How can it be so simple? Predicting the F0-contours of Mandarin words in spontaneous speech from their corresponding contextualized embeddings with linear mappings

Harald Baayen

Hong Kong, March 13, 2025

background: English homophony

- heterographic homophones (Susanne Gahl, Berkeley)

  time and thyme, wait and weight

- word-final s (Ingo Plag, Duesseldorf)

  (singular on verbs, plural on nouns, genitive, contracted auxiliary, stem-final segment, . . .)

affix homophony: English word-final s

HOMOPHONY AND MORPHOLOGY: WORD-FINAL S IN ENGLISH

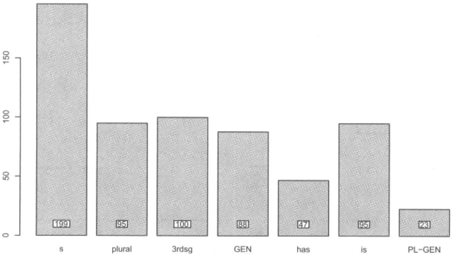
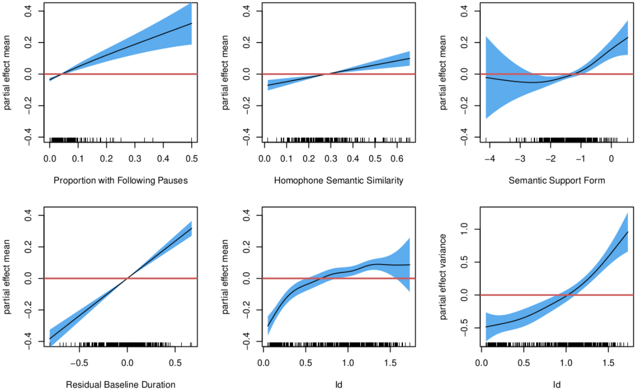


Figure 1
Distribution of different types of S in the data set. (Abbreviations: s = non-morphemic S, 3rdsg = 3rd person singular, GEN = genitive, PL-GEN = genitive-plural).

Plag, I., Homann, J., & Kunter, G. (2017). Homophony and morphology: The acoustics of word-final S in English. Journal of Linguistics, 53(1), 181–216.

phonetic detail of homophenes: English thyme and time



homophone-duration in English
(Gahl & Baayen, Language, 2024)

semantic support for form

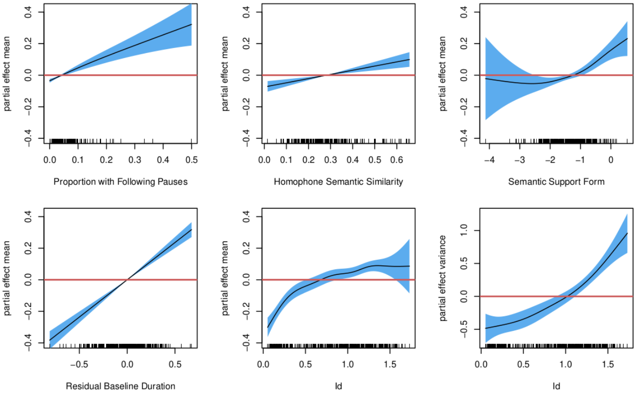
$$\underbrace{\boldsymbol{S}}_{\text{embeddings}} \cdot \underbrace{\boldsymbol{G}}_{\text{linear mapping}} = \underbrace{\boldsymbol{C}}_{\text{form vectors}}$$

given an estimate of $\boldsymbol{G}$, we obtain predicted form vectors $\hat{\boldsymbol{C}}$

the SemanticSupportForm measure is the sum of the predicted supports for the triphones of word i in $\hat{\boldsymbol{c}}_i$

