An investigation of Uyghur co-speech gestures: implications for metrical structure

Katie Franich¹, Connor Mayer², Travis Major³, and Gülnar Eziz¹

1. Harvard University, 2. UC Irvine, 3. USC

Tu+9
Cornell University
3/23/2024
Background on Uyghur stress and intonation
- ~10 million speakers
- Spoken primarily in Xinjiang, China and neighboring regions
- Southwestern Turkic language, most closely related to Uzbek
Stress in Turkic Languages

Turkish:

- A stress language with mostly final stress? (e.g. Kabak & Vogel 2001)
- A lexical pitch-accent language? (e.g. Levi 2005)
- A stress language with both edge- and head-marking intonation
  (e.g. Ipek 2015)
“[Uyghur stress is] remarkable for its complexity and instability.” Nadzhip (1971)

- If ultimate or penultimate syllable is heavy (CVV, CVC, CVVC, etc) it receives stress, otherwise final (Hahn 1991a, 1991b)
- Numerous exceptions, especially for loanwords.
- Stress falls on the first heavy syllable, otherwise final:
  - Duration is the only correlate of stress. (Engesaeth et al., 2009/2010)
- It is unclear how stress judgments were determined in each of the above resources.
Yakup (2013), Yakup & Sereno (2016)

Identified sets of minimal or near-minimal stress pairs with consistent stress judgments from Uyghur speakers.

- There were MANY other words where speakers disagreed on stress location!

<table>
<thead>
<tr>
<th>#</th>
<th>Initial stress</th>
<th>IPA</th>
<th>English gloss</th>
<th>Final stress</th>
<th>IPA</th>
<th>English gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Acha</td>
<td>/ˈatʃa/</td>
<td>‘elder sister’</td>
<td>aCHA</td>
<td>/aˈtʃa/</td>
<td>‘branching’</td>
</tr>
<tr>
<td>2</td>
<td>Ara</td>
<td>/ˈara/</td>
<td>‘fork’</td>
<td>aRA</td>
<td>/aˈra/</td>
<td>‘between’</td>
</tr>
<tr>
<td>3</td>
<td>TÕshük</td>
<td>/ˈtɒʃyk/</td>
<td>‘kitchen’</td>
<td>tôSHŪK</td>
<td>/tɔˈʃyk/</td>
<td>‘hole’</td>
</tr>
<tr>
<td>4</td>
<td>BAla</td>
<td>/ˈbala/</td>
<td>‘child’</td>
<td>baLA</td>
<td>/baˈla/</td>
<td>‘disaster’</td>
</tr>
<tr>
<td>5</td>
<td>CHAtaq</td>
<td>/ˈʧataq/</td>
<td>‘bad branch of tree’</td>
<td>chaTAQ</td>
<td>/ʧaˈtaq/</td>
<td>‘problem’</td>
</tr>
<tr>
<td>6</td>
<td>PACHaq</td>
<td>/ˈpaʧaq/</td>
<td>‘leg’</td>
<td>paCHAQ</td>
<td>/paˈʧaq/</td>
<td>‘piece’</td>
</tr>
</tbody>
</table>
Speakers produced disyllabic words with initial or final stress in a carrier phrase. **Duration** and **intensity** differed based on stress location; **f0 did not!**
Replicated and extended Yakup (2013) and Yakup & Sereno (2016)

Elicited disyllabic stress (near-)minimal pairs in sentence-initial AND sentence-medial position from eight speakers:

Initial: _____ bek yaxshi söz “_____ is a very good word”

Medial: Mahinur _____ deydu “Mahinur will say ______”

Measured duration, intensity, and f0 of both syllables
Duration results

Syllables are longer when they are:

- Stressed
- Word-final
- Sentence-initial

Major & Mayer (2018, in press)
Intensity results

Intensity is higher in syllables when they are:

- Word-final
- Sentence-initial

No effect of stress

Major & Mayer (2018, in press)
**f0 results**

f0 is higher in syllables when they are:

- Word-final
- Sentence-initial

**No effect of stress**

*Major & Mayer (2018, in press)*
Duration is the only acoustic cue consistently associated with stress

No studies have found that f0 correlates with stress

- Different sentence types lead to differences in f0
- Word-final syllables have higher f0
- It is unclear what can be said about intensity
Stress is really complicated in Uyghur

- Stress is often associated with final syllables.
- There is a tendency for non-final stress to associate with heavy syllables.
- When asked, speakers show a lot of variation with respect to where they identify stress.
Major & Mayer (2018; to appear) present an autosegmental metrical model (Pierrehumbert 1980, Beckman & Pierrehumbert 1986) of Uyghur intonation:
Both AP and ip end with a high tone (Ha and H-, respectively)

- Both undergo phrase-final lengthening
- H- final syllables are longer than Ha-final syllables.

