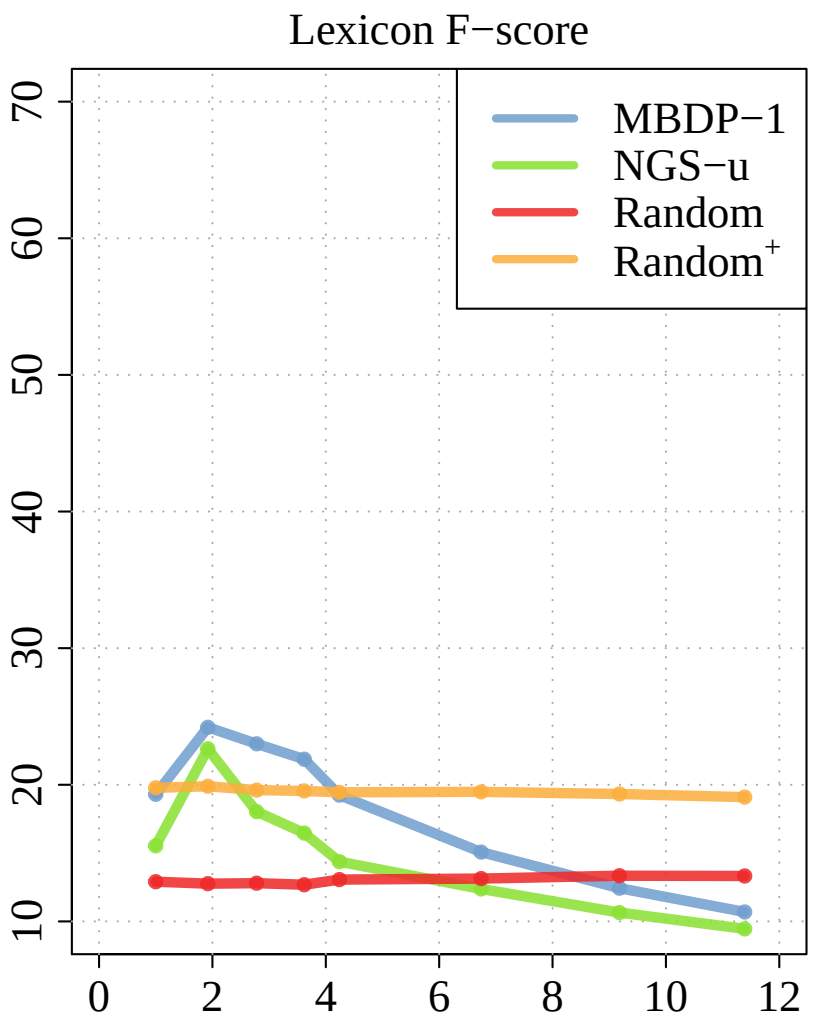
Results: French



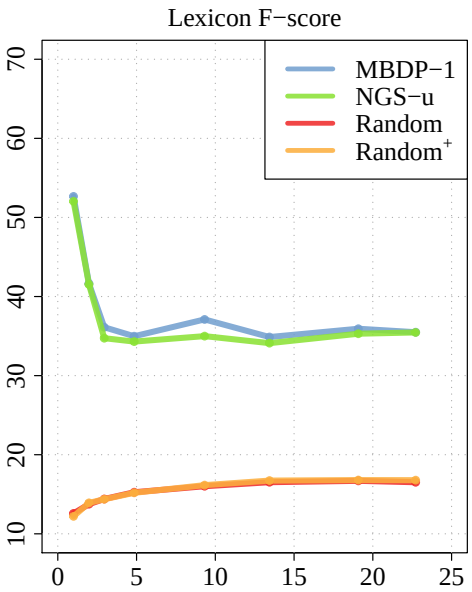


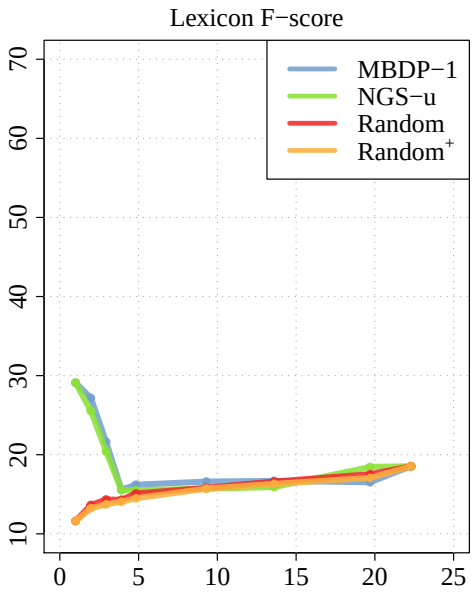
Results: Japanese

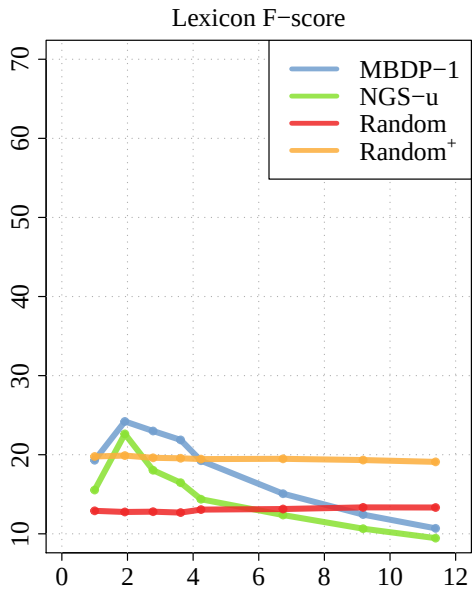




Effects of phonetic variation







Unsurprising results

- No mechanism for 'explaining away' allophonic variation;
- Any word form found by the models will be added to the lexicon.

Conclusion

Take-home message

- Cross-linguistic evaluation is not dispensable;
- Phonetic inputs impact word seg. models' performance.
- Phonological knowledge $<$ word segmentation?

Where to go from here?

- Incorporate some mechanism to handle noise and/or variation;
- Use the imperfect lexical knowledge to help learning a phonology.

$\rightsquigarrow$ *Combining indicators of allophony*, ACL'11 Student Session.

Conclusion

Take-home message

- Cross-linguistic evaluation is not dispensable;
- Phonetic inputs impact word seg. models' performance.
- Phonological knowledge $<$ word segmentation?

Where to go from here?

- Incorporate some mechanism to handle noise and/or variation;
- Use the imperfect lexical knowledge to help learning a phonology.

$\rightsquigarrow$ *Combining indicators of allophony*, ACL'11 Student Session.

/θæŋk͟ju:/