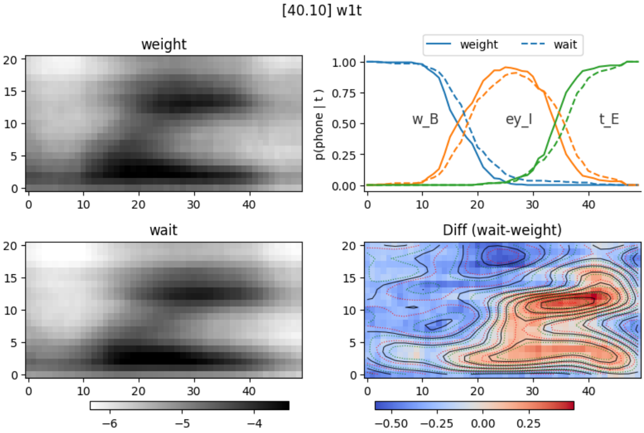
it provides an estimate of how precisely a word's form can be predicted from its embedding (meaning)

phonetic detail of homophones: English thyme and time



homophone-duration in English
(Gahl & Baayen, Language, 2024)

phonetic detail of homophenes: English thyme and time



homophone articulation in English and Mandarin (work in progress with Sean Tseng, Zhexuan Li, Mirjam Ernestus, Louis ten Bosch)

challenging three classical axioms in linguistics

1. the arbitrariness of the sign (de Saussure)

   English *dog* and French *chien*

2. the double articulation of language (Martinet)

   - a sound system with its own calculus (phonology)
   - a syntactic calculus operating with the units produced by the phonological calculus

3. language as a symbolic system

the weaver model of Levelt and colleagues
takes these axioms as givens

challenges for the three axioms

1. the arbitrariness of the sign (de Saussure)

   arbitrariness: the worst-case scenario for learning

2. the double articulation of language (Martinet)

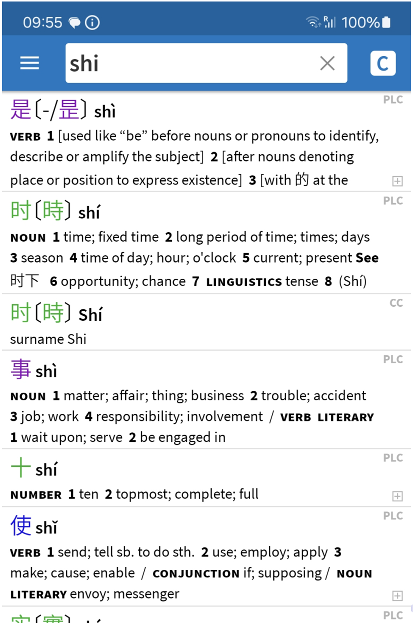
   there are remarkable isomorphies between
   fine details of meaning and phonetic realization

3. language as a symbolic system

   every abstraction comes with a cost

in what follows, I will show that, when these axioms are set aside, we can make substantial progress with predicting the realization of tones in conversational Mandarin

an L2 learner's perspective on tones in Mandarin Chinese



是〔-/昰〕 shì PLC

VERB **1** [used like “be” before nouns or pronouns to identify, describe or amplify the subject] **2** [after nouns denoting place or position to express existence] **3** [with 的 at the

时〔時〕 shí PLC

NOUN **1** time; fixed time **2** long period of time; times; days **3** season **4** time of day; hour; o'clock **5** current; present **See** 时下 **6** opportunity; chance **7** **LINGUISTICS** tense **8** (Shí)

时〔時〕 Shí CC

surname Shi

事 shì PLC

NOUN **1** matter; affair; thing; business **2** trouble; accident **3** job; work **4** responsibility; involvement / **VERB** **LITERARY** **1** wait upon; serve **2** be engaged in