A surprising conclusion: Uyghur has stress, but the intonational system does not interact with stress.
Both AP and ip end with a high tone and show final lengthening:

- Stress-conditioned duration in word-final syllables is obfuscated by phrase-final lengthening
- Informally, speakers often select non-final syllables as “the most prominent” despite those syllables being shorter
- Speaker judgments are often inconsistent about prominence

Back to square one:

What is stress in Uyghur and how can we measure it?!!?
Co-speech gestures
WHAT ARE CO-SPEECH GESTURES?

Movements of the hands/arms, head, shoulders, etc. which accompany spoken language and which are **temporally coordinated to speech**

- Occur even in the physical absence of an interlocutor (Wei 2006)
- Are acquired by both blind and sighted speakers (Özçalışkan et al. 2018)
‘Beat gestures’: a subset of co-speech gesture which are tied closely to prosodic structure

• ‘Non-meaningful’

• Conditioned by prosodic structure (Kendon 1980; McNeill 1992)

• Influence speech perception (Leonard & Cummins 2011; Bosker & Peeters 2021)
WHAT ARE CO-SPEECH GESTURES?

Other types of gestures:
- Iconic / depictive
- Deictic
- Conventional
WHAT ARE CO-SPEECH GESTURES?

Other types of gestures:
• Iconic / depictive
• Deictic
• Conventional

• All gestures have prosodically important timing (McClave, 1998; Krahmer & Swerts, 2005, 2007; Loehr, 2004, 2007; Prieto 2018)
• Gestures preferentially aligned with **metrically-prominent** syllables

• The apex (max. extension) of a gesture aligns closely with **pitch peaks** of pitch-accented syllables in several Western European languages

• Gestures preferentially aligned with **metrically-prominent** syllables

• The *apex* (max. extension) of a gesture aligns closely with **pitch peaks** of pitch-accented syllables in several Western European languages

In Uyghur, we have proposed that stress and intonation do not interact (as they do in English).

- Gestures preferentially aligned with *metrically-prominent* syllables.
- The *apex* (max. extension) of a gesture aligns closely with *pitch peaks* of pitch-accented syllables in several Western European languages.

• Gestures preferentially aligned with **metrically-prominent** syllables

• The **apex** (max. extension) of a gesture aligns closely with **pitch peaks** of pitch-accented syllables in several Western European languages

• In Uyghur, we have proposed that stress and intonation **do not interact** (as they do in English)

• Gestures tend to gravitate to rhythmically-prominent syllables cross-linguistically

---

HOW DOES ALIGNMENT WORK IN TURKIC LANGUAGES?

Prediction:

Gestures in Uyghur will be attracted to stressed syllables (as reflected in duration), rather than pitch peaks (which mark phrase boundaries, rather than rhythmic prominence)

• Gestures preferentially aligned with *metrically-prominent* syllables

• The apex (max. extension) of a gesture aligns closely with *pitch peaks* of pitch-accented syllables in several Western European languages


They’re BEAUTiful!
GESTURE CORPORAS

- Three Uyghur speakers video- and audio-recorded (accessed via wikitongues through YouTube) describing aspects of Uyghur language and culture

- Audio data transcribed and force-aligned at word-, syllable- and phone-levels using Montreal Forced Aligner (McAuliffe et al. 2017) and subsequently checked for alignment accuracy

- ~3200 phones total
- 165 gesture-accompanied phones (more data collection currently in progress!)
GESTURE PHASES AND ANNOTATION

Pre-Gesture

Preparation

Stroke Onset

Stroke Apex
Consonant-aligning apex

Time

Vowel Onset

Vowel-aligning apex
Co-speech gestures in Uyghur
Results: Apex Timing
Gesture rates across speakers
Question: What syllable features predict a co-occurring gesture?

