featurality as a byproduct of script inheritance

{Nikita Bezrukov, Ronan Soni}
Princeton University

AWLL14 . Temple University Rome .
10 Nov 2023

Introduction

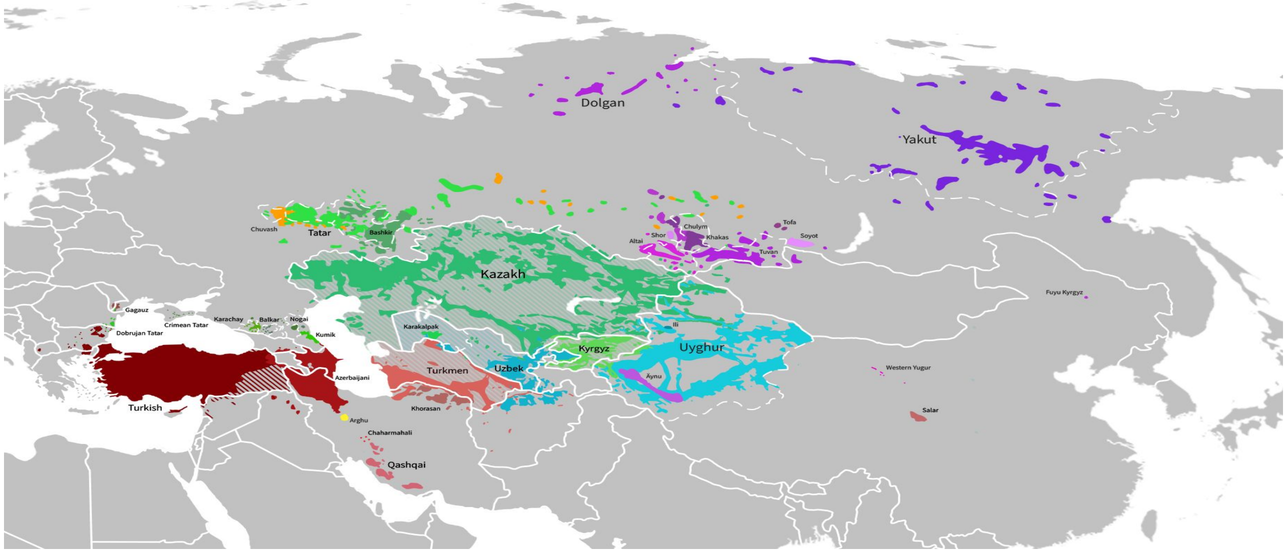
- Featurality:
  - Encountered most often in discussions of Hangeul
    - Feature marking was present from the script's creation
  - However, frequent elsewhere
    - Seems to often appear due to adaptation of an existing script to a new phonological system

Introduction: Featurality

- Contrasts in **manner** and **degree** (compare to morphography and phonography)
    - For the former, compare the construction of syllables (or syllable-level structures) from phonemes to the construction of phonemes from features
    - For the latter, note that featurality and morphography are alike in that they cannot alone comprise a system

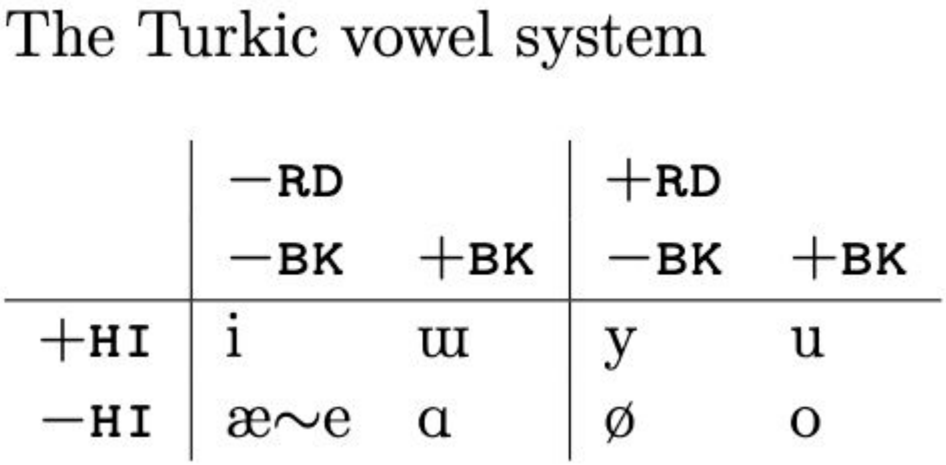
Turkic

- This talk: a sample of orthographies for Turkic
  - Large language family, lots of standardized languages, rich history of writing in different scripts



Turkic: Core vocalic contrasts

- We’ll focus on how the core vocalic contrasts are encoded
    - 3 features (height, backness, roundedness), 8 vowels