十 shí PLC

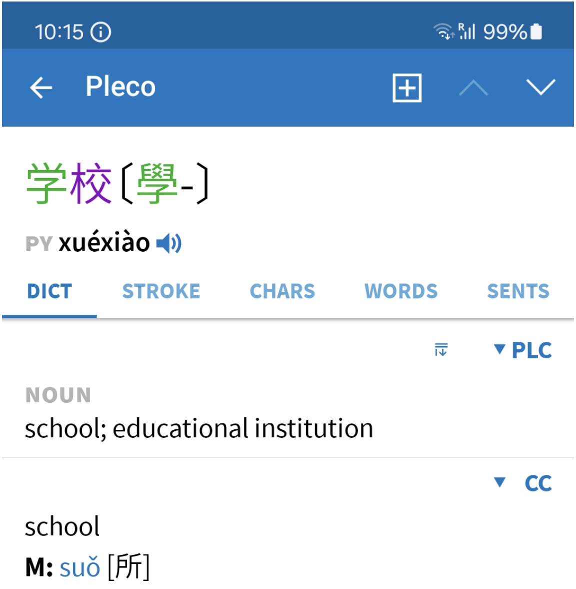
NUMBER **1** ten **2** topmost; complete; full

使 shǐ PLC

VERB **1** send; tell sb. to do sth. **2** use; employ; apply **3** make; cause; enable / **CONJUNCTION** if; supposing / **NOUN** **LITERARY** envoy; messenger

PLC

tones in Mandarin Chinese



10:15 ⓘ 99%

← Pleco ⊞

学校 (學-)

PY xuéxiào

DICT STROKE CHARS WORDS SENTS

▼ PLC

NOUN

school; educational institution

▼ CC

school

M: suǒ [所]

examples of pronunciations of xuexiao, ‘school’

- standard pronunciation
- normal pronunciation
- reduced pronunciation

(from Janice Fon’s corpus of conversational Taiwan Mandarin)

why study Mandarin tones?

as a struggling L2 learner, I am bumping into discrepancies between

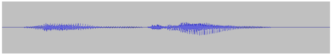
- the tones indicated by the pinyin transliteration
  (and in the tones as realized in the Pleco dictionary)

- and the tones that are actually produced

example: 文章 wen2 chang1 (high-rise followed by high)

pleco dictionary app

octave upglide on wen



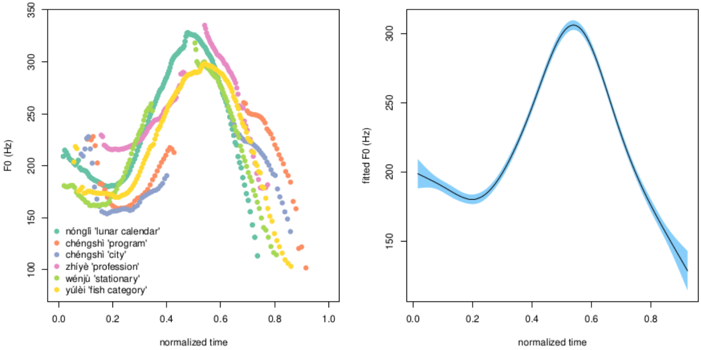
ninchanese app

level tone followed by a second level tone that is a fifth higher

why study Mandarin tones?

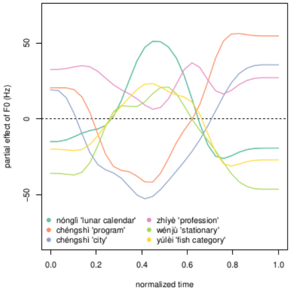
- to better understand what is going on with 文章 and many others
- hypotheses
  1. Mandarin words have statistical tonal signatures
  2. these tonal signatures are driven by words’ meanings

