Information-Theoretic Characterization of Vowel Harmony: A Cross-Linguistic Study on Word Lists

Julius Steuer[1],Badr Abdullah[1], Johann-Mattis List[2], Dietrich Klakow[1]

[1]Language Science and Technology, Saarland University
[2]MPI-EVA / Univ. of Passau

SIGTYP @ EACL 2023

IDeaL SFB 1102

UNIVERSITÄT DES SAARLANDES

Vowel Harmony

- Constraint on the vowels in a word form
- Vowels need to agree w. r. t. a feature
- Backness (Finnish, Hungarian, Turkish, Korean), Roundness (Turkish, Mongolian), Nasality (Guaraní), Tongue root position (Mongolian)
- Surfaces mainly in inflectional and derivational morphology

Vowel Harmony

- Constraint on the vowels in a word form
- Vowels need to agree w. r. t. a feature
- Backness (Finnish, Hungarian, Turkish, Korean), Roundness (Turkish, Mongolian), Nasality (Guaraní), Tongue root position (Mongolian)
- Surfaces mainly in inflectional and derivational morphology

Turkish Front, Back harmony:

Genitive	-In	= [ɯn]/[un]/[yn]/[in]
Plural	-lAr	= [lɛr]/[lar]

Vowel Harmony

- Constraint on the vowels in a word form
- Vowels need to agree w. r. t. a feature
- Backness (Finnish, Hungarian, Turkish, Korean), Roundness (Turkish, Mongolian), Nasality (Guaraní), Tongue root position (Mongolian)
- Surfaces mainly in inflectional and derivational morphology

Turkish Front, Back harmony:

Genitive -In = [ɯn]/[un]/[yn]/[in]

Plural -lAr = [lɛr]/[lar]

[kɯz] ‘girl’ + -In (genitive) = [kɯzɯn]

[jyz] ‘face’ + -In (genitive) = [jyzyn]

Vowel Harmony

- Constraint on the vowels in a word form
- Vowels need to agree w. r. t. a feature
- Backness (Finnish, Hungarian, Turkish, Korean), Roundness (Turkish, Mongolian), Nasality (Guaraní), Tongue root position (Mongolian)
- Surfaces mainly in inflectional and derivational morphology

Turkish Front, Back harmony:

Genitive -In = [ɯn]/[un]/[yn]/[in]

Plural -lAr = [lɛr]/[lar]

[kɯz] ‘girl’ + -In (genitive) = [kɯzɯn] + -lAr (plural) = [kɯzlarɯn]

[jyz] ‘face’ + -In (genitive) = [jyzyn] + -lAr (plural) = [jyzlɛrin]

Motivation

Previous Work:

- Quantify vowel harmony based on estimates from large corpora of inflected word forms
  - E.g. Goldsmith & Riggle (2012), Baker (2009), Mayer et al. (2010)
- Requires specific and on a specific type of data (precompiled lists of inflected word forms, running text)
  - Cannot be applied to low-resource languages

Motivation

Previous Work:

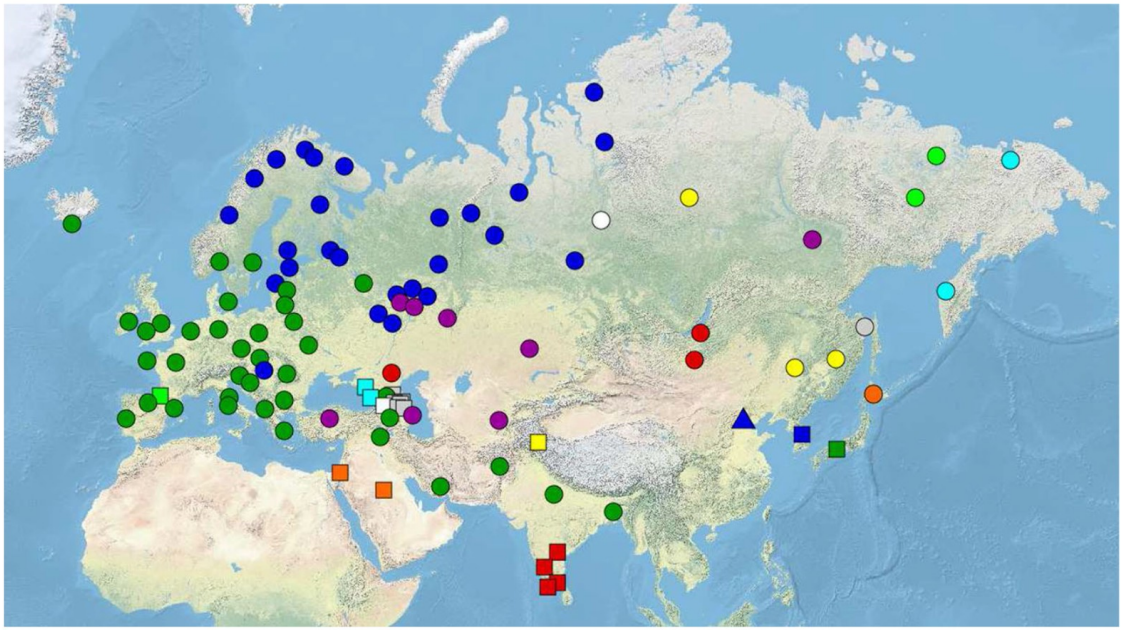
- Quantify vowel harmony based on estimates from large corpora of inflected word forms
  - E.g. Goldsmith & Riggle (2012), Baker (2009), Mayer et al. (2010)
- Requires specific and on a specific type of data (precompiled lists of inflected word forms, running text)
  - Cannot be applied to low-resource languages

Our Approach:

- Use concept-based word lists instead of large corpora
  - (Recover vowel harmony even if it is no more present in inflectional morphology)
- NLMs are a smart way to parametrize a probability distribution
  - Quantify vowel harmony based on the performance of a NLM

Data

NorthEuraLex



Language Family	#
Indo-European	37
Uralic	26
Turkic	8
Nakh-Daghestanian	6
Dravidian	4
Eskimo-Aleut, Mongolic, Tungusic	3
Afro-Asiatic, Abkhaz-Adyge, Chukotko-Kamchatkan, Yukaghir	2
Nivkh, Ainu, Koreanic, Japonic, Burushaski, Kartvelian, Basque, Yeniseian, Sino-Tibetan	1

- Dataset by Dellert et al. (2021), Lexibank (List et al. 2022) version by Dellert (2021)
- Concept-based word lists
- 107 Languages, 9 language families (without isolates)
- 677 (Italian) to 1513 (Manchu) lemmata per language

NorthEuraLex

No.	Name	English	German	Russian	Concepticon
Search	Search	Search	Search	Search	Search
1	EYE	eye [[anatomy]]	Auge [[Anatomie]]	глаз [[анатомия]]	EYE
2	EAR	ear [[anatomy]]	Ohr [[Anatomie]]	ухо [[анатомия]]	EAR
3	NOSE	nose [[anatomy]]	Nase [[Anatomie]]	нос [[анатомия]]	NOSE
4	MOUTH	mouth [[anatomy]]	Mund [[Anatomie]]	рот [[анатомия]]	MOUTH
5	TOOTH	tooth [EX:human incisor]	Zahn [BSP:menschlicher Schneidezahn]	зуб [НАПР:человека]	TOOTH
6	TONGUE	tongue [[anatomy]]	Zunge [[Anatomie]]	язык [орган в полости рта]	TONGUE
7	LIP	lip [[anatomy]]	Lippe [[Anatomie]]	губа [[анатомия]]	LIP
8	CHEEK	cheek [[anatomy]]	Wange [[Anatomie]]	щека [[анатомия]]	CHEEK
9	FACE	face [of a human]	Gesicht [des Menschen]	лицо [человека]	FACE
10	FOREHEAD	forehead [of a human]	Stirn [des Menschen]	лоб [человека]	FOREHEAD
11	HAIR	hair [of human head]	Haar [Kopfhaar des Menschen]	волос [на голове человека]	HAIR
12	MOUSTACHE	moustache [of a man]	Schnurrbart [eines Mannes]	усы [мужчины]	MOUSTACHE
13	BEARD	beard [generic]	Bart [allgemein]	борода [волосяной покров нижней части лица]	BEARD
14	CHIN	chin [[anatomy]]	Kinn [[Anatomie]]	подбородок [[анатомия]]	CHIN

