Towards large(r)-scale cross-linguistic study of speech:

prosodic case studies

Morgan Sonderegger

McGill University

Northwestern University

6/1/2018

Introduction

- Speech is highly variable

- Structure and sources of variability
  - Central Qs in linguistics/speech sciences
  - Decades of work → much known
  - Scale: mostly handful of cues (VOT, formants), languages (English), hand measurement

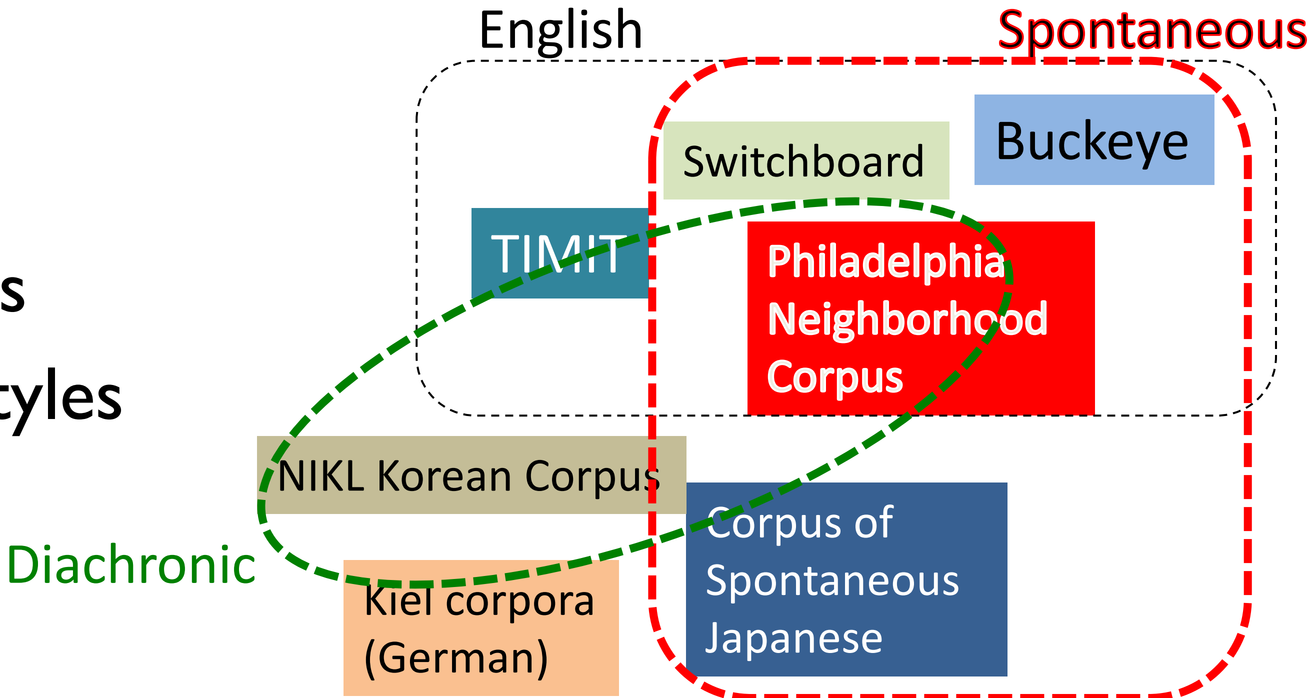
- Most of what we know is from fine-grained studies

Introduction

- Huge amount of annotated speech data exists
  - Corpora
  - Academic labs
  - Web

At least orthography + audio

- Across
  - Languages
  - Speech styles
  - Time



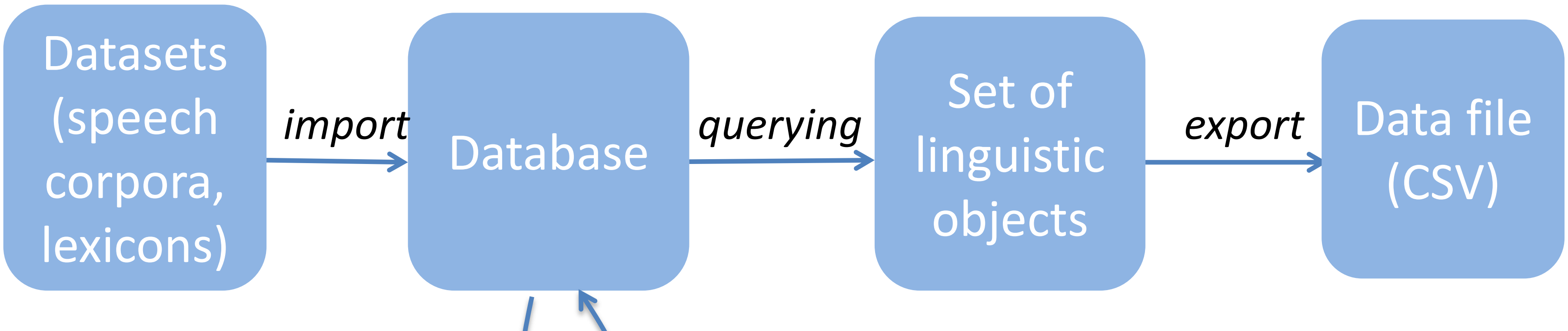
Introduction

- Large(r)-scale studies
  - corpora + (semi) automatic analysis + statistical modeling
  - Scale up
  - Less careful

- Today: two case studies

- Enabled by software facilitating large-scale studies
- Claims:
  - New insights
  - Complementary to fine-grained studies

Polyglot-Speech Corpus Tools



Datasets (speech corpora, lexicons)

import

Database

querying

Set of linguistic objects

export

Data file (CSV)

enrichment

- Implementation
  - Python module
  - Graphical interface (under redevelopment)

McAuliffe et al. (2017) *Interspeech*

montrealcorpustools.github.io/speechcorpustools/

Polyglot-SCT: Goals

1. Scalable

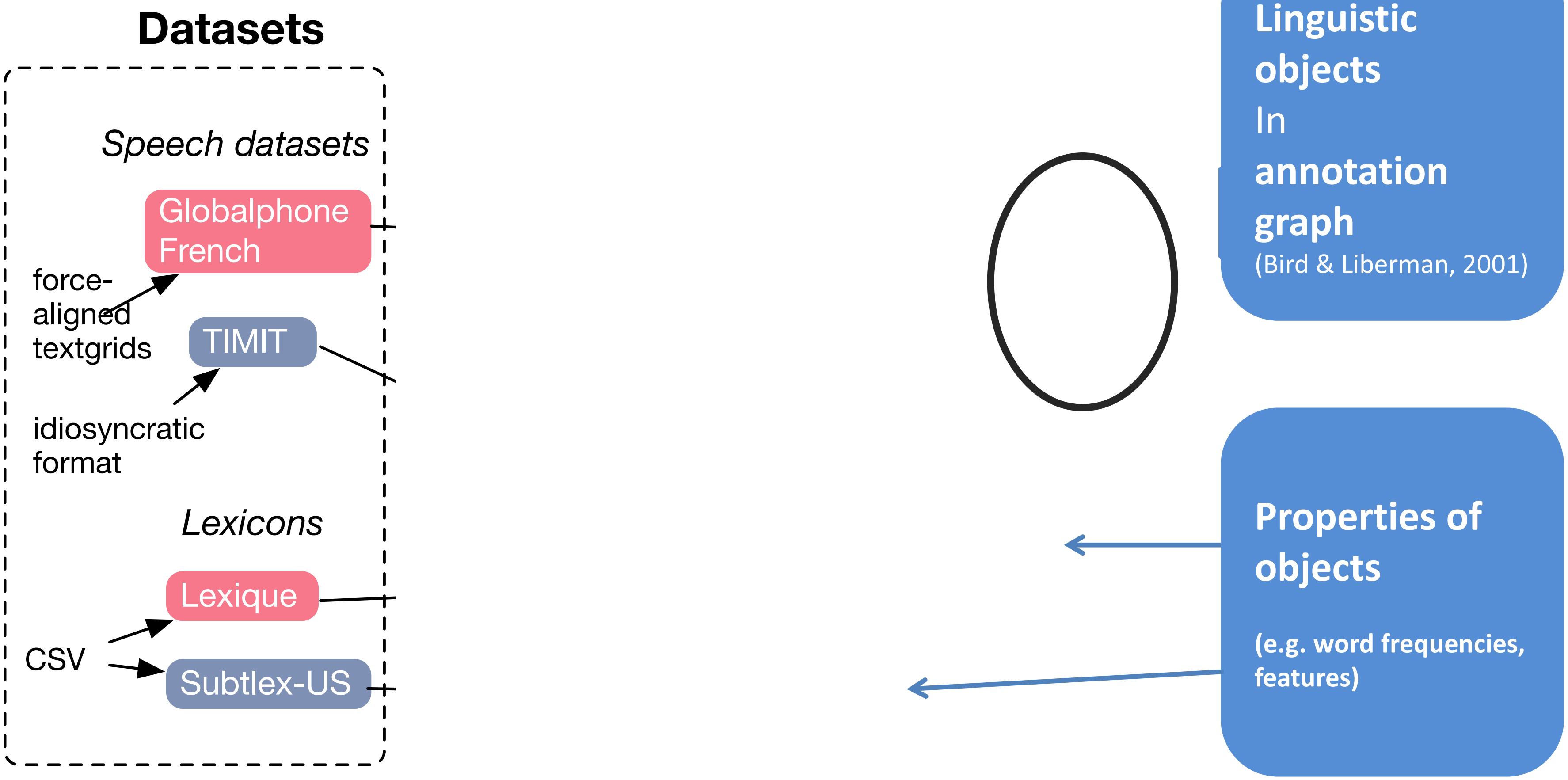
2. Require minimal technical skill from user