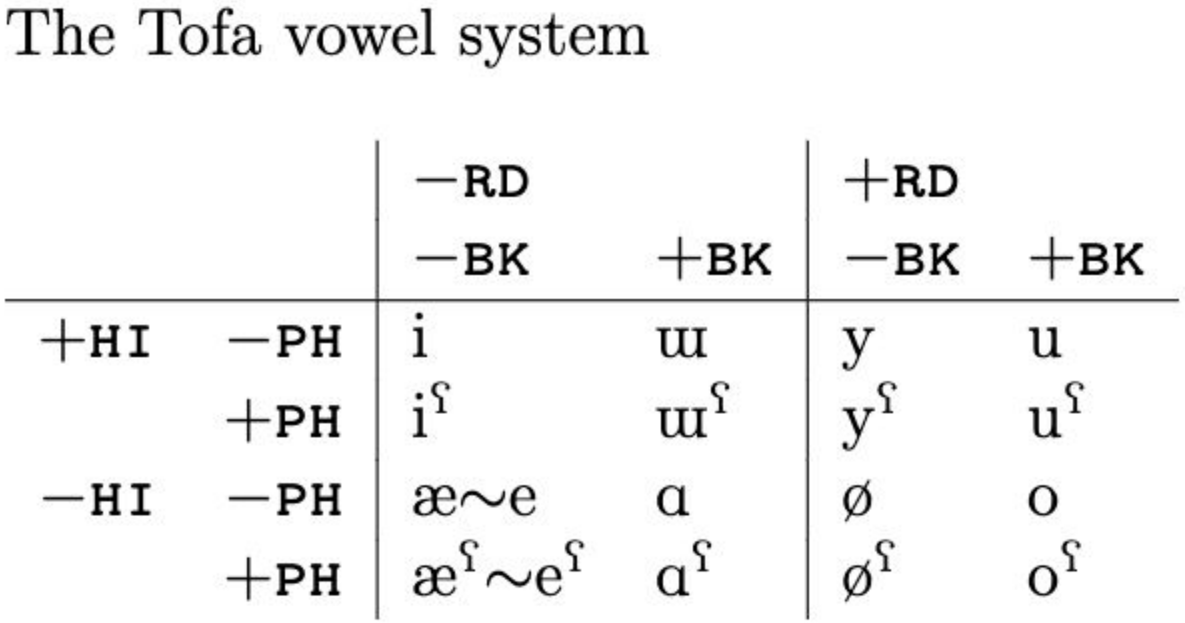


The Turkic vowel system

	−RD		+RD	
	−BK	+BK	−BK	+BK
+HI	i	ɯ	y	u
−HI	æ∼e	ɑ	ø	o

Turkic: Tuvan-Tofa

- Tuvan and Tofa have a fourth feature, pharyngealization (will treat it on par, 16 vowels)

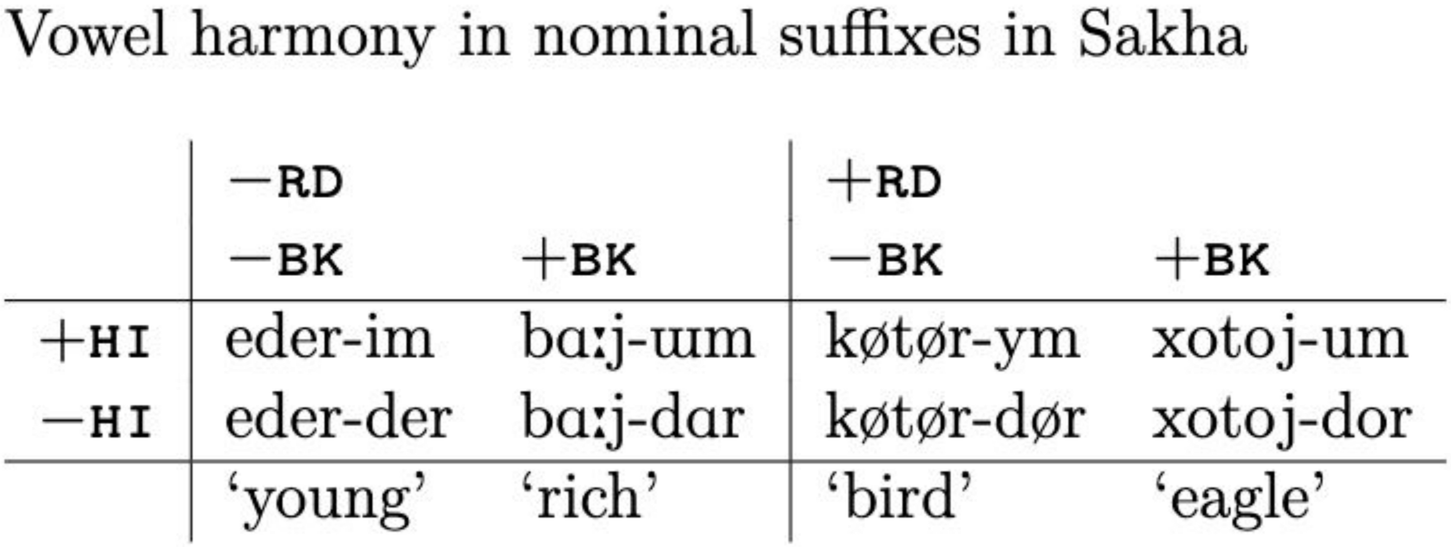


The Tofa vowel system

		−RD		+RD	
		−BK	+BK	−BK	+BK
+HI	−PH	i	ɯ	y	u
	+PH	i^{ʕ}	ɯ^{ʕ}	y^{ʕ}	u^{ʕ}
−HI	−PH	æ∼e	ɑ	ø	o
	+PH	æ^{ʕ}∼e^{ʕ}	ɑ^{ʕ}	ø^{ʕ}	o^{ʕ}

Turkic: Vowel harmony

- Backness and roundedness are spread dynamically through vowel harmony: a marking bias?



Vowel harmony in nominal suffixes in Sakha

	−RD		+RD	
	−BK	+BK	−BK	+BK
+HI	eder-im	bɑːj-ɯm	køtør-ym	xotoj-um
−HI	eder-der	bɑːj-dɑr	køtør-dør	xotoj-dor
	‘young’	‘rich’	‘bird’	‘eagle’

Turkic

- Hypotheses:
  - We investigate the interplay between
    - phonological (parameters of harmony)
    - orthographic factors (script inheritance)
  - in the degree of featurality of an orthographic system

Data: sample of 76 orthographies from Turkic (19-21 centuries)

Script inheritance plays a pivotal role in featurality.

Model

- This talk: limited to alphabets (i.e., phonographic, segments are all represented, and combined linearly)

- We’re concerned with **orthographic spell-out**:
  - i.e., mapping phonological feature bundles to orthographic characters.

- Typical orthographic spell-out is **segmental**.