Language Sample

- Subset of NorthEuraLex
- 5 languages with vowel harmony
- 5 languages without vowel harmony
- "Harmonic Groups" defined by features

Language	Harmonic Groups		
Finnish	−BACK {y, ø, æ}	+BACK {u, o, ɑ}	BACK neutral {e, i}
Hungarian	−BACK {y, ø}	+BACK {u, o, ɒ}	BACK neutral {e, i}
Manchu	−BACK {e/ɤ}	+BACK {ɑ, ɔ}	BACK neutral {i,u}
Khalkha Mongolian	−ATR {e, u, o} −ROUND {e, a, i}	+ATR {a, ʊ, o} +ROUND {o}	ATR neutral {i} ROUND neutral {u, ʊ}
Turkish	−BACK {i, e, y, œ} −ROUND {i, e, u, o}	+BACK {ɯ, a, u, o} +ROUND {ɯ, a, y, œ}	BACK neutral - ROUND neutral -
Arabic			
Ainu			
Estonian			
Armenian			
Basque			

Methodology

Feature Surprisal from Word Lists

- Average feature surprisal over vowel positions *t* for harmonic group *H* in word list *W:*

$$\overline{\eta}(\mathcal{H}) = -\frac{1}{|W|}\sum_{w \in W}\sum_{t \in \{\tau,\cdots,T\}}\eta(\mathcal{H},t)$$

- Analogical for disharmonic vowels
- Relative strength of vowel harmony is indicated by the difference in average feature surprisal:

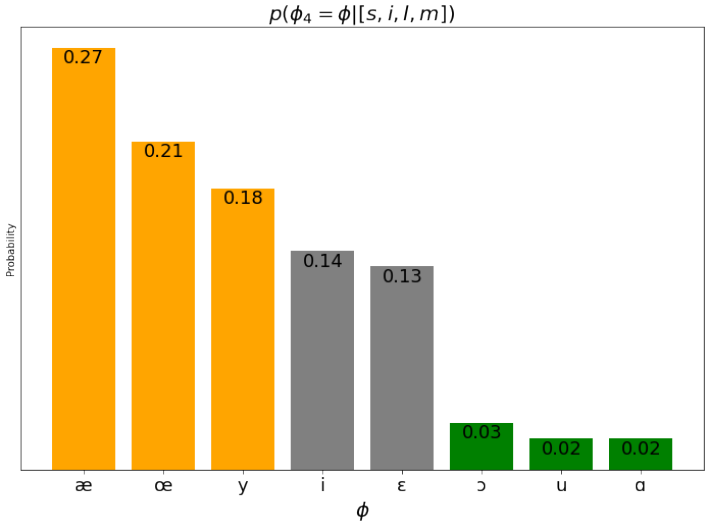
$$\Delta_\eta = \overline{\eta}(\mathcal{H}) - \overline{\eta}(\neg\mathcal{H})$$

- Difference quantifies the relative strength of the vowel harmony constraint
  - Diff $\gg 0$: strong vowel harmony constraint
  - Diff ≈ 0: no vowel harmony
  - Diff $\ll 0$: should not occur

Example

Phoneme surprisal

- Finnish *silmässä* [silmæs:æ] ‘eye (locative)’
- [i] triggers –BACK harmony
- First vowel [i] is ignored (no context)
- Surprisal at [æ]: -log(0.27) = 1.8889