3. Abstraction away from dataset format

4. Querying dataset without access to raw data

- Aim to address barriers to large-scale corpus studies

Polyglot-SCT: Import



Datasets

Speech datasets

Globalphone French

force-aligned textgrids

TIMIT

idiosyncratic format

Lexicons

Lexique

CSV

Subtlex-US

Linguistic objects

In

annotation graph

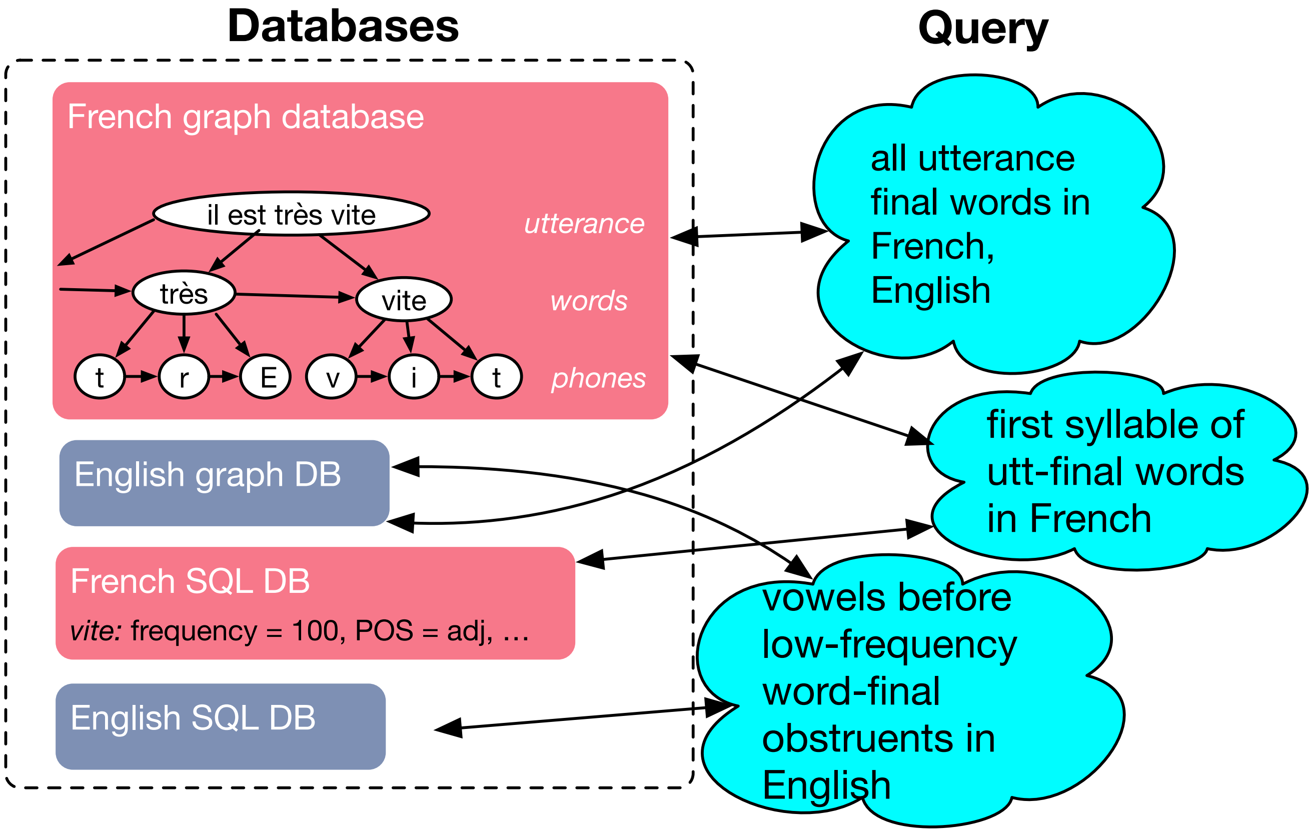
(Bird & Liberman, 2001)

Properties of objects

(e.g. word frequencies, features)

- Speech, text datasets → queryable databases

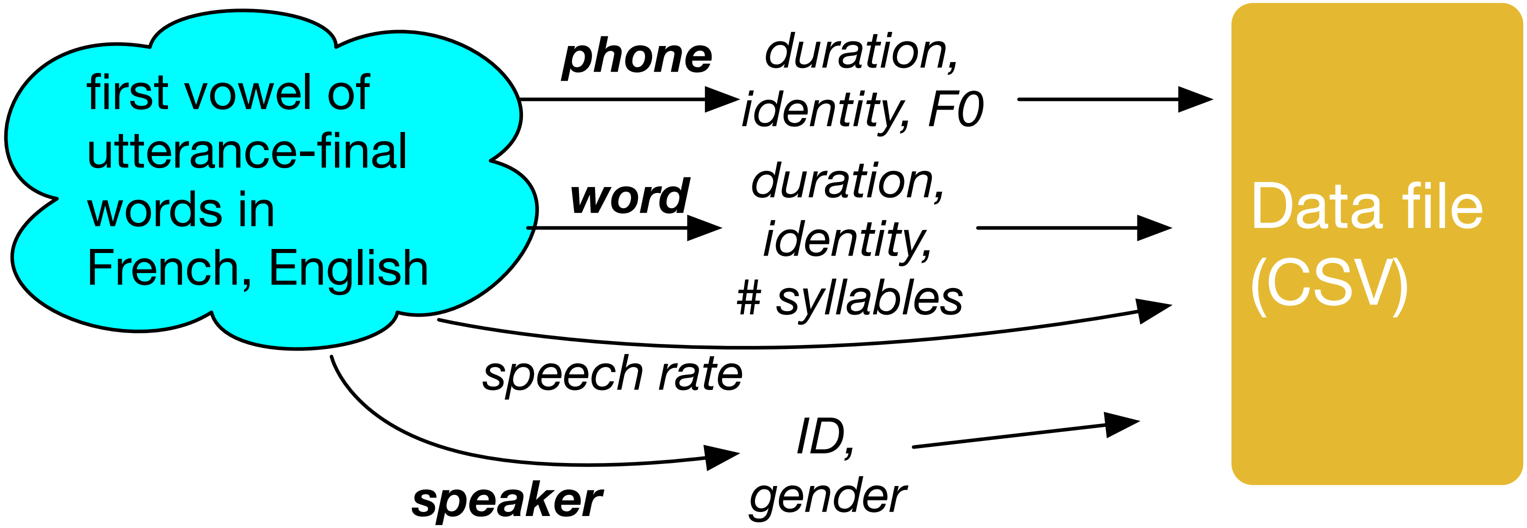
Polyglot-SCT: query



Databases
Query
French graph database
il est très vite
très
vite
t
r
E
v
i
t
utterance
words
phones
English graph DB
French SQL DB
vite: frequency = 100, POS = adj, …
English SQL DB
all utterance final words in French, English
first syllable of utt-final words in French
vowels before low-frequency word-final obstruents in English

- Find subset of linguistic objects

Polyglot-SCT: export



- Properties of objects → spreadsheet
  - (→ R, Excel)

Study 1: intrinsic F0 effects

Introduction

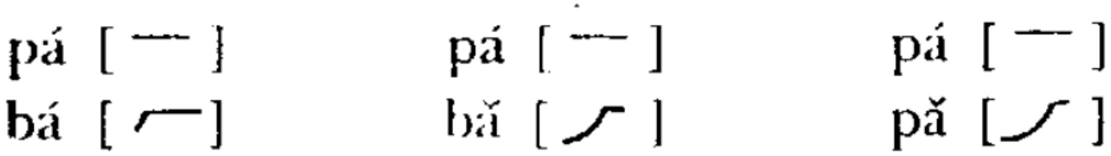
- Where does sound change come from?
- Most common:

  phonetic effect → phonological pattern

"phonetic precursors"

- Ex: tones
  - Often (e.g. Chinese):

    F0 perturbations → lexical tone



Introduction

But: most phonetic precursors never lead to sound change!

What kind of precursor can be a source of change?

- robust
  - Across speakers, languages
- … but variable
  - Individual differences, language-specific phonetics

tension

(e.g. Hombert et al. 1979, Ohala 19XX; Baker et al., 2011; Labov, 1967; Kingston, 2007; Yu, 2013)

Introduction

How robust/variable is each phonetic precursor, across languages and individuals?

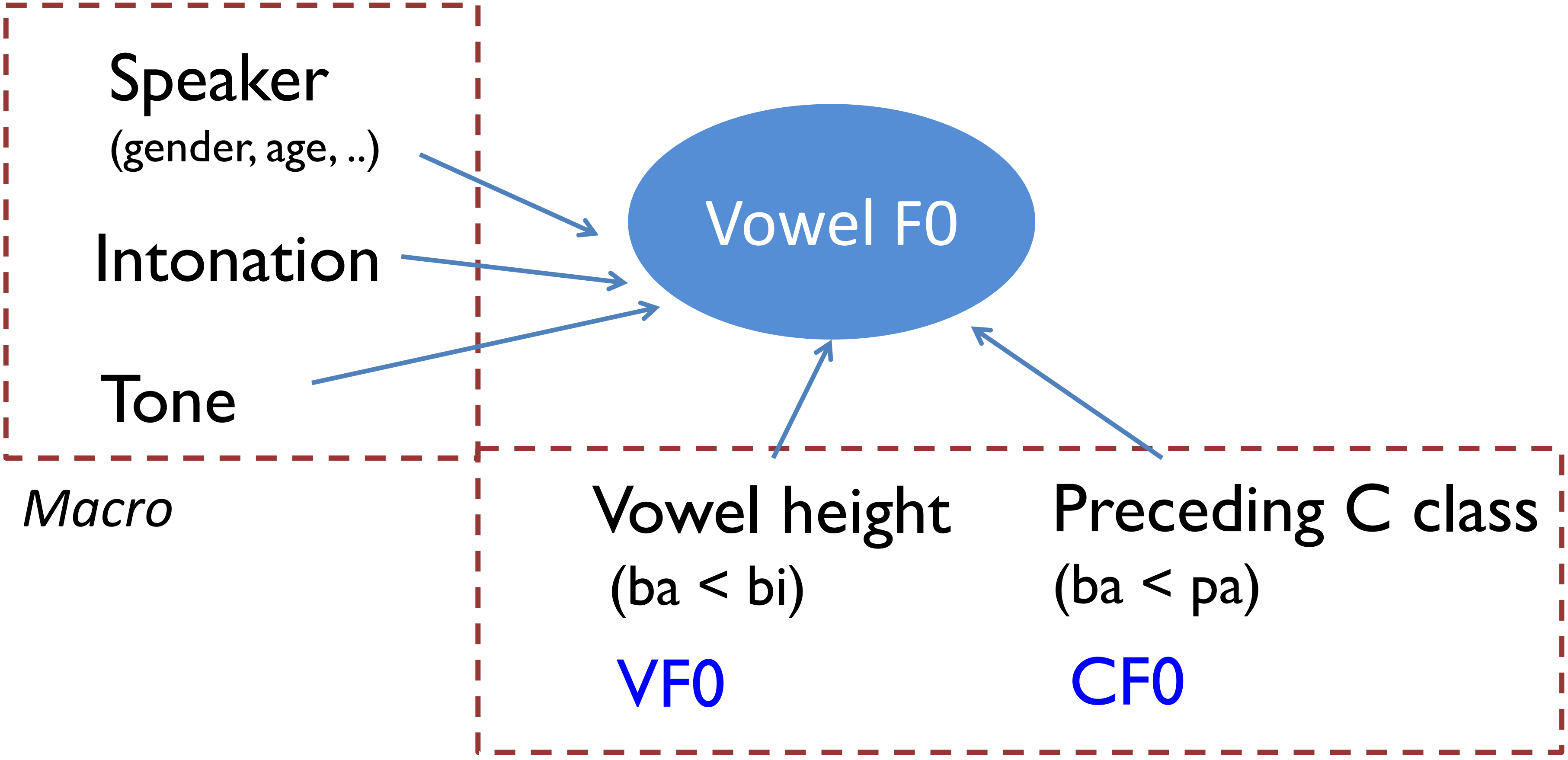
Introduction

- Methodologically hard
  - Need: big and comparable data: many languages, speakers
  - small effects, big confounds

- Approach:
  cross-linguistic corpora + automatic analysis + statistical modeling

- Q1: can a "phonetic precursor" be detected in corpus data across languages & speakers?

Influences on vowel F0



Speaker
(gender, age, ..)

Intonation

Tone

Vowel F0

Macro

Vowel height
(ba < bi)

VF0

Preceding C class
(ba < pa)

CF0

Intrinsic F0 effects

Micro

(e.g. Chen, 2011; Connell 2002; Fischer-Jørgenson, 1990; Hanson, 2009; Hoole & Honda, 2011; House & Fairbanks, 1953; Kingston & Diehl, 1994; Kirby & Ladd, 2016; Kingston, 2007; Ladd & Silverman, 1994; Meyer, 1896; Whalen & Levitt, 1995)

Intrinsic F0

- Huge literature
  - primarily: small *n*, lab speech
  - focus: mechanism (automatic vs. controlled)

Across languages:

- CF0
  - "voiced"<"voiceless": most languages
- VF0
  - [-high] < [+high] : (near-)universal
- Effect size: variable
  - Tonal ⇒ smaller effect?