Model

- **Segmental** orthographic spell-out:

$$\left[\begin{array}{c} +\ \alpha \\ -\ \beta \\ \hline +\ \gamma \end{array}\right] \xrightarrow{\text{segmental}} \text{a.}$$

$$\left[\begin{array}{c} +\ \alpha \\ -\ \beta \\ \hline -\ \gamma \end{array}\right] \xrightarrow{\text{segmental}} \text{e.}$$

Model

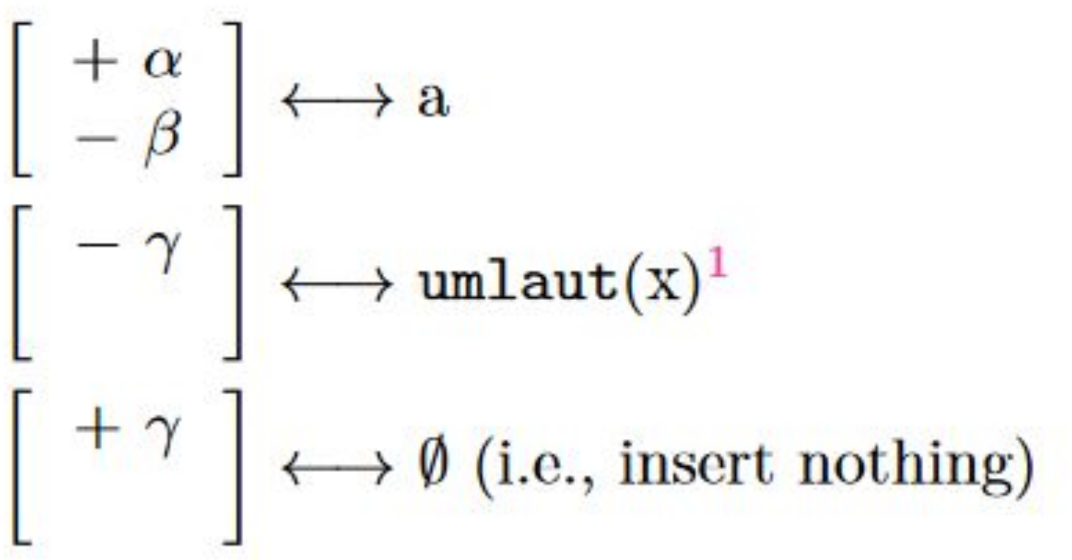
- We’re interested in **sub-segmental** orthographic spell out:

$$\left[\begin{array}{c} +\,\alpha \\ -\,\beta \\ \hline +\,\gamma \end{array}\right] \xrightarrow{\text{sub-segmental}} \mathrm{a}\left[\;+\,\gamma\;\right] \xrightarrow{\text{sub-segmental}} \mathrm{a}.$$

$$\left[\begin{array}{c} +\,\alpha \\ -\,\beta \\ \hline -\,\gamma \end{array}\right] \xrightarrow{\text{sub-segmental}} \mathrm{a}\left[\;-\,\gamma\;\right] \xrightarrow{\text{sub-segmental}} \ddot{\mathrm{a}}.$$

Model

- Grammar (rewrite rules, featuring subsets):



$$\begin{bmatrix} +\,\alpha \\ -\,\beta \end{bmatrix} \longleftrightarrow \text{a}$$

$$\begin{bmatrix} -\,\gamma \\ \end{bmatrix} \longleftrightarrow \texttt{umlaut(x)}^{1}$$

$$\begin{bmatrix} +\,\gamma \\ \end{bmatrix} \longleftrightarrow \emptyset \text{ (i.e., insert nothing)}$$

Model: Degree of featurality

- Degree of featurality:
  - First pass:
    - Measure length in steps of the derivation of the corresponding symbol from each phoneme
    - Segmental systems: minimum featurality (1)

$$\begin{bmatrix} +\alpha \\ -\beta \\ -\gamma \end{bmatrix} \xrightarrow{\text{sub-segmental}} \text{a}\begin{bmatrix} +\alpha \\ -\beta \end{bmatrix} \xrightarrow{\text{sub-segmental}} \text{ä}\begin{bmatrix} +\alpha \end{bmatrix} \xrightarrow{\text{sub-segmental}} \text{ä'}.$$

Model: Degree of featurality

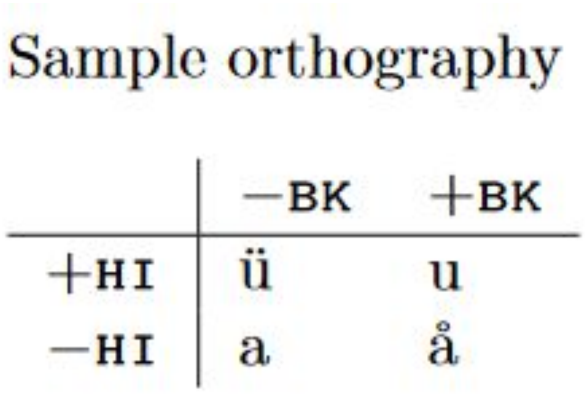
- Intermediate case (2)

$$\left[\begin{array}{c} +\ \alpha \\ -\ \beta \\ \hline +\ \gamma \end{array}\right] \xrightarrow{\text{sub-segmental}} \mathrm{a}\left[\ +\ \gamma\ \right] \xrightarrow{\text{sub-segmental}} \mathrm{a}.$$

$$\left[\begin{array}{c} +\ \alpha \\ -\ \beta \\ \hline -\ \gamma \end{array}\right] \xrightarrow{\text{sub-segmental}} \mathrm{a}\left[\ -\ \gamma\ \right] \xrightarrow{\text{sub-segmental}} \ddot{\mathrm{a}}.$$