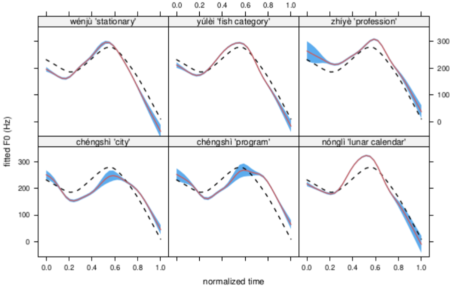
modeling pitch contours with GAMs



- the tokens analyzed here come from a dictionary (careful speech)
- we use the generalized additive model (GAM, Wood, 2017) to obtain an estimate of the pitch contour for the pitch contours of words sharing the same RF tone pattern

modeling pitch contour with GAMss





we decompose words' pitch contours into

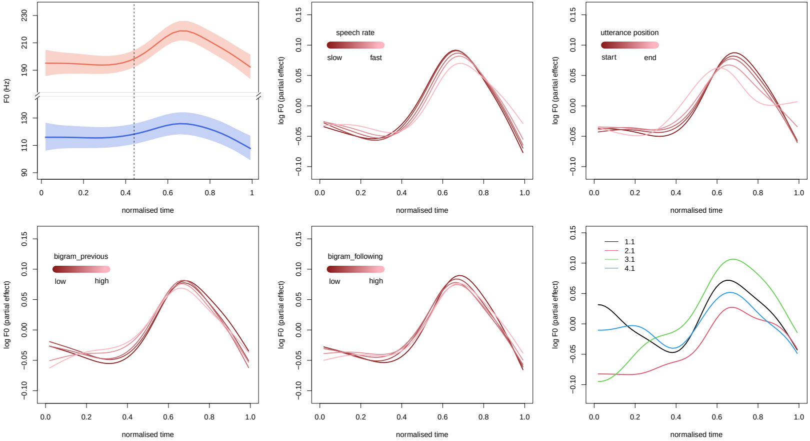
- a general smooth for the tone pattern (RF)
- word-specific smooths that present the deviations from the general tone pattern (the 'grand mean curve')

wait a moment!

how tones are realized depends on lots of factors
that you are not taking into account!

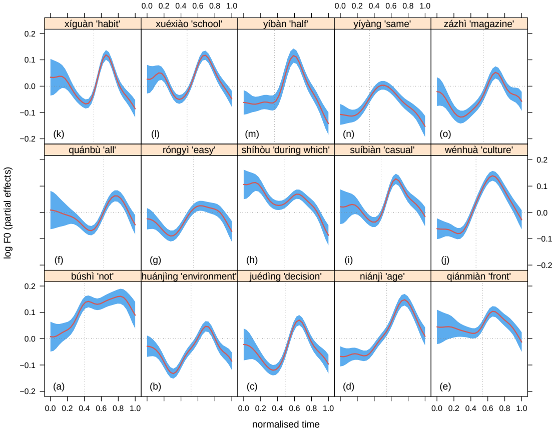
- speech rate
- position in the utterance
- probability given preceeding and following word
- tone sandhi with adjacent words
- words' segments
- gender
- speaker-specific habits

careful decomposition of pitch contours



all tokens have, according to the dictionary, the rise-fall tone pattern

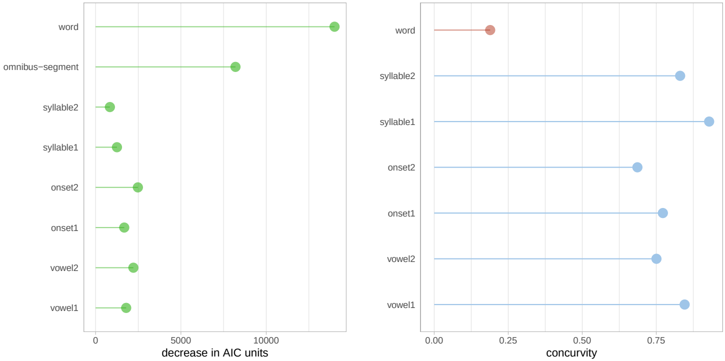
GAMs strongly support by-word wiggly random effects



Mandarin words indeed seem to have their own tonal signatures

(53 types, 9496 tokens, Yu-Ying Chuang)

statistical evidence: segmental components



word form or word meaning

where does this ‘word superiority’ effect come from?