Q2: How much variability in IF0 across 14 languages?

Intrinsic F0

- Strongly affected by:
  - – "Intonation"
  - – Gender (VF0)

  …
- Interspeaker variability:
  - – Often noted

Q3: How much variability in IF0 across speakers?

- Relationship to sound change:
  - – CF0 ⇒ sound change ("tonogenesis")
  - – VF0 ⇏ sound change
  - – Why?

Datasets

English
French
German
Korean
Russian
Polish
Spanish
Turkish

Hausa
Mandarin
Thai
Vietnamese

Swedish

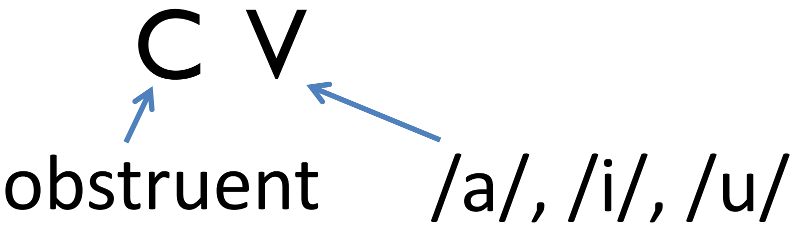
- Read sentence corpora
  - ~20 hours each
  - Force-aligned

Montreal Forced Aligner: trainable for different languages

GlobalPhone (Schulz et al., 2013), Librispeech (Panyatov et al., 2015)

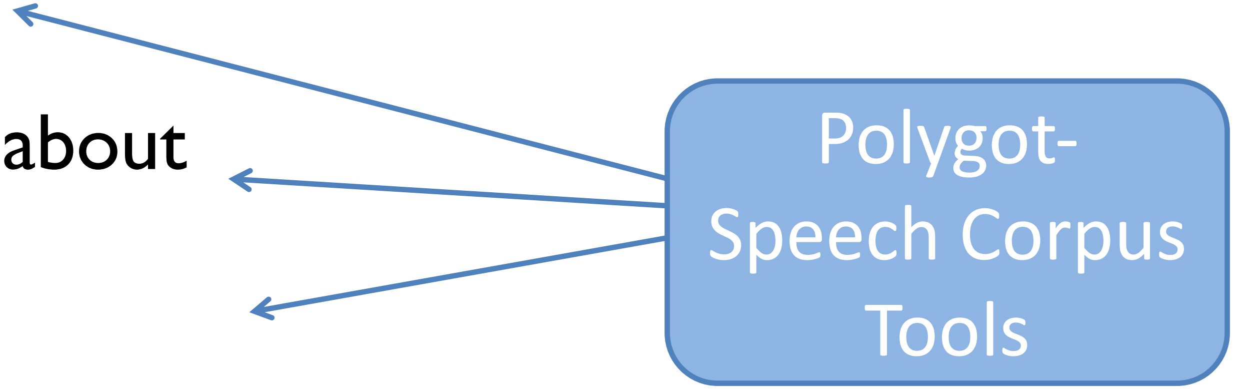
Datasets

- “Utterance-initial”



- vowel F0 (Praat)
  - F0 histogram → speaker min, max → re-extract F0

- Controls : info about
  - Speaker
  - Utterance
  - Context
  - Word



Datasets

- Data cleaning: minimize F0 errors, reduced vowels

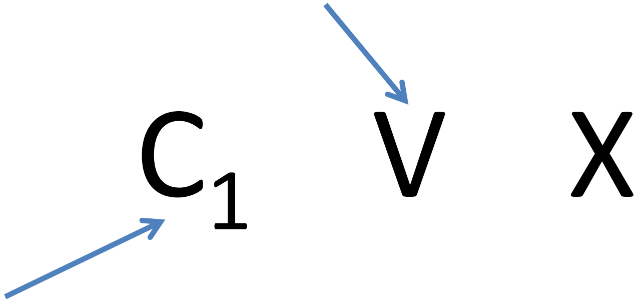
- Exclusions:
  - "bad" speakers
  - "bad" tokens (e.g. too short)

- Data per language:
  - 1.9-9.5k tokens (~2000)
  - ~100 speakers

CF0: Analysis

- One linear mixed effects model / language
- Main terms:

Response: mean F0 in first 50 ms



C_1 V X

consonant "voicing"*

overall effect + interspeaker variability

fixed effect by-speaker random slope

* Ex: French p/b, Mandarin p/p^h

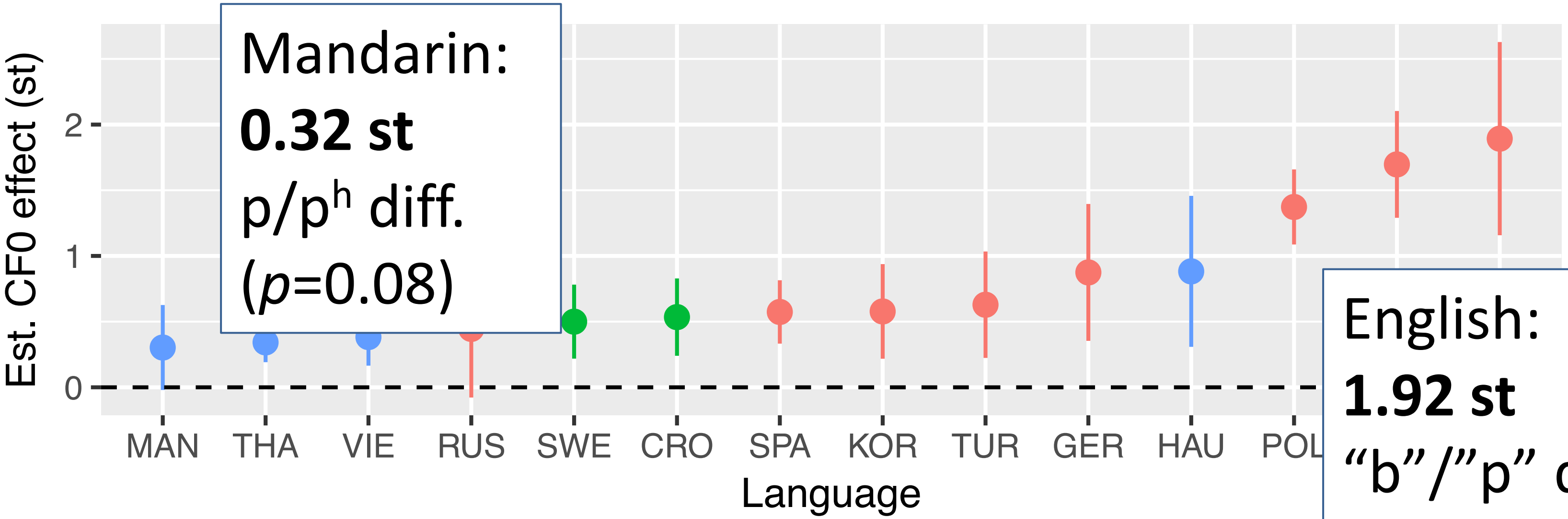
CF0: analysis

- Other terms: extra slides

- Conservative model structure

CF0: across languages

- "most voiceless" – "most voiced" effect:



Mandarin:
0.32 st
p/p^h diff.
(*p*=0.08)

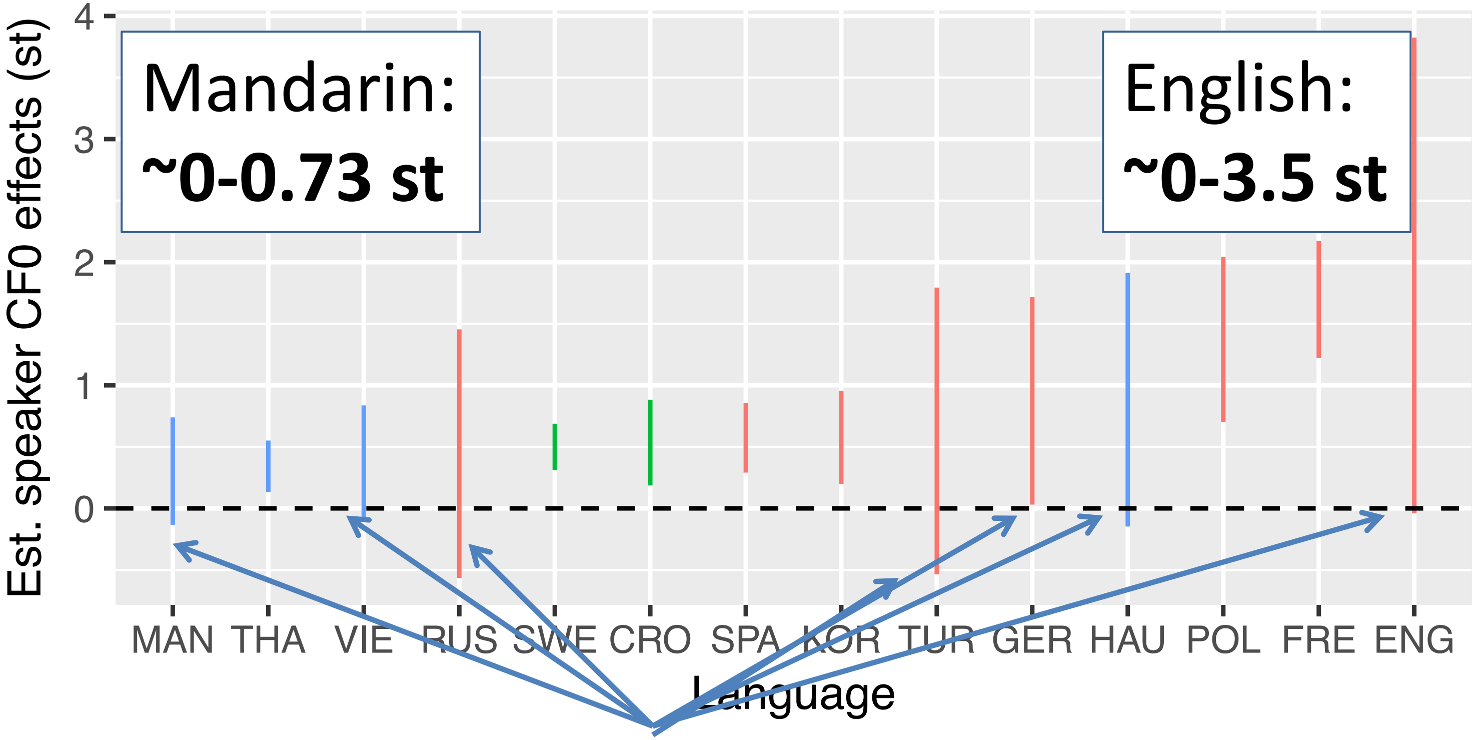
English:
1.92 st
"b"/"p" diff.
(*p*<0.001)

tonal pitch accent non-tonal

- Robust across languages
- Variable effect size
  - Non-tonal ⇒ larger effect

CF0: across speakers

- Predicted effects for 95% of individuals:

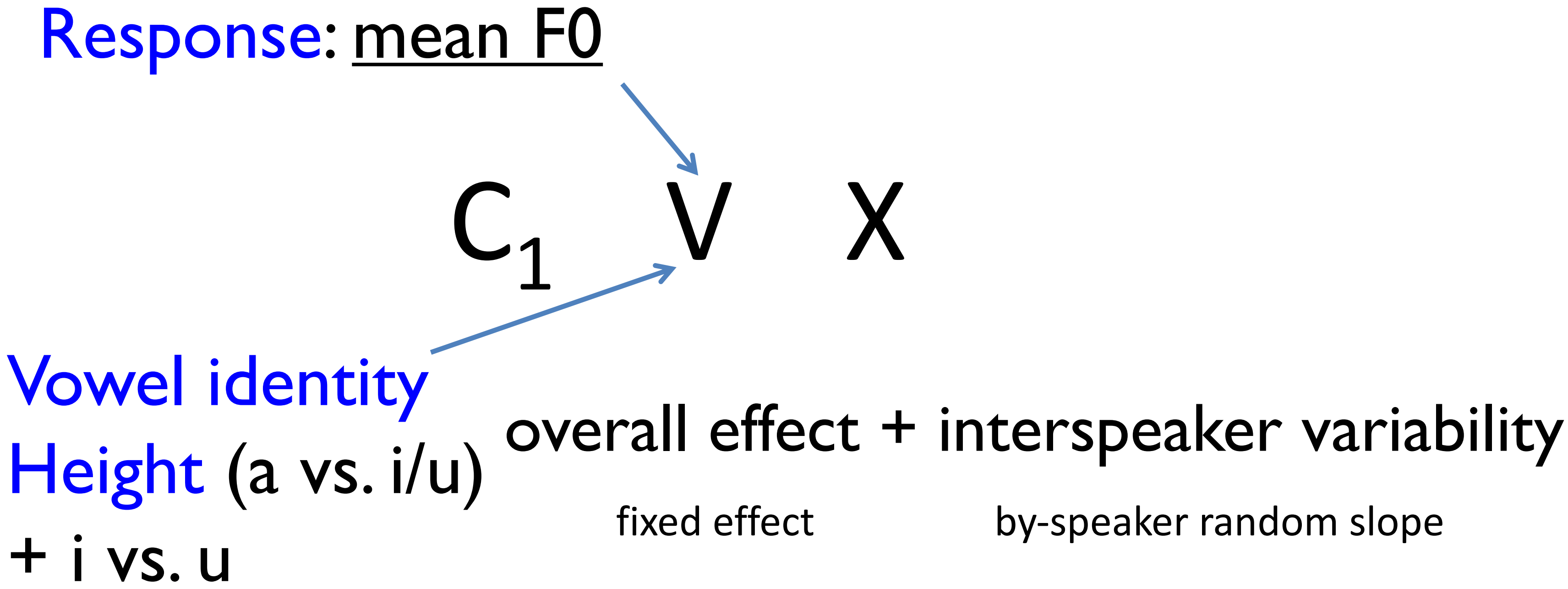

Mandarin: **~0-0.73 st**

English: **~0-3.5 st**

- Common: large interspeaker variability

VF0: Analysis

- One linear mixed effects model / language
- Main terms:



Response: mean F0

C_1 V X

Vowel identity
Height (a vs. i/u)
+ i vs. u

overall effect + interspeaker variability

fixed effect

by-speaker random slope

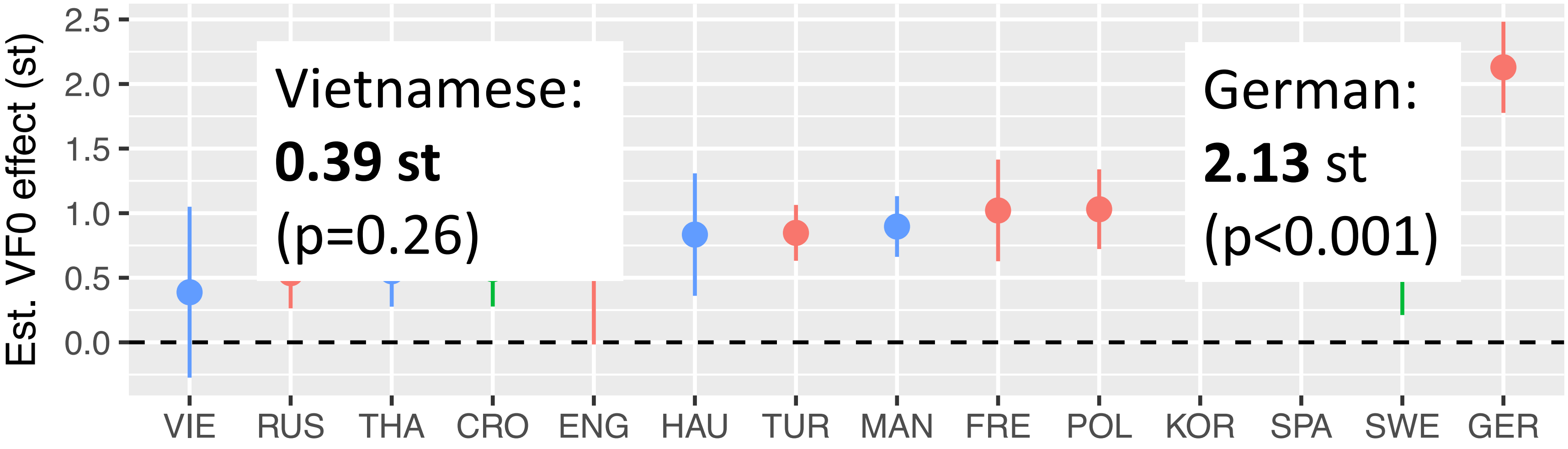
VF0: analysis

- + various controls

- Conservative model structure

VF0: across languages

- High – low vowel effect:



Vietnamese:
0.39 st
(p=0.26)

German:
2.13 st
(p<0.001)

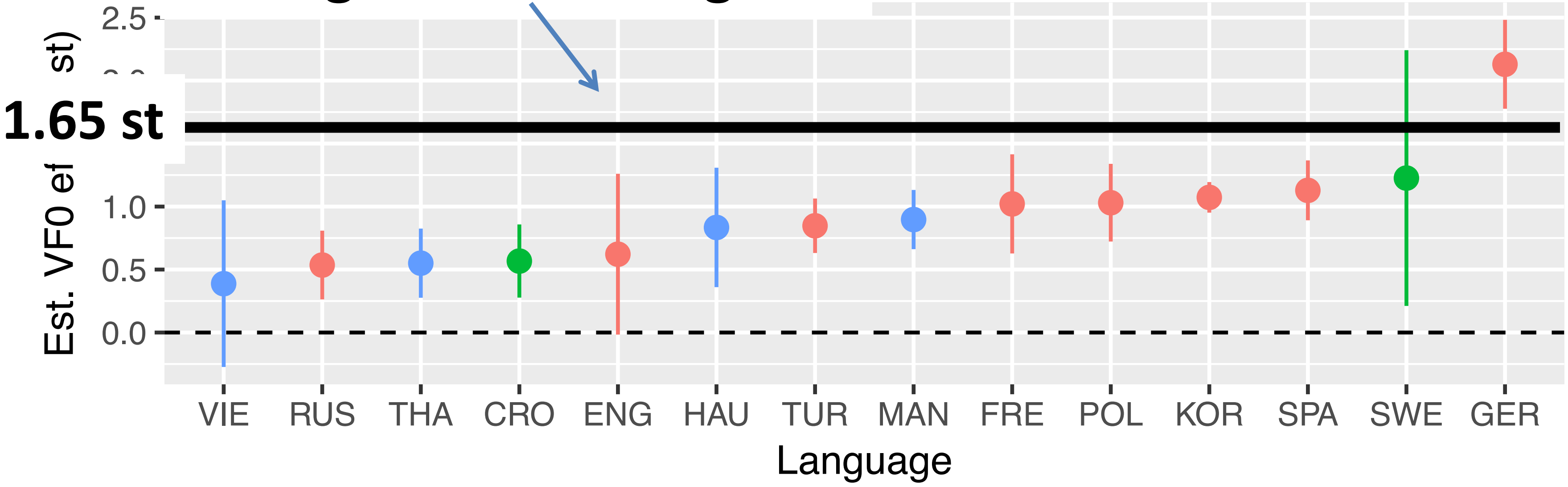
tonal pitch accent non-tonal

- mostly robust across languages
- variable effect size
  - – Non-tonal ⇒ generally larger effect

Average effect across gender, tone, etc.

VF0: across languages

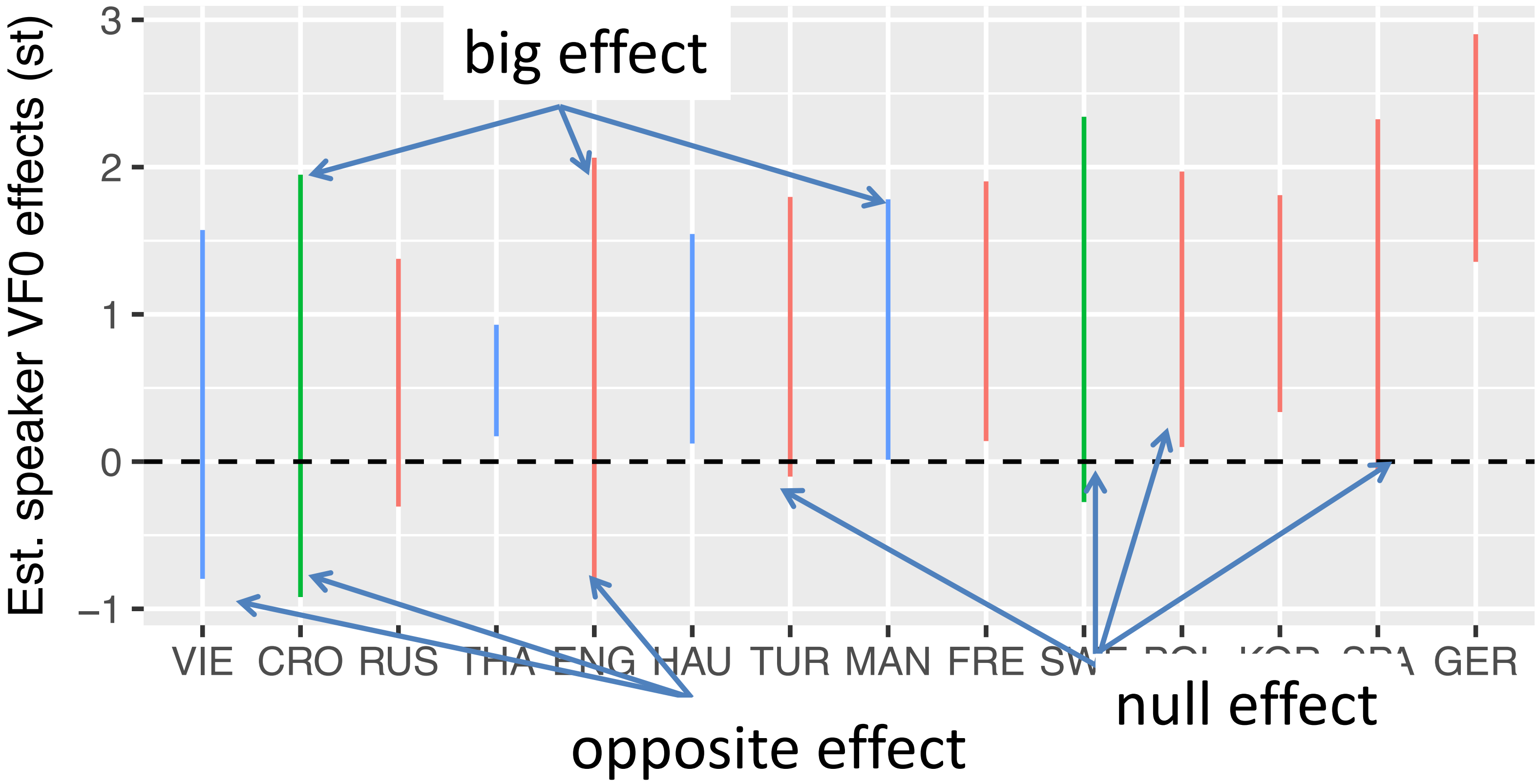
Whalen & Levitt (1995)
average over 31 langs:



- Sentences vs. lab speech?
- (or artifact of methodology?)