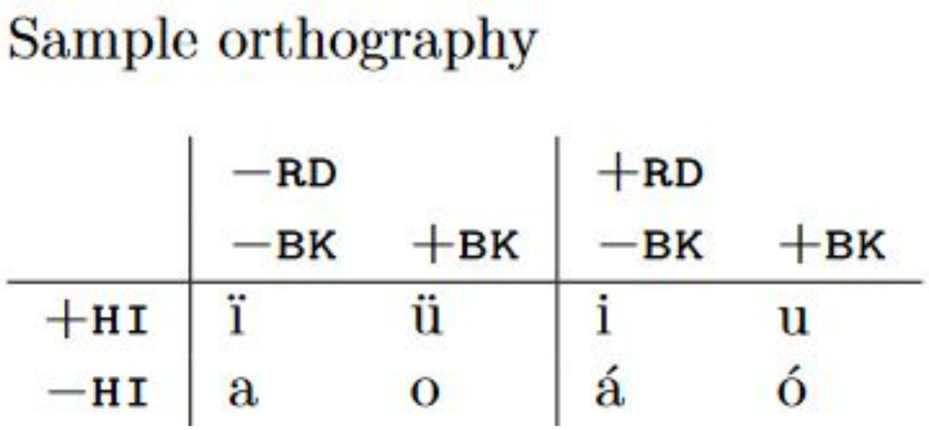
Model: Degree of featurality

- Possible shortcomings:
  - How to explain these systems?
  - Solution: introduce residuals to derivation steps



Sample orthography

	−BK	+BK
+HI	ü	u
−HI	a	å



Sample orthography

	−RD		+RD	
	−BK	+BK	−BK	+BK
+HI	ï	ü	i	u
−HI	a	o	á	ó

Model: Degree of featurality

- This resolves some of the issues with regards to this system
  - ○ To match intuition, a step which leaves a residual feature could be considered a half-step

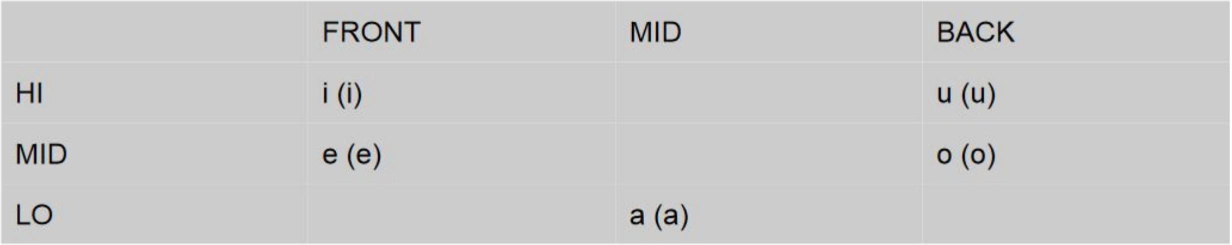
$$\begin{bmatrix} +\ \text{RD} \\ -\ \text{BK} \\ +\ \text{HI} \end{bmatrix} \xrightarrow{\text{sub-segmental}} \text{u}\begin{bmatrix} \text{- BK} \\ (\text{res})\ +\ \text{HI} \end{bmatrix} \xrightarrow{\text{sub-segmental}} \text{ü}.$$

Orthographies

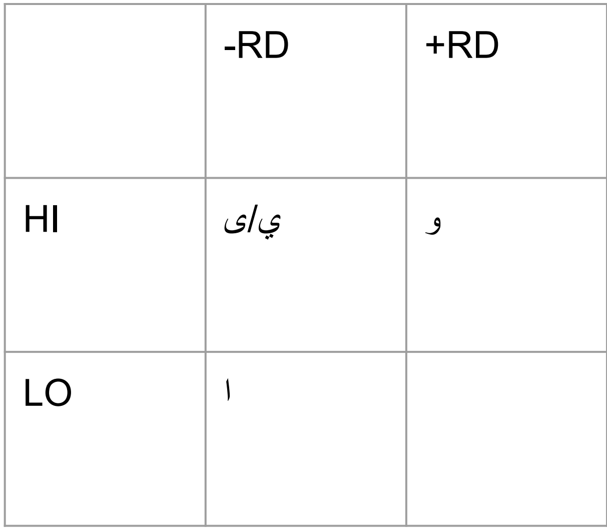
- The basis of the development of featurality seems to be the inheritance of the script, and featural properties are dependent on the script origin.
    - Therefore we can model script inheritance based on this observation:
        - In the first step, glyphs with “predetermined” values are assigned
        - Then, the system is extended to the complete inventory of the language
        - This generates a set of derivation rules

Initial Step

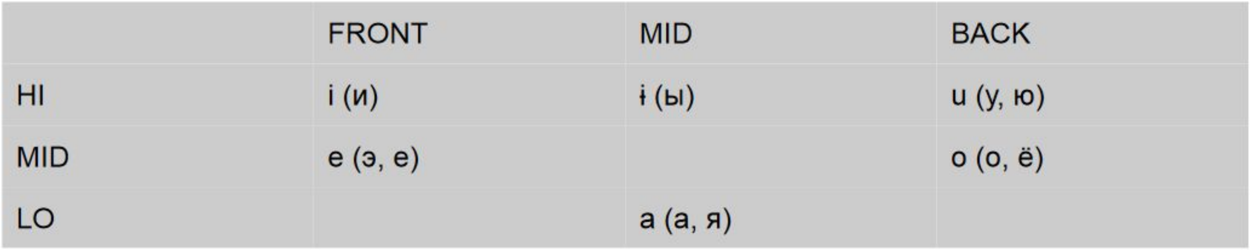
- The main scripts Turkic languages use are Latin, Cyrillic, and Arabic, from which are often taken some variation of these initial values:



	FRONT	MID	BACK
HI	i (i)		u (u)
MID	e (e)		o (o)
LO		a (a)	