- is this a form effect (i.e., word = word form)?
- or is this a semantic effect (i.e., word = word meaning)?

if this is a semantic effect, word meaning should be a better predictor for tone contours than word form

word senses

- we used contextualized embeddings (based on GPT-2) in combination with a sense disambiguation algorithm to couple each word token with a specific sense
- e.g., “school” as building,
  as opposed to “school” as organization
- a GAM with a factor smooth for sense instead of word provides an improved fit to the data

word senses

- sense labels are crude
- you never know where one sense ends and the next one begins (Adam Kilgarrif)
- can we improve on discrete word senses?

micro-analysis of form and meaning: replacing senses by contextualized embeddings

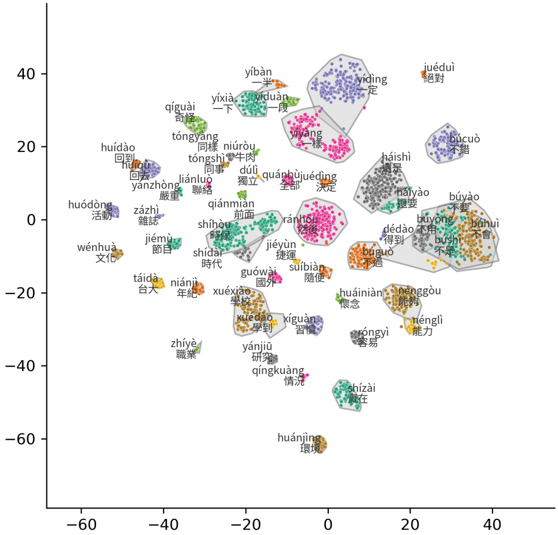
form:

the pitch contour of a given token in the spoken corpus

meaning:

the contextualized embedding calculated using GPT-2 for this token given its context in the Corpus of Spontaneous Taiwan Mandarin

the contextualized embeddings seem decent



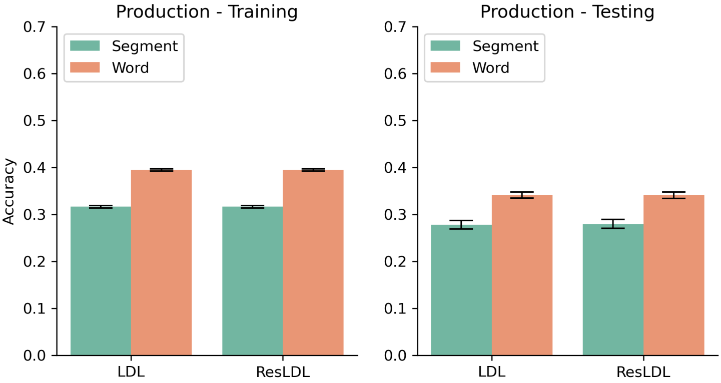
given vectors for form and meaning, we can set up a mapping between them!

$$\underbrace{\boldsymbol{S}}_{\text{contextualized embeddings}} \cdot \underbrace{\boldsymbol{G}}_{\text{linear mapping}} = \underbrace{\boldsymbol{C}}_{\text{pitch contour vectors}}$$

3,778 tokens (representing 51 word types)
50-dimensional pitch vectors
768-dimensional contextualized embeddings

prediction accuracy

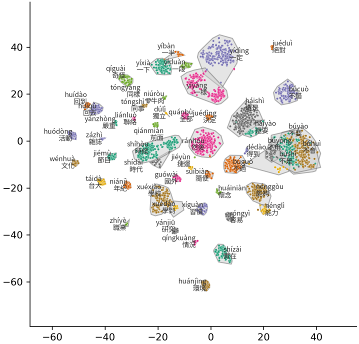
prediction for a token is taken to be accurate when the nearest neighbor pitch contour belongs to the same word type (using Euclidean distance)



the probability of a CE and pitch vector belonging to the same word type by chance is 0.038