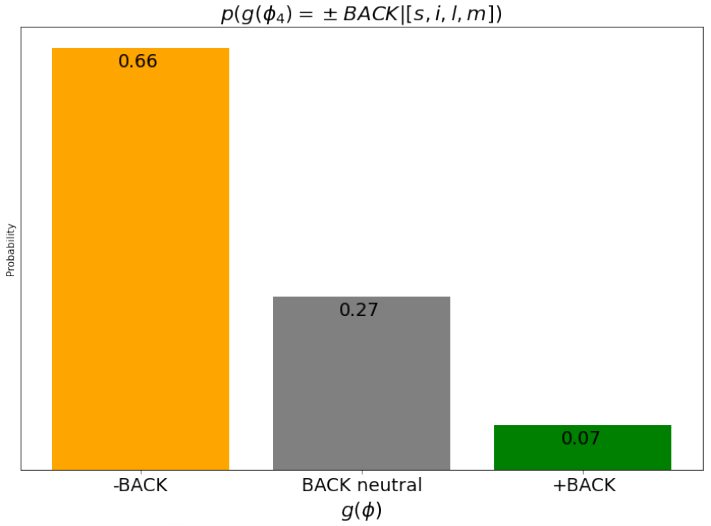


Feature surprisal

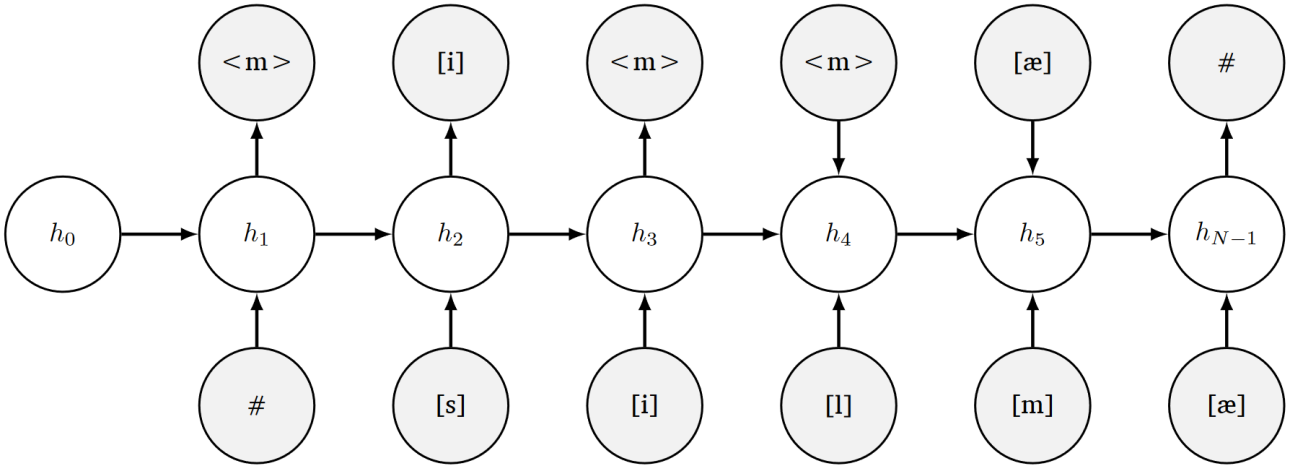
- Surprisal @ -BACK: -log(0.66) = 0.5995
- Surprisal @ +BACK: -log(0.07) = 3.8365
- Surprisal reduction: 3.8365-0.5995 = 3.2370

$$\eta(\mathcal{H}, t) = -\log_2 \sum_{\pi \in \mathcal{H}} p(\pi \mid t, \varphi_{<t})$$