VF0: across speakers

- Predicted effects for 95% of individuals



big effect

opposite effect

null effect

- Common: large interspeaker variability

Discussion

- IF0 effects can be detected using
  - Corpus data
  - Fully automatic analysis
  - Basic statistical controls
  - *n* =~2-4k
- Not obvious!

- Demonstrates feasibility of large-scale studies of phonetic precursors (involving F0)

Discussion

- Robust group-level IF0 effects across languages
  - same direction
  - "universality" (Whalen & Levitt, 1995)

- Very different effect sizes
  - One reason: tonal/pitch accent language
    ⇒ smaller IF0 more likely
    (hypothesized for VF0: Connell 2002)

- Fits with automatic + controlled mechanism
  (c.f. Hoole & Honda, 2011)

Discussion

- Large interspeaker variability in IF0 magnitude common, within language
  - ⇒ there are some speakers with null/large effects
  - Still, most speakers show effect in same direction

- Overall: IF0 effects
  - robust across languages
  - variable across speakers
- Both important for sound change

- Related to actuation: why sound changes from IF0 possible, but rare? (Kingston, 2007)

Study 2: duration compression effects

With: Michael McAuliffe

Michael Wagner

Introduction

- Major aspect of speech timing: longer linguistic unit ⇒ compressed sub-parts

- Ex: *stick*, *sticky*, *stickiness* (Lehiste, 1972)

- cover term: duration compression effects

Introduction

- Menzerath's Law (Menzerath, 1928, 1954)
  - 'The longer the whole, the shorter the parts'
  - Domain-general_(not just speech)
  - longer words ⇒ shorter average syllable duration
  - phonetic Menzerath effect

- Related: Polysyllabic shortening (Lehiste, 1972)
  - Syllable/V durations shorter in bigger words/prosodic domains

Introduction

- Extensive work on DCEs
  - individual languages, controlled settings

- Unclear:
  - Are DCEs universal?
    (Siddins et al., 2014; Suomi, 2007; White & Turk, 2010)

Q1: can we observe duration compression effects across typologically-diverse languages?

- Today: test for phonetic Menzerath effect

Introduction

- Unclear: are DCEs just reducible to other factors?

- Fewer segments per second:
  - Speech rate
  - Longer words ⇒ fewer segments/syllable ("Structural Menzerath effect")

- Prosodic effects on syllable duration:
  - Accent
  - Initial position
  - Final position

(e.g. Sluijter, 1995; Fougeron & Keating, 1997; Oller, 1973; Klatt, 1973, 1975; White & Turk, 2010; Windmann et al., 2015)

Introduction

- Q2: can DCEs be reduced to fewer segments/second?

- Q3: can DCEs be reduced to a prosodic lengthening effect?

Datasets

English
German
Russian
Swahili
Ukrainian
Bulgarian
Hausa
Polish
Portuguese
Spanish
Swahili
Turkish

Mandarin
Vietnamese
Thai

Diverse (word) prosody

French
Korean

Croatian
Swedish

- Read sentence corpora
  - ~20 hours each
  - Force-aligned

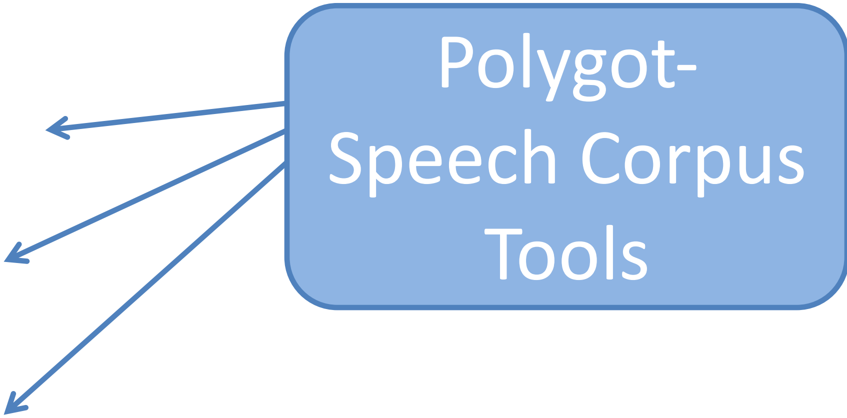
Montreal Forced Aligner

, 2013), TIMIT (Garofolo et al., 1993)

Datasets

- "Utterance" final "words"

- Measures
  - Word length (# syllables)
  - Mean syllable duration


Polygot-Speech Corpus Tools

- Controls
  - Speech rate
  - Expected syllable duration
  - Speaker, word ID
  - etc.

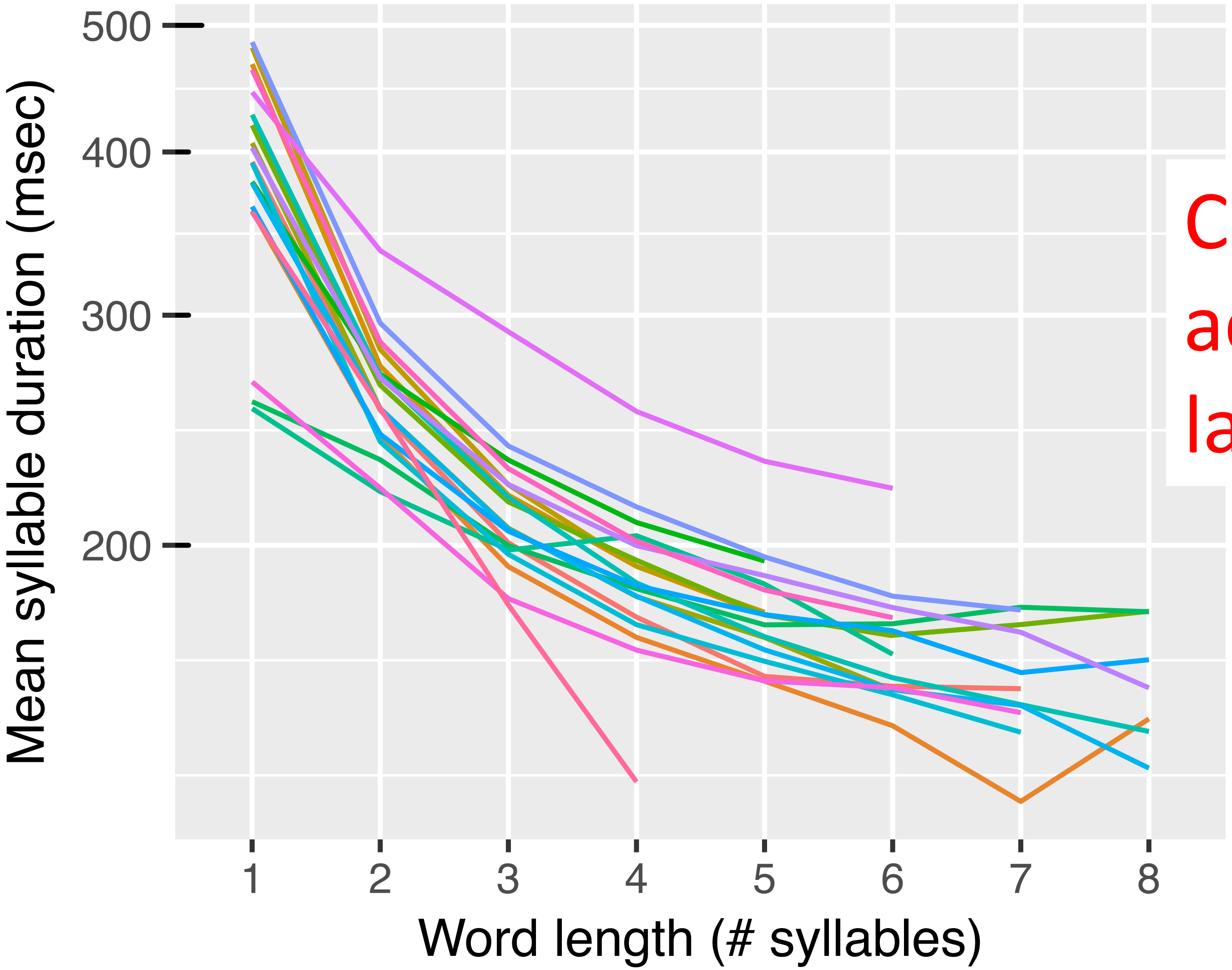
Given segmental content, for individual speaker

(Ernestus, Gahl)

Datasets

- Pruning
  - Words above language-specific cutoff length

Results: Mean syllable duration



Clear effect across languages

Analysis 1: controlling for segments/second

- Does mean syllable duration ~ word length, beyond effects of
  - Speech rate
  - Expected syllable duration
  - Who’s talking
  - Particular words
  - Utterance length

  ?