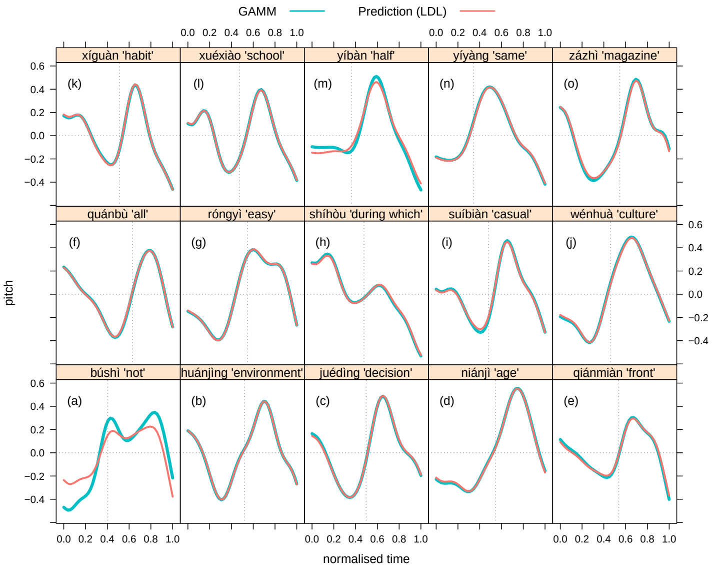
permutation baselines are 3.7% for training and 3.5% for testing

predicting word-specific tone contours



*if we take the centroid of the embeddings of a word, and map this centroid embedding onto a pitch contour using our mapping **G**, we should get the word-specific contour identified by the GAM (general curve + word signature curve)*

GAM-based and DLM-based predicted pitch contours



follow-up studies

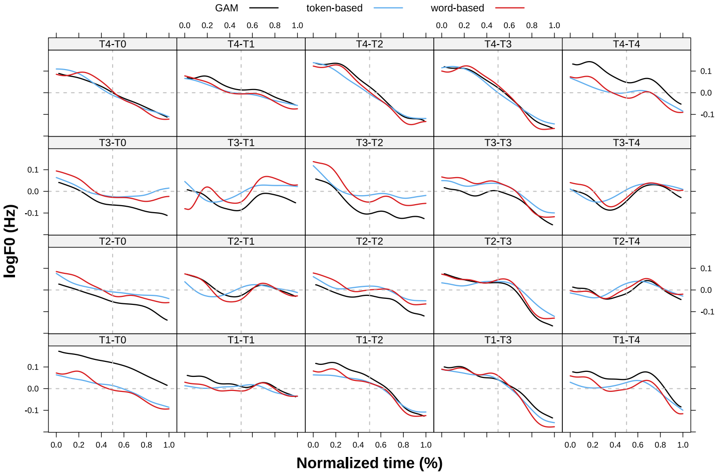
1. Mandarin 2-syllable words, all 20 tone patterns
2. Mandarin 1-syllable words
3. English left-stressed 2-syllable words

study (1) all 20 tone patterns

- thus far, we have only considered the rise-fall tone pattern
- we now extend our dataset and include words from all 20 tone patterns that are possible for bisyllabic words
- prediction:

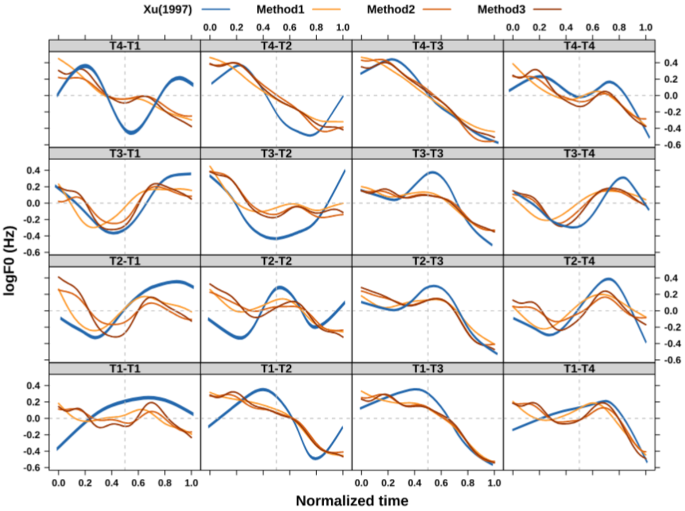
  when we consider all 20 tone patterns together, and take the centroid embedding for each tone pattern, we will obtain the tone-pattern specific pitch contours identified by a GAM that is given access to all tone patterns

20 tone patterns predicted from the corresponding semantic centroids in embedding space



(988 words, 4283 tokens, Yu-Xin Lu)

comparison with lab speech (Xu 1997)

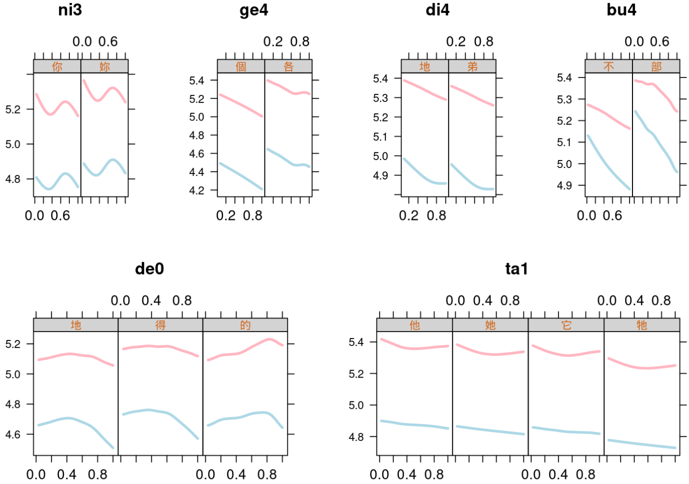


the variable importance of tone pattern is modest

the difference in AIC for models with and without tone pattern as predictor range from 3 to 40 (assessed for 4 frequent tonal contexts)

the corresponding differences in AIC for models with and without word as predictor range from 221 to 595

study (2): monosyllabic words with selected vowels

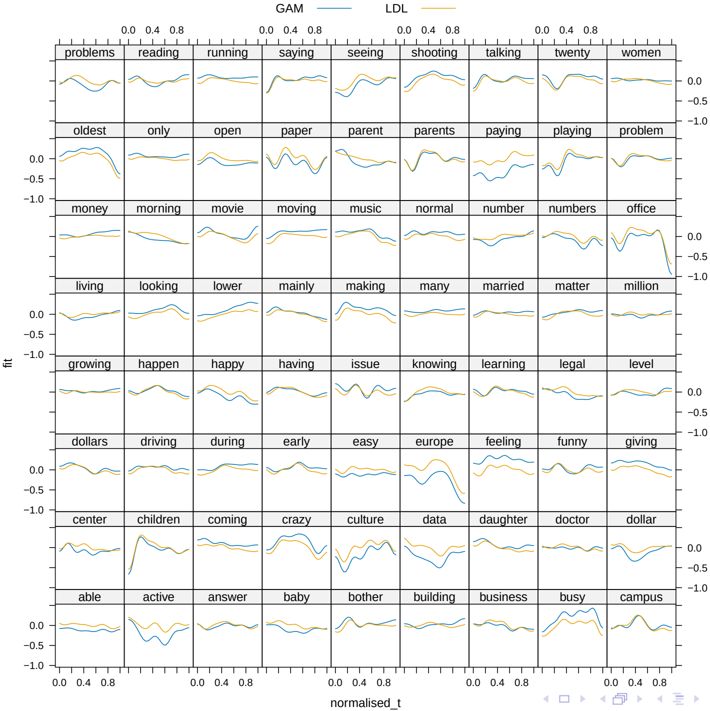


(63 types, 3824 tokens, Xiao-Yun Jin)