- Surprisal of harmonic group *H* given context



Neural Language Model

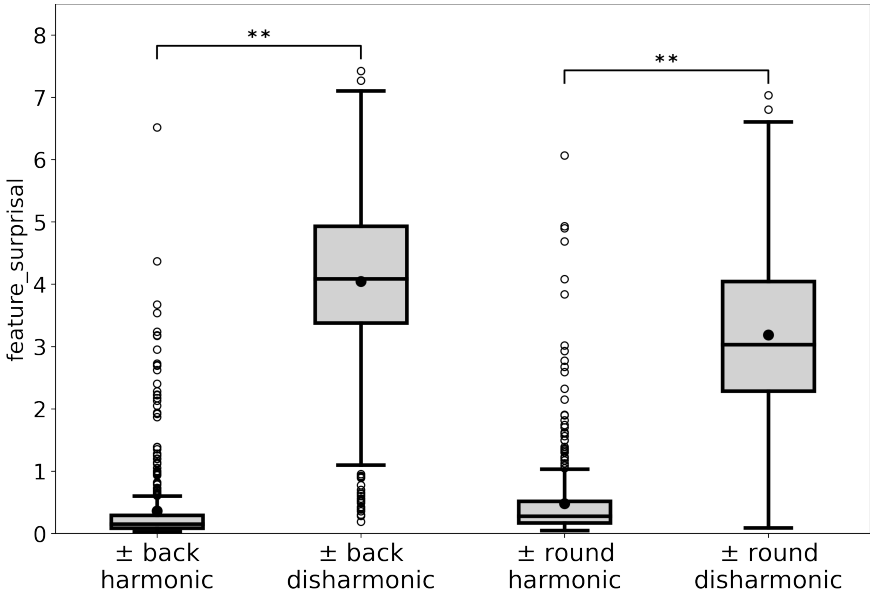


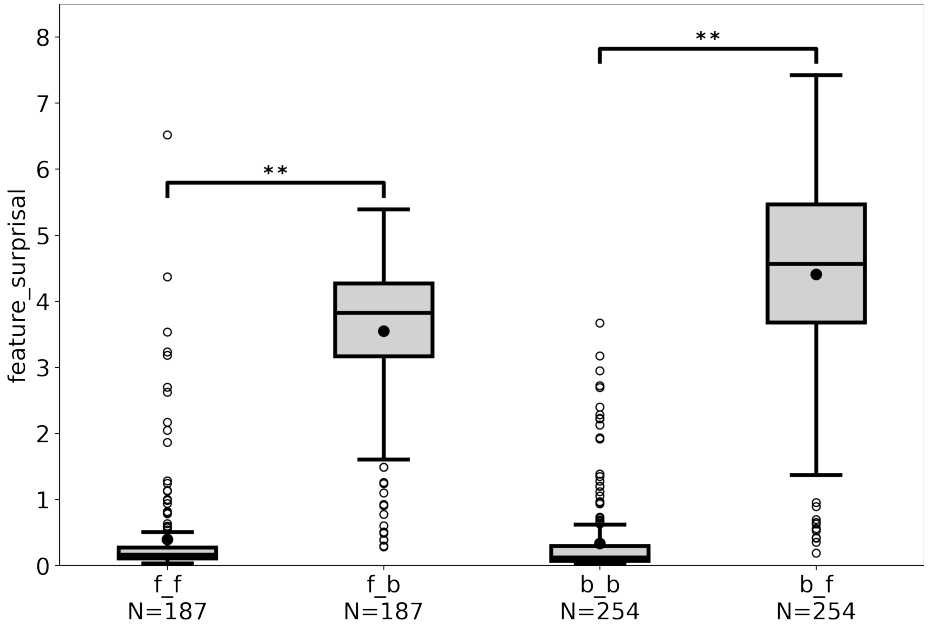
- Probabilities are parametrized via a NLM
  - Based on Feedforward LSTM and hyperparameters in Pimentel et al. (2021)
  - Task: Next phoneme prediction, minimize NLL loss
  - Output restricted to vowel inventory (consonants replaced by mask symbol)
- Separate model trained for each language