Enhanced “# segments in syllable”

Analysis I

- Menzerath-Altmann law: $y = ax^b e^{cx}$

Mean syll duration

syllables

- Our case:

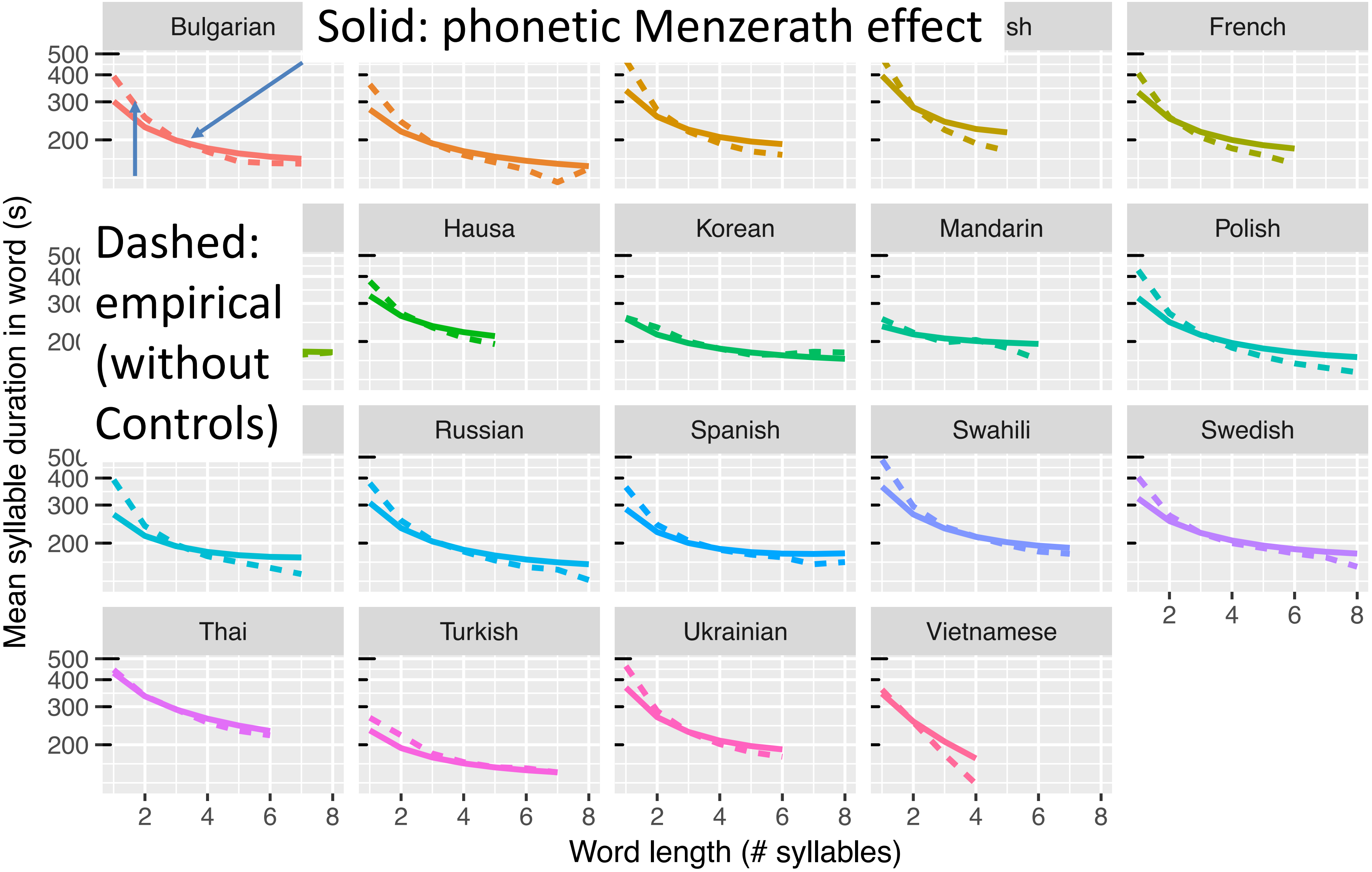
$$\log(y) = \boxed{\beta_1 x + \beta_2 \log(x)} + \boxed{\ldots}$$

Phonetic ME alone

Effect of speech rate,
expected syll
duration, etc.

- Linear mixed-effects model

Results



Solid: phonetic Menzerath effect

Dashed: empirical (without Controls)

Results: analysis 1

- Clear phonetic Menzerath effect across languages

- Q2: are DCEs reducible to segments/second?
  - No

- Empirical relationship "steeper":
  - Phonetic M effect (compressed syllables), plus
  - Structural M effect (compressed # segments)

Analysis 2: prosodic lengthening effects

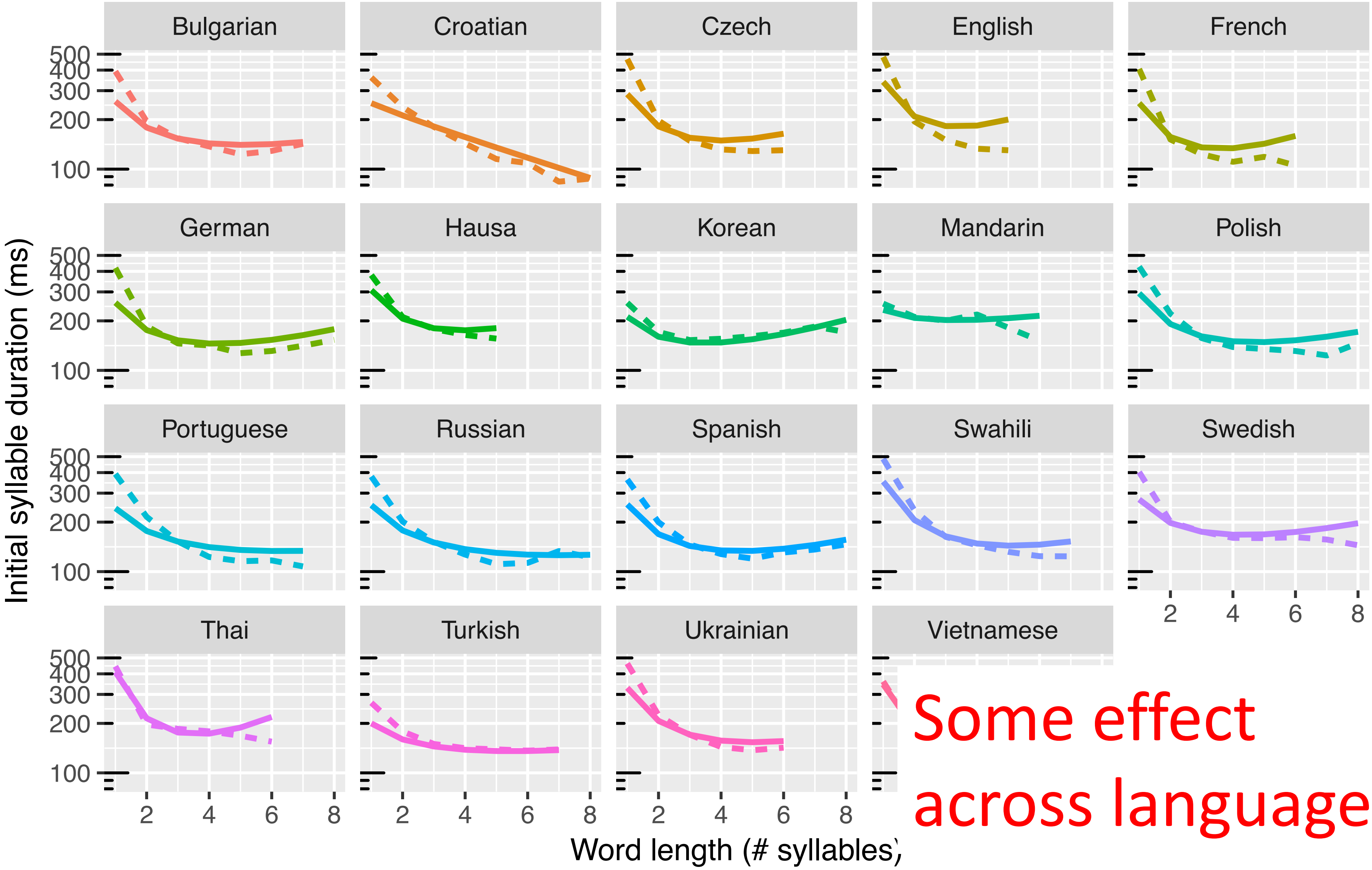
(White & Turk, 2010; Windmann et al., 2015)

- Observed effect due to
  - Initial strengthening
  - Final lengthening
  - Accentual lengthening
  - …

  across languages?

- Check:
  - Same analysis for <u>initial syllable duration</u> only, etc.