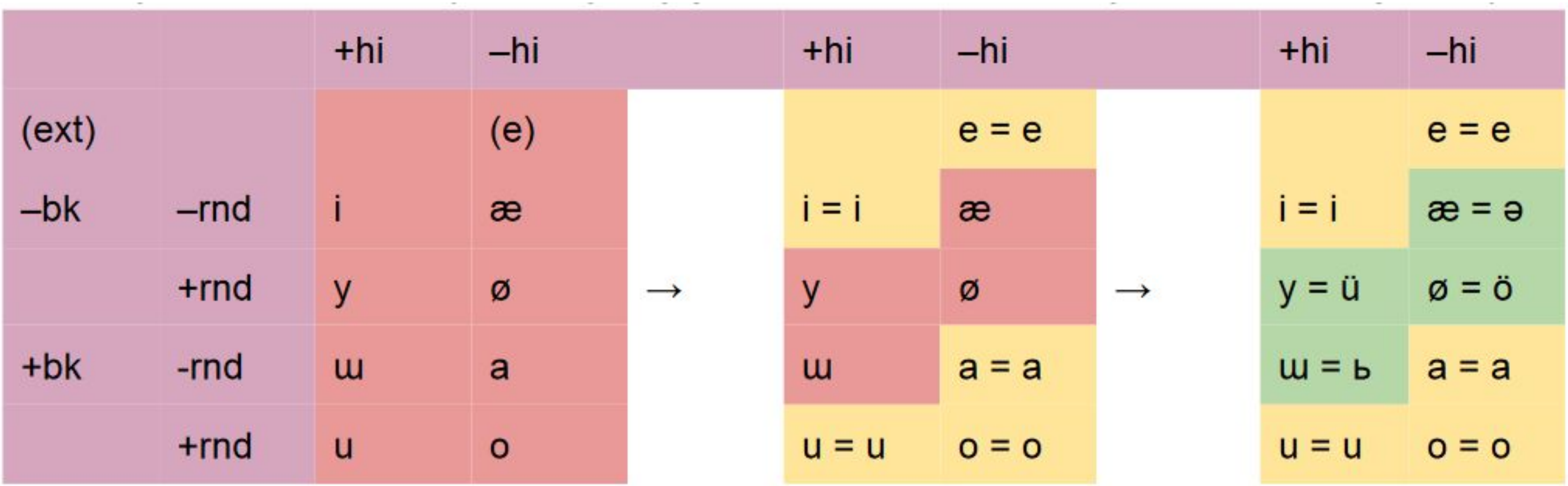
	-RD	+RD
HI	ي/ى	و
LO	ا	



	FRONT	MID	BACK
HI	i (и)	ɨ (ы)	u (у, ю)
MID	e (э, е)		o (о, ё)
LO		a (а, я)	

Orthographies: Azerbaijani (1929)

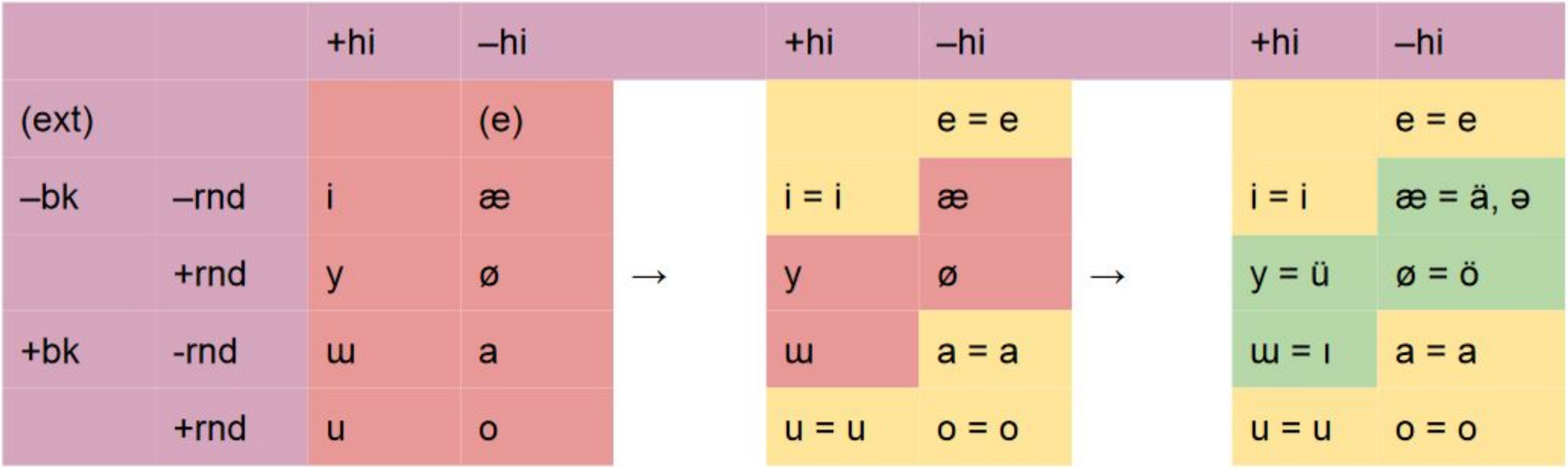
- The additional symbols are taken from multiple sources, with some containing featural elements



		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			(e)			e = e			e = e
–bk	–rnd	i	æ		i = i	æ		i = i	æ = ə
	+rnd	y	ø	→	y	ø	→	y = ü	ø = ö
+bk	-rnd	ɯ	a		ɯ	a = a		ɯ = ь	a = a
	+rnd	u	o		u = u	o = o		u = u	o = o

Orthographies: Malqar (1994)