Logistic regression model fit to all syllables

\[ \text{gesture} \sim \text{syl. weight} \times \text{position} + \text{maxf0} + \text{intensity} + \text{duration} + (1|\text{speaker}) + (1|\text{syllable}) \]
Gestures prefer heavy syllables
Gestures do not prefer final position
Position x weight

Two syllable words by syllable weight

Three syllable words by syllable weight

Proportion of syllables with gestures
Gestures are not predicted by f0…
Gestures are not predicted by intensity…
But syllables with gestures are longer
Gestures prefer to be associated with heavy syllables

Gestures have no preference for final position

Syllables with gestures are longer
**Question:** What factors predict syllable duration?

**Linear regression** model fit to all syllables

duration ~ syl. weight + position * maxf0 + intensity + gesture * position + (1|syllable)
Heavy syllables are longer
Duration: final > initial > medial
Non-final syllables with gestures are longer
f0 and duration only correlated in final position
Takeaway points

- Heavy syllables are longer
- Final syllables > initial syllables > medial syllables
- Syllables with gestures are longer (weakest in final position)
- f0 and duration covary only at word boundaries
  - Correlates of prosodic boundaries (Major & Mayer 2018, in press)
Why are gestured syllables longer?

Do gestures preferentially align with longer syllables?

Or does an aligned gesture cause lengthening?
Compare only syllables that occur in tokens of the same word with and without gestures

**Paired t-test** between mean duration for each syllable/word in gestured/ungestured tokens shows no difference ($t(59) = |1.32|$, $p = 0.19$)
Paired gestured/ungestured syllables
Takeaway points

No significant difference in duration between syllables in different tokens of the same word with/without gestures

Supports claim that gestures gravitate to long syllables rather than lengthening them
Discussion & Conclusion
Consistent with past accounts, f0 and duration do not covary

- Except in word-final position where both are correlates of prosodic boundary strength

Syllables undergo word-final lengthening
Gestures are attracted to **heavy syllables** and **longer syllables**

Pitch, intensity, and position in word did not predict gesture location

Consistent with account where stress in Uyghur is

- attracted to heavy syllables
- signaled primary by durational differences
Provides further evidence that gesture timing is determined primarily by *rhythmic factors*, rather than pitch ‘prominence’
Outstanding questions

Data from more speakers

Evaluate between different accounts of heavy syllable sensitivity?

Method for eliciting better stress judgments?
Future directions

- Look at more data in Uyghur
  - Targeted elicitation of particle constructions
  - Look at pragmatic properties of the gestures we’ve already coded.
  - Controlled experimental investigation?
- Expand to other Turkic languages
Thanks to Gulnisa Nazarova and Mustafa Aksu for permission to use the YouTube video and images. We would also like to thank the Harvard PhonLab for feedback.
Extra slides…
US English: Probabilistic relationship between pitch accent type and gesture presence

Im & Baumann 2020
US English: Probabilistic relationship between *pitch accent type* and *gesture presence*

Gesture Less Likely

Gesture More Likely

L* !H* H* L-H*

*Im & Baumann 2020*
US English: Probabilistic relationship between **pitch accent type** and **gesture presence**

*Why?*

<table>
<thead>
<tr>
<th>L*</th>
<th>!H*</th>
<th>H*</th>
<th>L-H*</th>
</tr>
</thead>
</table>

Gesture Less Likely  Gesture More Likely

*Im & Baumann 2020*
US English: Probabilistic relationship between pitch accent type and gesture presence

Why?

Associated w/ Given Information

L* !H* H* L-H*

Associated Discourse-New/Unused Information

Gesture Less Likely Gesture More Likely

Pierrehumbert & Hirschberg 1990; Baumann & Riester 2013, Im & Baumann 2020
‘PROMINENCE’ AND CO-SPEECH GESTURE

US English: Probabilistic relationship between **pitch accent type** and **gesture presence**

*Why?*

**Less perceptually prominent**

- L*
- !H*

**More perceptually prominent**

- H*
- L-H*

**Gesture Less Likely**

Baumann & Roehr 2015
<table>
<thead>
<tr>
<th>as_factor(gesture)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

The table contains rows with various entries, but the content is not clearly legible due to the quality of the image.