Results: initial syllables

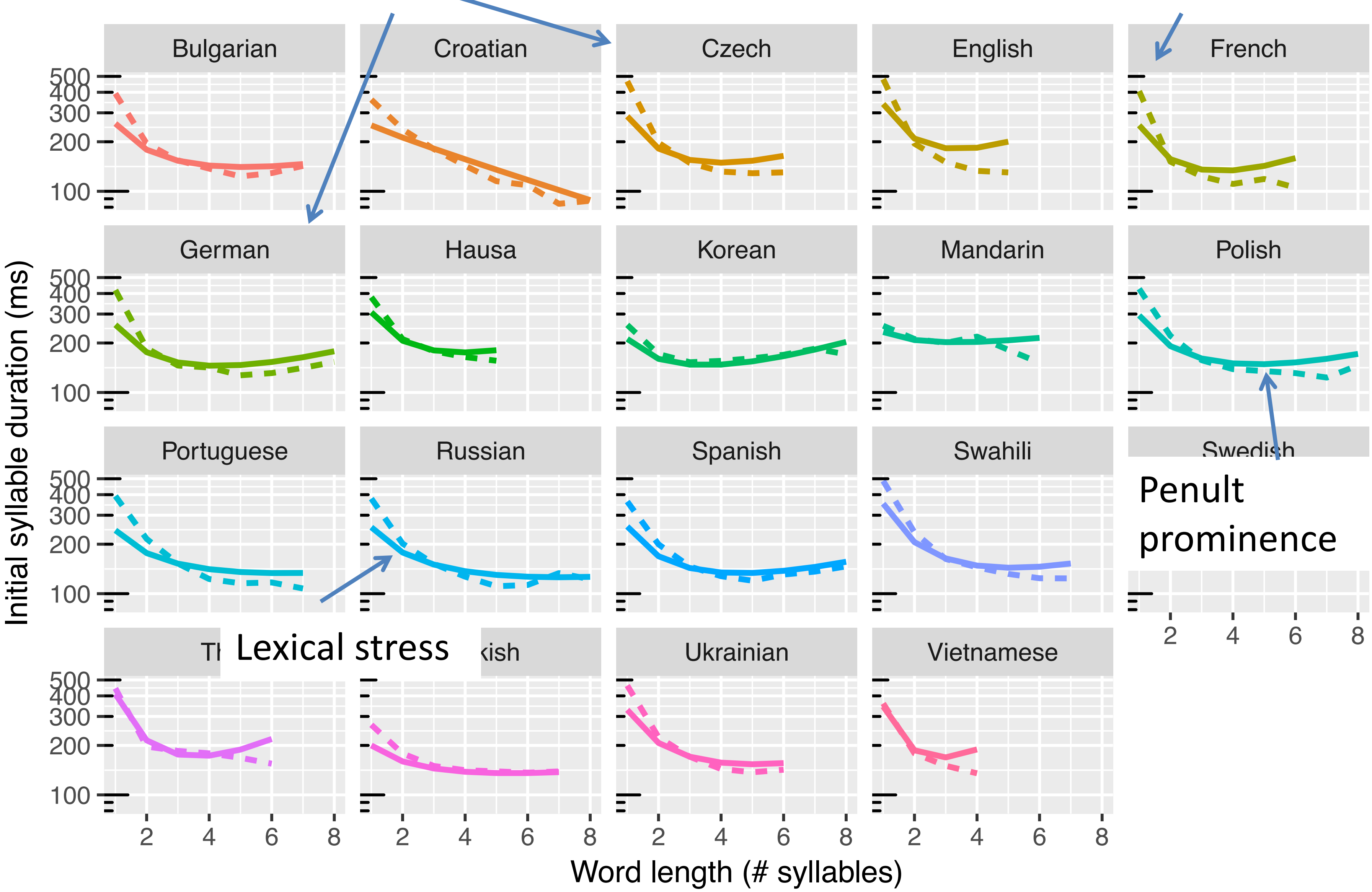


Some effect across languages

Results: initial syllables

~ initial prominence

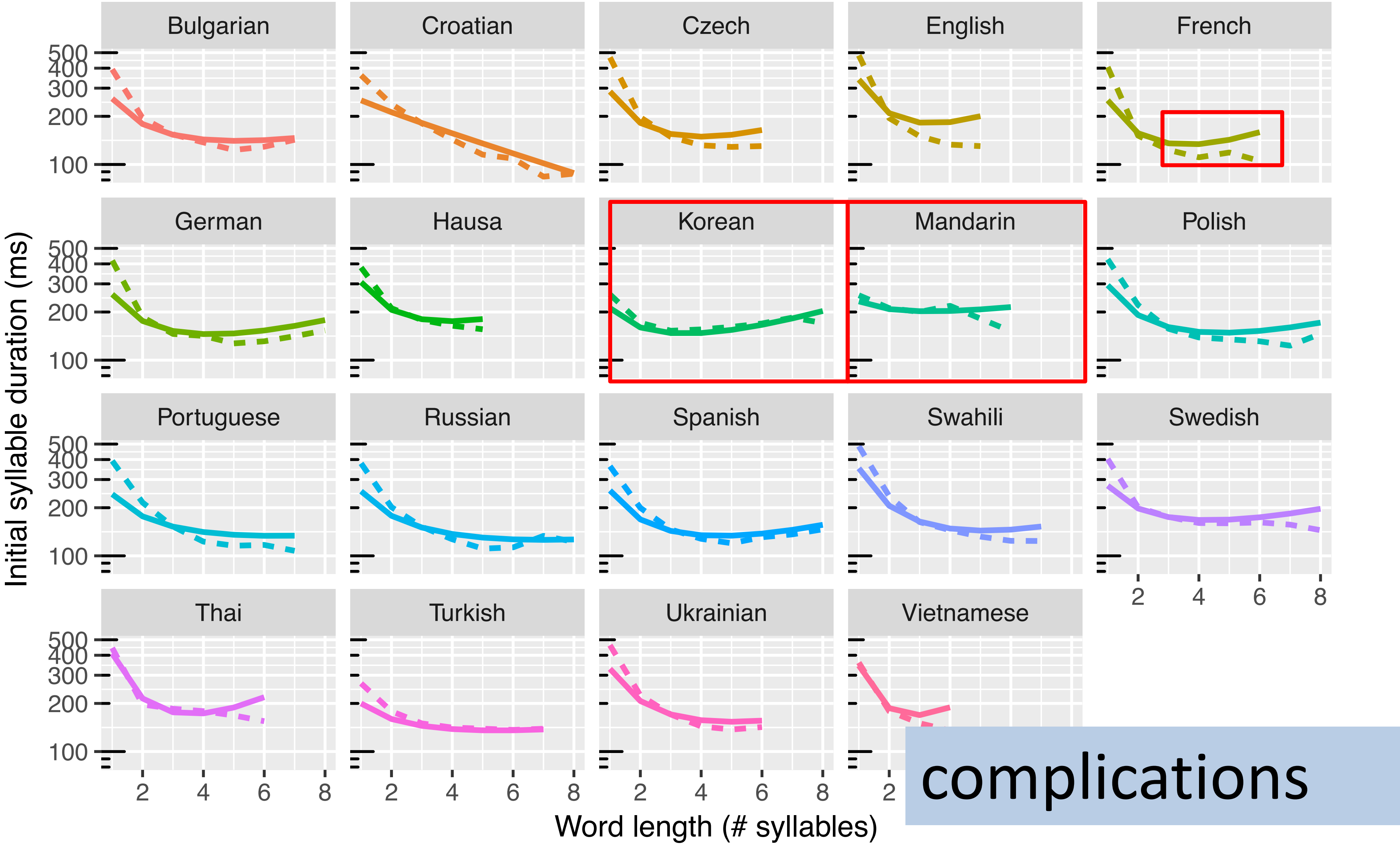
Final prominence



Penult prominence

Lexical stress

Results: initial syllables



complications

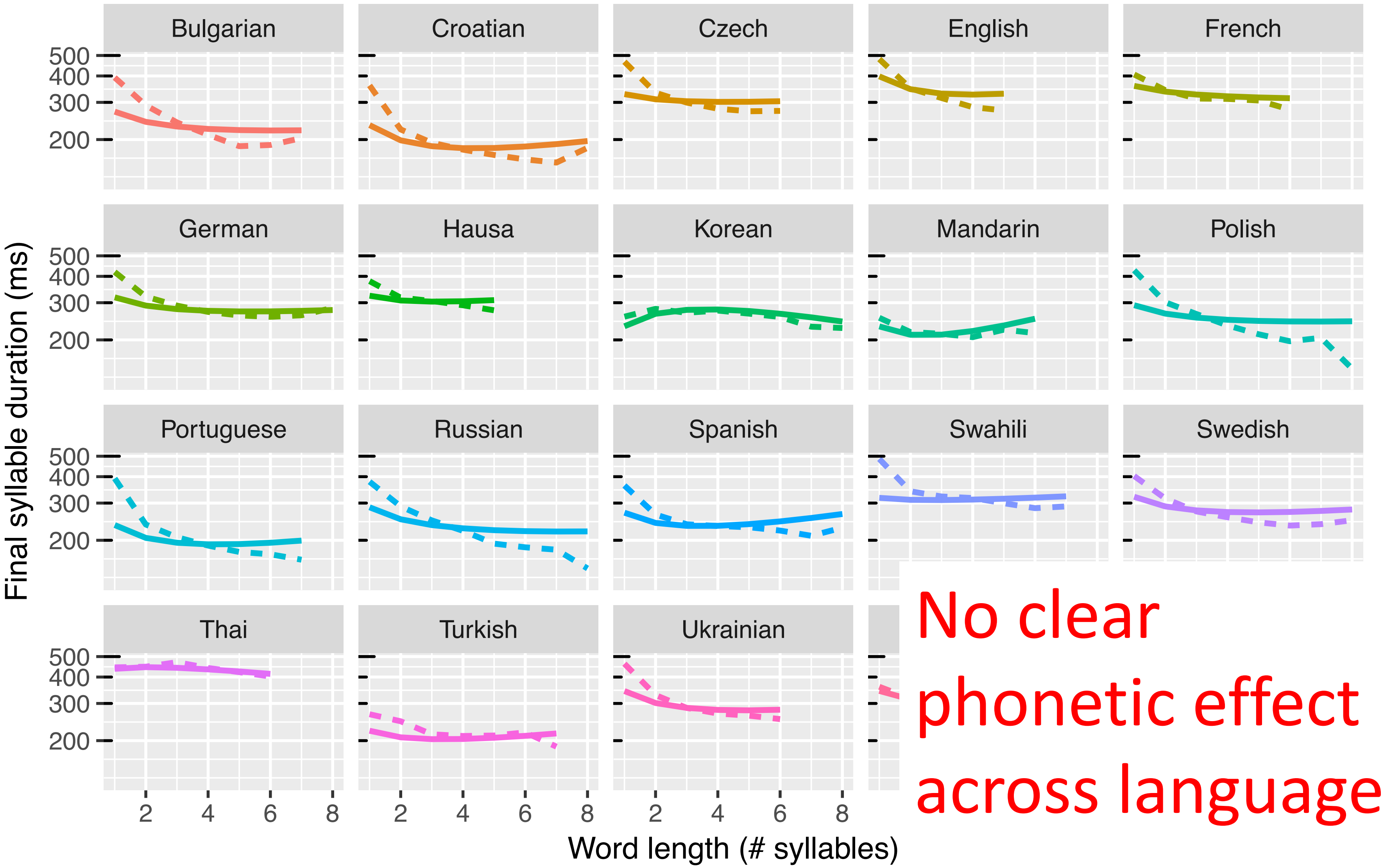
Results: initial syllables

- Consistent compression effect
  - (at least: 1-3 syllables)

- Very different prosodic systems

- Can't be just
  - Accentual lengthening
  - Initial strengthening
  - PSS on accented syllables only

Results: final syllables



No clear phonetic effect across languages

Results: final syllables

- No consistent phonetic compression effect
  - = phonetic ME

- Overridden by other factors?
  - final lengthening, language-specific prosody

- Aside: Much of empirical effect is actually due to fewer segments/syllable
  - = structural ME

Discussion

1. Duration compression effects may be universal
   - At least phonetic Menzerath effect

2. DCEs not reducible to (some) other factors

- Not obvious!

- (1)+(2) ⇒ DCEs reflect something deep about processing/planning
  - Mechanism?

Thanks

- Michael McAuliffe, Elias Stengel-Eskin, Arlie Coles
- Comments: James Kirby, Simon King, Montreal Language Modeling Lab members

- Funding:

SSHRC CRSH

Fonds de recherche
Société et culture
Québec

INNOVATION.CA
CANADA FOUNDATION FOR INNOVATION | FONDATION CANADIENNE POUR L'INNOVATION

Questions

Extra slides

Barriers to large-scale corpus studies

- Speech datasets:
  - Large
  - Complex
  - Diverse formats

- Access to many speech datasets
  - Costly or ethically restricted

- Result: requires lots of specialized code, $$, effort

CF0: analysis

- Other terms
  - "Voicing" interactions: gender
  - Controls:
    - Speaker gender, mean F0
    - Utterance length
    - V identity (incl. height)
    - Speaker, word, preceding/following phone

- Conservative model structure

VF0: analysis

- Other terms
  - V height interactions: gender
  - Controls:
    - Speaker gender, mean F0
    - C_1 "voicing"
    - Utterance length
    - V identity
    - Speaker, word, preceding/following phone