- Here the additional symbols are more generalized



		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			(e)			e = e			e = e
–bk	–rnd	i	æ		i = i	æ		i = i	æ = ä, ə
	+rnd	y	ø	→	y	ø	→	y = ü	ø = ö
+bk	-rnd	ɯ	a		ɯ	a = a		ɯ = ı	a = a
	+rnd	u	o		u = u	o = o		u = u	o = o

Orthographies: Other Latin

		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			(e)			e			e = ė
–bk	–rnd	i	æ		i = i	æ		i = i	æ = e
	+rnd	y	ø	→	y	ø	→	y = ü	ø = ö
+bk	-rnd	ɯ	a		ɯ	a = a		ɯ = y	a = a
	+rnd	u	o		u = u	o = o		u = u	o = o

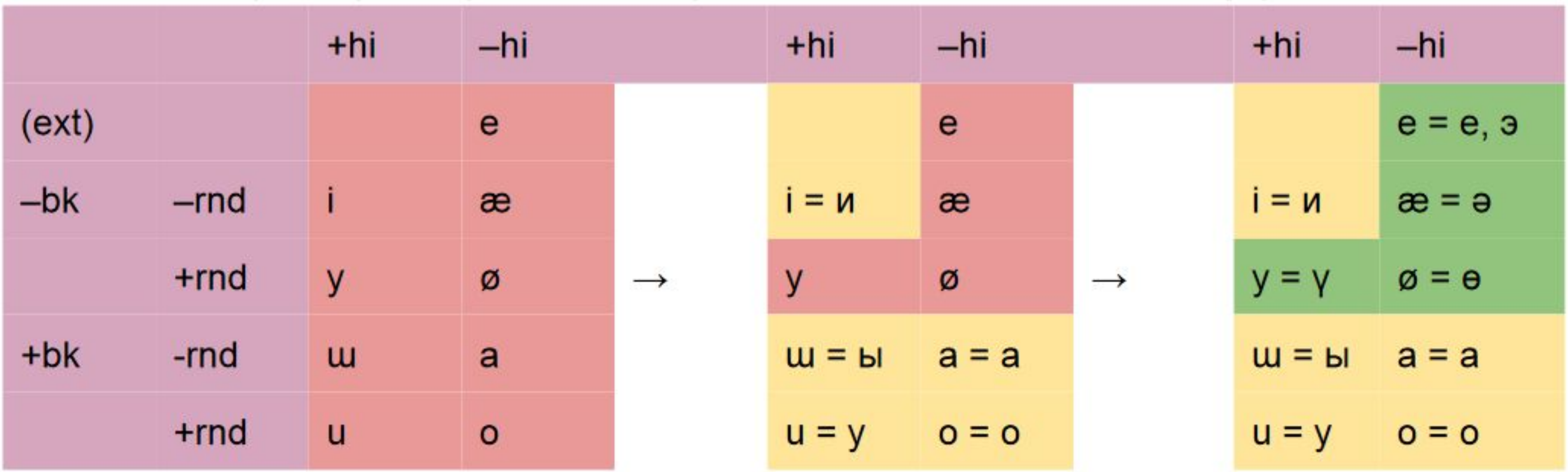
		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			(e)			e = e			e = e
–bk	–rnd	i	æ		i = i	æ		i = i	æ = ä, ə
	+rnd	y	ø	→	y	ø	→	y = ü	ø = ö
+bk	-rnd	ɯ	a		ɯ	a = a		ɯ = y	a = a
	+rnd	u	o		u = u	o = o		u = u	o = o

		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			(e)			e = e			e = e
–bk	–rnd	i	æ		i = i	æ		i = i	æ = ea
	+rnd	y	ø	→	y	ø	→	y = v	ø = q
+bk	-rnd	ɯ	a		ɯ	a = a		ɯ = x	a = a
	+rnd	u	o		u = u	o = o		u = u	o = o

- The systems surveyed display variation in many symbols, but others are effectively universal

Orthographies: Azerbaijani

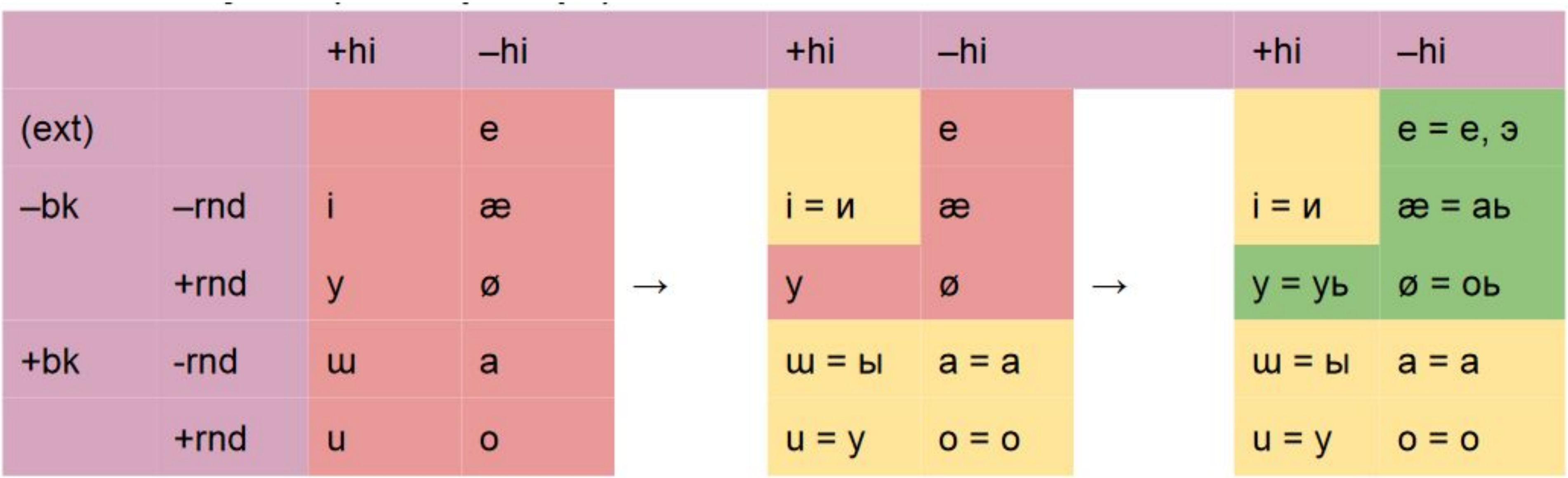
- A common type of system, with some light featural elements