Results

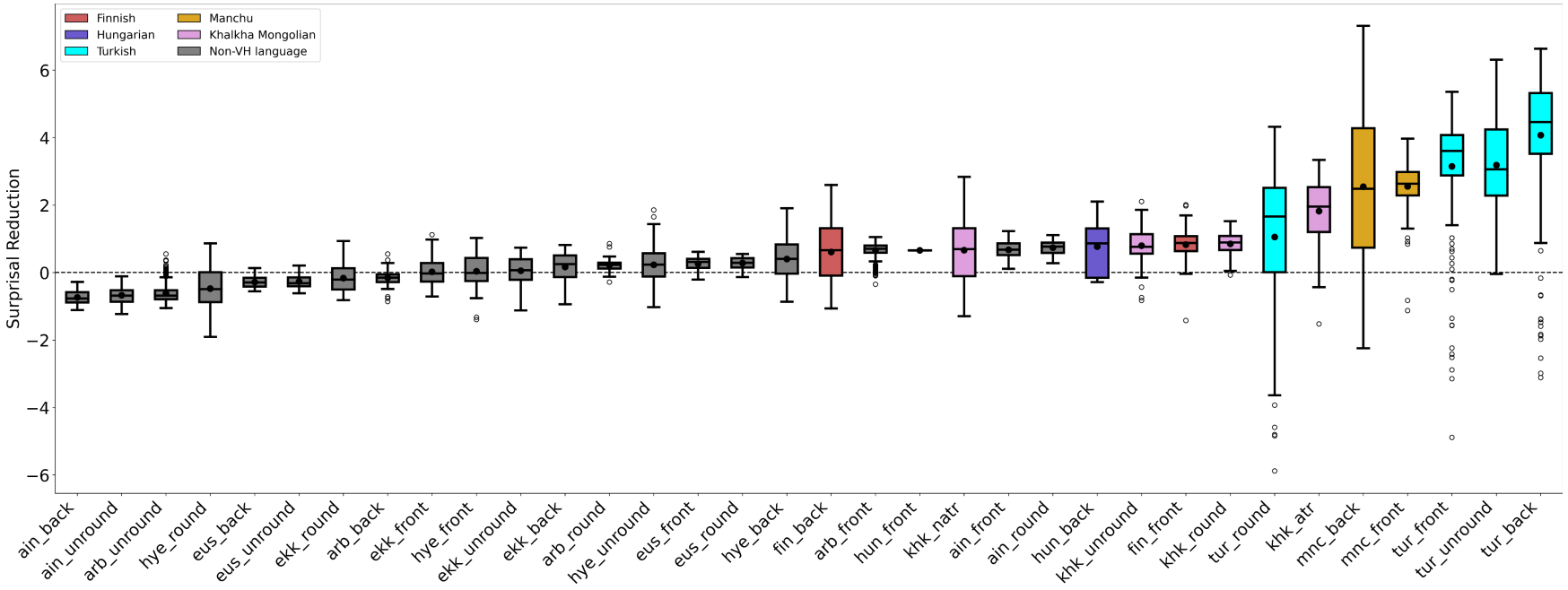
Turkish

- Turkish has strong BACK and ROUND harmony
- +BACK harmony stronger than -BACK
- ROUND harmony not as strong as BACK
- Expected, since some suffixes lack +ROUND forms





Surprisal Reduction



- Difference between surprisal in the harmonic and disharmonic context
- Turkish + Manchu as expected, Finnish & Hungarian closer to non-VH languages
- Khalkha Mongolian indecisive