- Conservative model structure

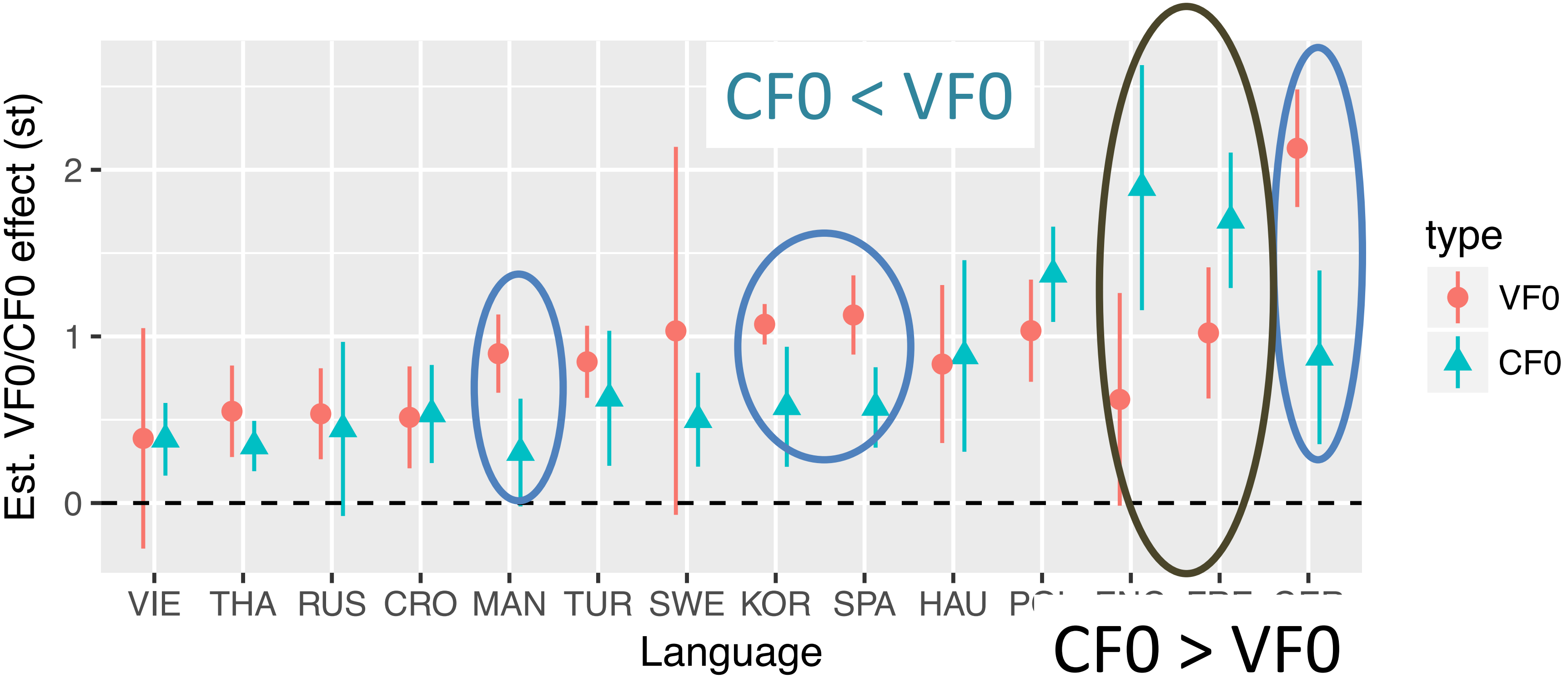
Extra: VF0 vs. CF0

- Asymmetry between IF0 effects w.r.t. sound change:
  - CF0: many attested changes
  - VF0: ~none

- Why?
  - VF0/CF0 magnitude roughly similar? (Hombert et al., 1979)
  - Perhaps perception is different (Hombert, 1979)
  - VF0 effects show more variability? (Kingston, 2011)

- Q4: Relative magnitude, variability of CF0 & VF0 across languages?

VF0 vs. CF0: effect size



CF0 < VF0

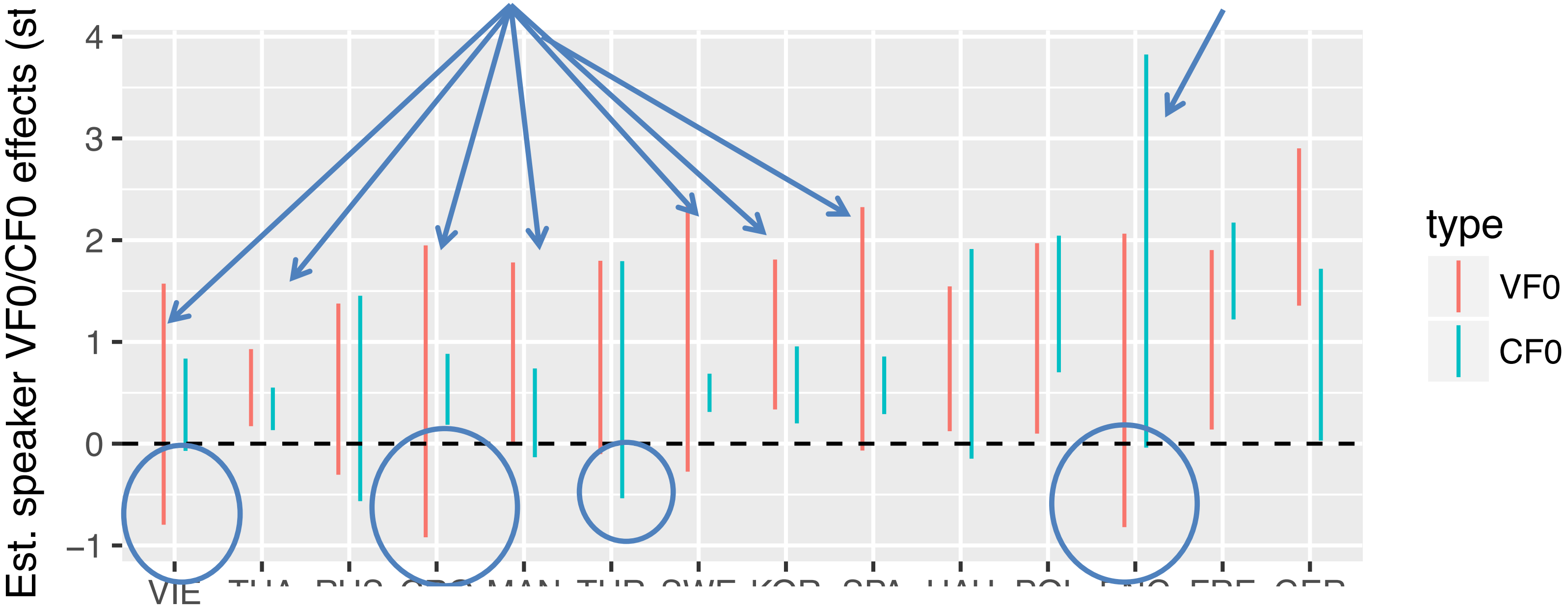
CF0 > VF0

- No clear pattern
- CF0, VF0 of ~comparable size

VF0 vs. CF0: speaker variability

VF0 >> CF0

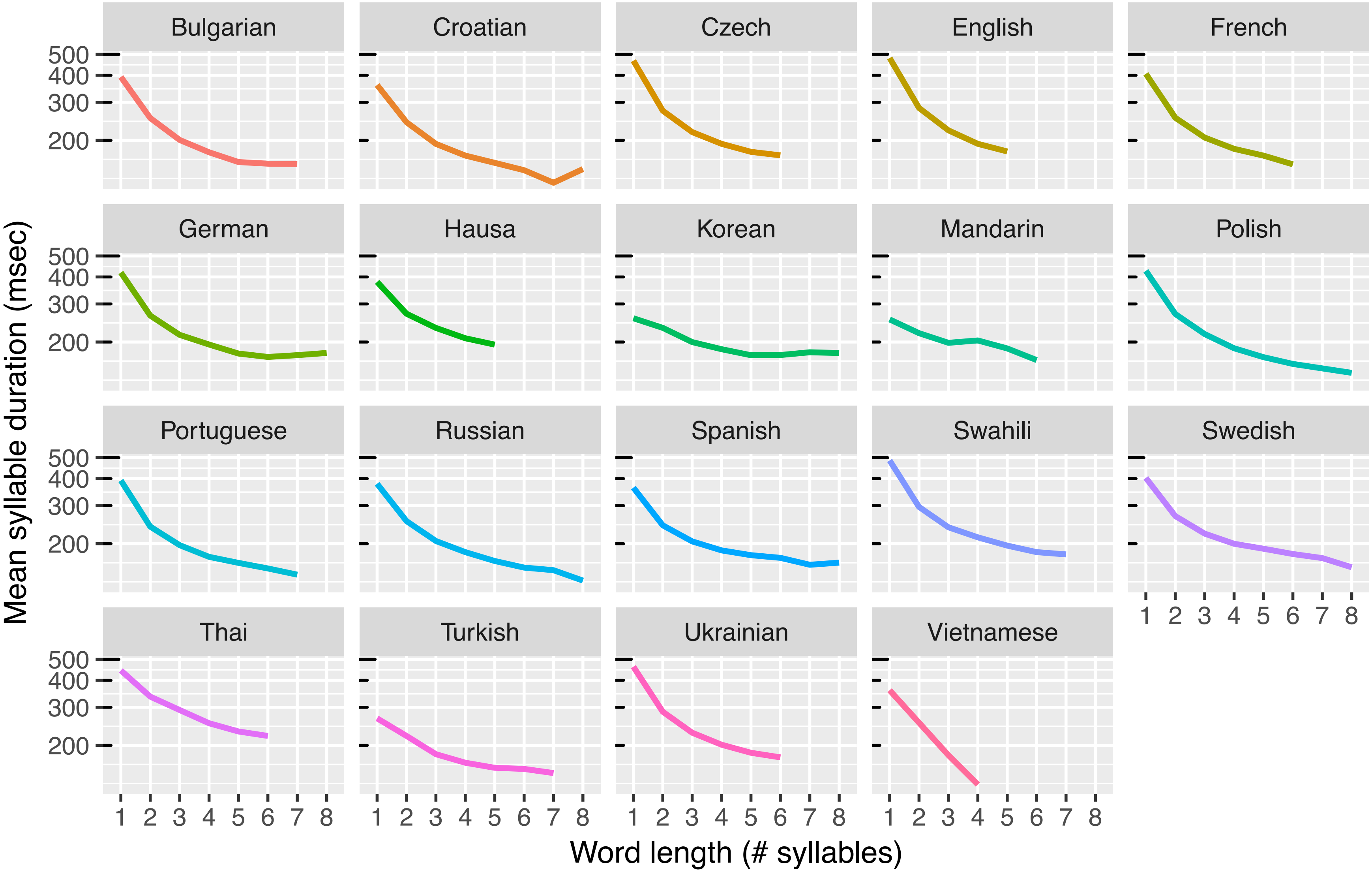
CF0 >> VF0



Minority of speakers show reverse effects

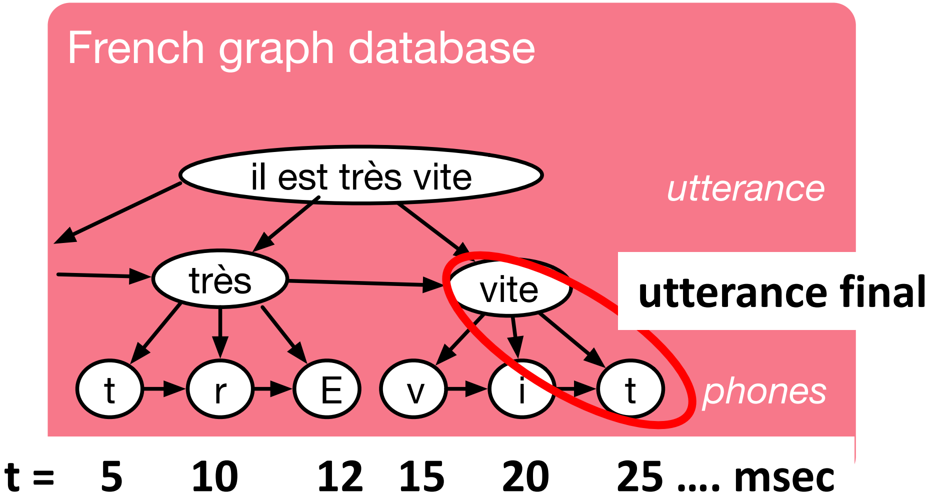
- Overall: no obvious pattern
- But: some evidence that VF0 "more variable" than CF0

Mean syllable duration



SCT: representation & enrichment

- DBs: contains properties of objects, relationships between them:
  - Positional:
    - Ex: Utterance position
  - Hierarchical
    - Ex: containing word
  - Temporal
    - Begin, end, duration



- Enrich with additional information:
  - Suprasegmental: pauses, speech rate, ..
  - Acoustic: F0, formants..