		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			e			e			e = e, э
–bk	–rnd	i	æ		i = и	æ		i = и	æ = ə
	+rnd	y	ø	→	y	ø	→	y = ү	ø = ө
+bk	-rnd	ɯ	a		ɯ = ы	a = a		ɯ = ы	a = a
	+rnd	u	o		u = у	o = o		u = у	o = o

Orthographies: Qumuq-Malqar

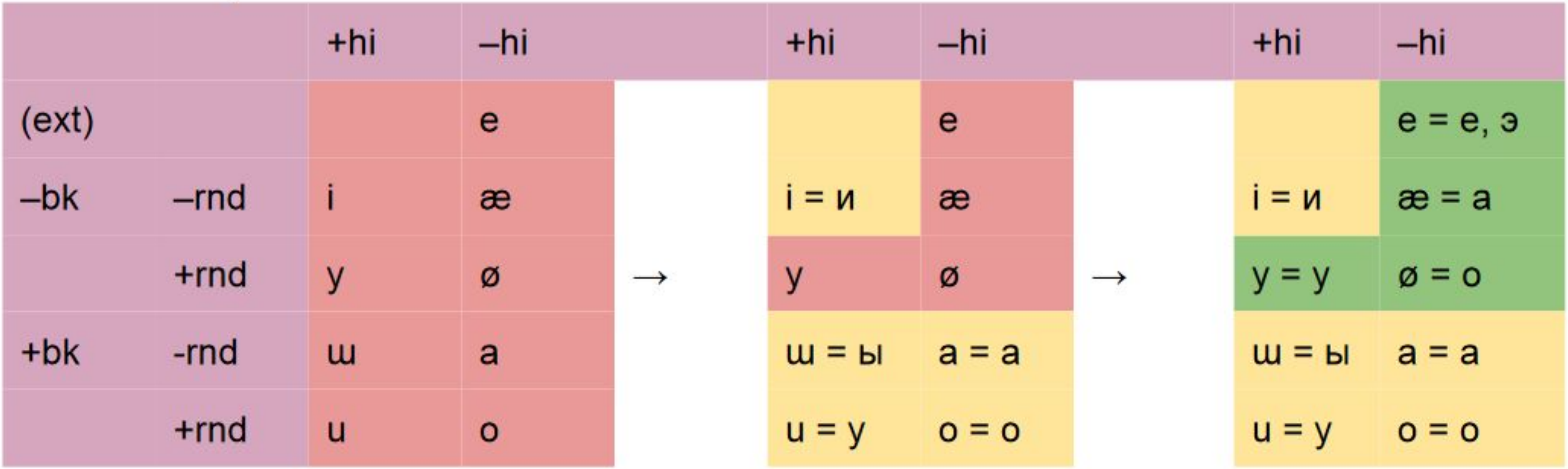
- The additional symbols are filled in using a linearly combining glyph. Tuvan and Tofa use this system but with another glyph for pharyngealization!



		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			e			e			e = e, э
–bk	–rnd	i	æ		i = и	æ		i = и	æ = аь
	+rnd	y	ø	→	y	ø	→	y = уь	ø = оь
+bk	-rnd	ɯ	a		ɯ = ы	a = a		ɯ = ы	a = a
	+rnd	u	o		u = y	o = o		u = y	o = o

Orthographies: Qırım (1938)

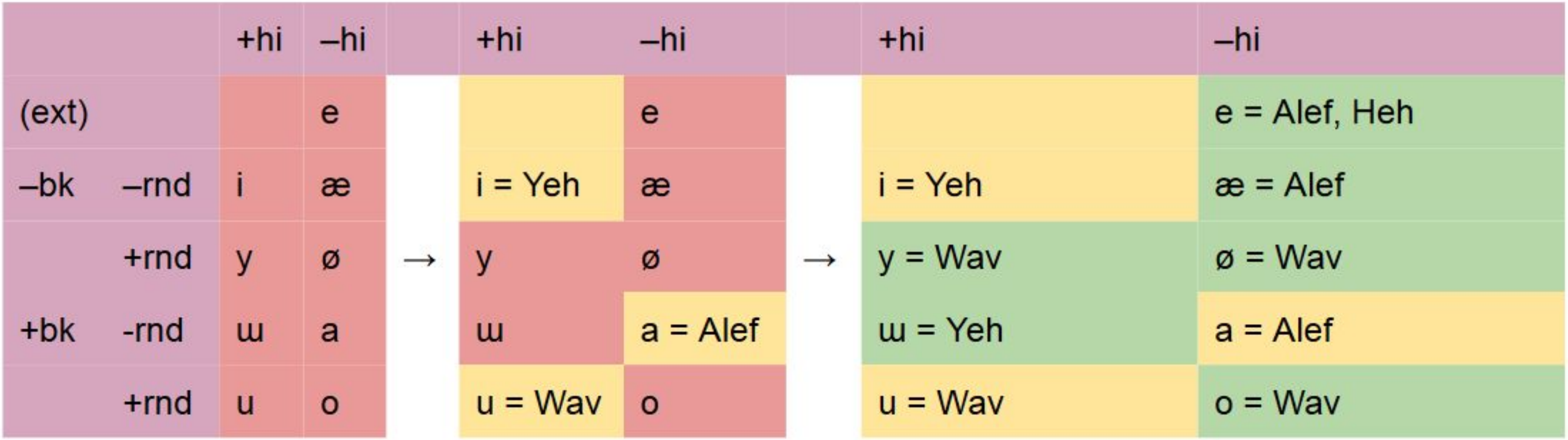
- This system uses underrepresentation of frontness/backness - but it is still represented when there is a distinction in the inherited symbols