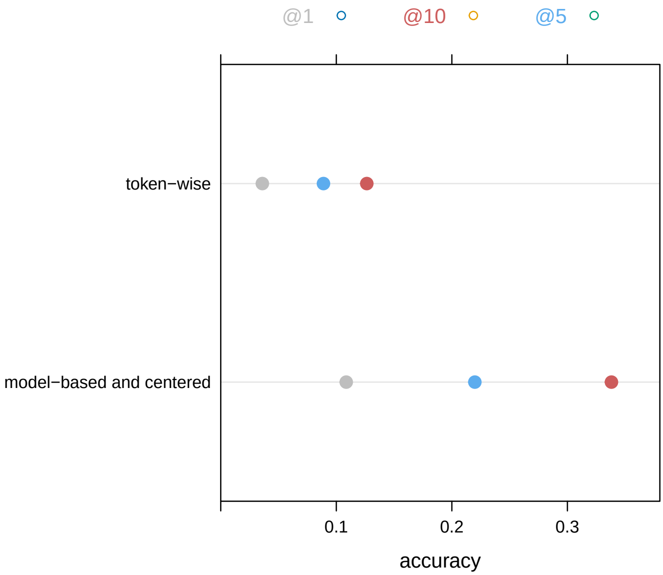
study (3) English two-syllable left-stressed words

Buckeye corpus

(72 types, 4724 tokens, Yu-Ying Chuang, Melanie Bell)


GAM
LDL
problems
reading
running
saying
seeing
shooting
talking
twenty
women
oldest
only
open
paper
parent
parents
paying
playing
problem
money
morning
movie
moving
music
normal
number
numbers
office
living
looking
lower
mainly
making
many
married
matter
million
growing
happen
happy
having
issue
knowing
learning
legal
level
dollars
driving
during
early
easy
europe
feeling
funny
giving
center
children
coming
crazy
culture
data
daughter
doctor
dollar
able
active
answer
baby
bother
building
business
busy
campus
fit
normalised_t
0.0
0.4
0.8
0.0
−0.5
−1.0

effects are more modest for English



conclusions

- pitch and meaning in context are deeply entangled

- (contextualized) embeddings are very context-dependent, they soak up all kinds of subtle co-occurrence information at the word (or subword) level

- the pitch contours of word tokens are also very context-dependent: cf. tonal co-articulation, the effects of speech rate, emotion, position in the sentence, . . .

- the present results suggest that there is considerable isomorphy between the form space and the semantic space

discussion

- the evidence for the entanglement of phonetics and semantics helps explain why current AI/NLP/LLMs are so successful

- abstract units (phonemes, morphemes, even ‘words’) are harmful for both machine learning and, I propose, also for human subliminal, automatized error-driven learning as well:

- abstraction (e.g., 20 Mandarin tones, the same forms for “homophones”) inevitably goes hand in hand with loss of detail that, however, is part and parcel of language experience and human cognition

general discussion

the challenge for linguistic theory for the coming years is to better understand how the fine details of form and meaning interact, at many levels, including

- a descriptive level (cf. English plurals and homophones)
- a high-level computational level (cf. modeling Mandarin tone)
- a cognitive-biological level (exploiting, for instance, the efficiency afforded by spiking neurons)

the challenge for applied linguistics is how to make use of these insights to improve L2 acquisition of tone

the question in the title of this talk:
how can it be so simple?

we observe linear mapping from CEs to F0 contours in normalized time

- does it work because we map from a high dimensional space (768) into a low dimensional space (50)?
- does it work because CEs are additive?
- does it work because of parallel contextual conditioning?
- is the mapping linear because linearity is more energy-efficient than non-linearity, offering evolutionary and biological advantages?