		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			e			e			e = е, э
–bk	–rnd	i	æ		i = и	æ		i = и	æ = а
	+rnd	y	ø	→	y	ø	→	y = у	ø = о
+bk	-rnd	ɯ	a		ɯ = ы	a = а		ɯ = ы	a = а
	+rnd	u	o		u = у	o = о		u = у	o = о

Orthographies: Qırım (1921)

- This system adapts the Arabic script, and so it shows an even higher degree of featural marking particularly through underspecification



		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			e			e			e = Alef, Heh
–bk	–rnd	i	æ		i = Yeh	æ		i = Yeh	æ = Alef
	+rnd	y	ø	→	y	ø	→	y = Wav	ø = Wav
+bk	-rnd	ɯ	a		ɯ	a = Alef		ɯ = Yeh	a = Alef
	+rnd	u	o		u = Wav	o		u = Wav	o = Wav

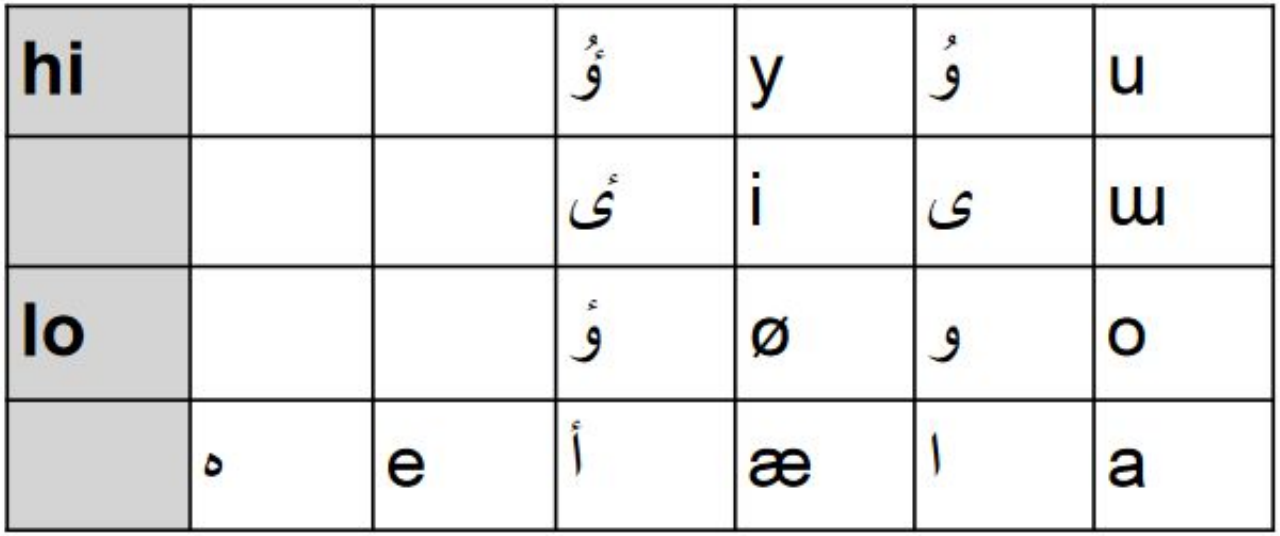
Orthographies: Qumuq (1921)

- This system avoids underspecification for rounded vowels only, through 2-feature diacritics

		+hi	–hi		+hi	–hi		+hi	–hi
(ext)			e			e			e = Heh
–bk	–rnd	i	æ		i = Yeh	æ		i = Yeh	æ = Alef
	+rnd	y	ø	→	y	ø	→	y = Wav+dot	ø = Wav+two-dots
+bk	-rnd	ɯ	a		ɯ	a = Alef		ɯ = Yeh	a = Alef
	+rnd	u	o		u = Wav	o		u = Wav	o = Wav+circumflex

Orthographies: Kazakh (1920s)

- This system makes systematic use of Arabic diacritics to disambiguate: the ḍamma (ʊ) indicates [+high], and the hamza (ʔ) is [-back]



hi			ؤُ	y	وُ	u
			ئ	i	ى	ɯ
lo			ؤ	ø	و	o
	ه	e	أ	æ	ا	a

Takeaway

- Featurality:
  - Arises as a byproduct of script inheritance
  - Can be measured and formally modelled
  - Exists in various configurations that can be explored in large data sets (the comparative cases we introduced)
- Future work:
  - Extending to other scripts and crucially consonants
  - Refinement of the calculation of degree

References

DeFrancis, John. *Visible Speech: The Diverse Oneness of Writing Systems*. University of Hawaii Press, 1989.

Gnanadesikan, Amalia. “Towards a typology of phonemic scripts”*. Writing Systems Research*, 2017, vol. 9, no. 1, pp. 14–35

MacMahon, Michael. “Phonetic Notation.” *The World’s Writing Systems*, edited by Peter Daniels and William Bright, Oxford University Press, 1996, pp. 821–46.

Rogers, Henry. *Writing Systems: A Linguistic Approach*. Blackwell Pub, 2005.

Sampson, Geoffrey. Writing Systems. Stanford University Press, 1985.

Sproat, Richard, and Alexander Gutkin. “The Taxonomy of Writing Systems: How to Measure How Logographic a System Is.” *Computational Linguistics*, vol. 47, no. 3, Nov. 2021, pp. 477–528.

Sproat, Richard William. *A Computational Theory of Writing Systems*. Cambridge University Press, 2000.