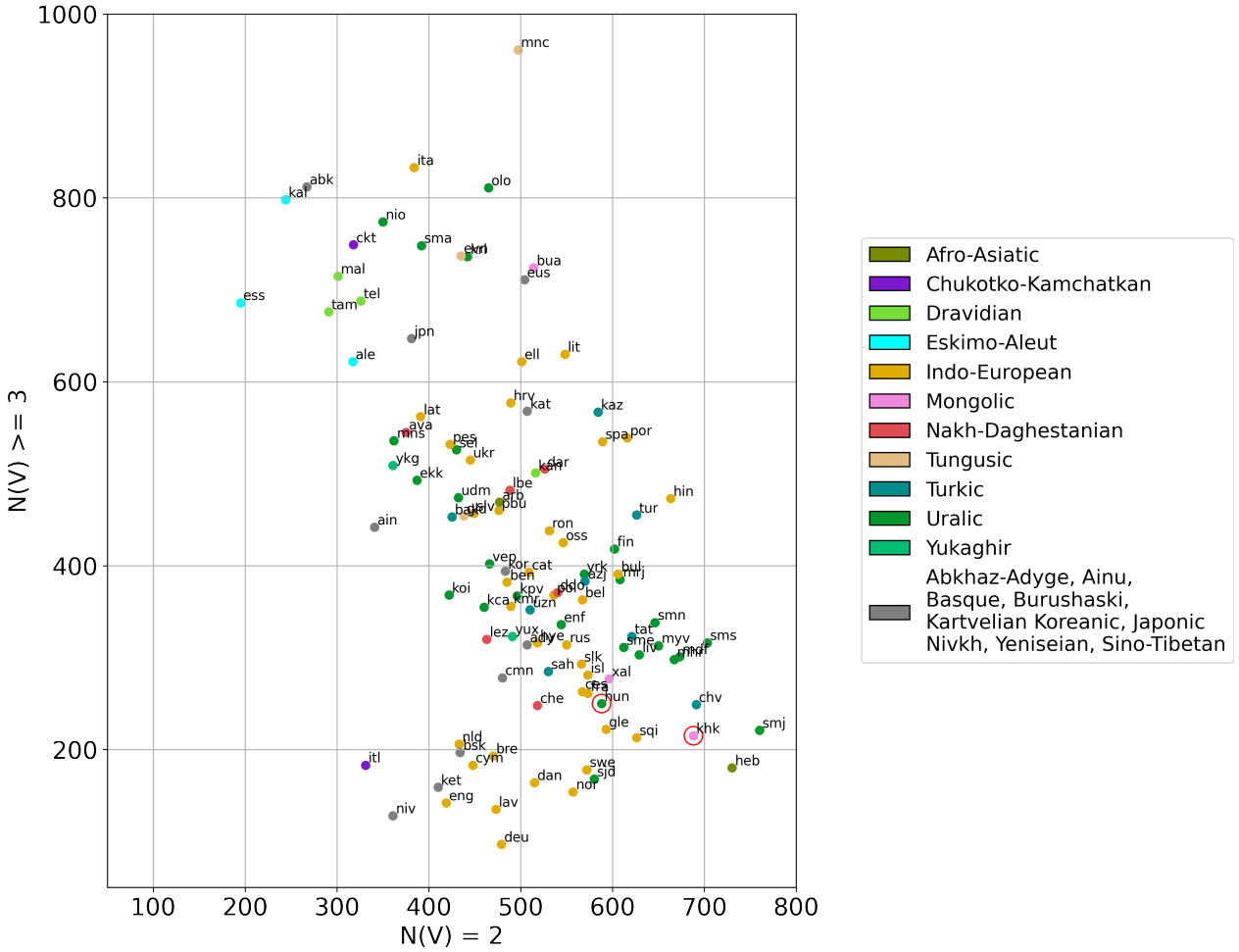
Discussion

Findings

- NLM could learn vowel harmony constraints from word list data
  - Word lists were on the “larger” side
- Found vowel harmony constraints in
  - Turkish, Manchu, Khalkha Mongolian
  - To a lesser degree in Finnish and Hungarian
- Few items with $>=$ 3 vowels for Hungarian
  - Makes it difficult to analyze behavior of neutral vowels
  - Even fewer for Khalkha Mongolian

Limits of Word List Data

- Test set of ~300 word forms
- Majority of the data needed for NLM training
- For complex interactions more data is needed to observe them in the test set
- Neutral vowels, weaker constraints, loanwords, opaque vowels…
- Supports Dockum et al. (2019)



Questions?

References

- Baker, Adam C. “Two Statistical Approaches to Finding Vowel Harmony.” University of Chicago, 2009. http://www.cs.uchicago.edu/research/publications/techreports/TR2009-03.
- Dellert, J. (2021). CLDF dataset derived from Dellert et al.’s ”NorthEuraLex (Version0.9)” from 2020 (v.4.0) [Dataset]. Zenodo. https://doi.org/10.5281/zenodo.5121268.
- Dellert, J., Daneyko, T., Münch, A., Ladygina, A., Buch, A., Clarius, N., Grigorjew, I., Balabel, M., Boga, H. I., Baysarova, Z., Mühlenbernd, R., Wahle, J., and Jäger, G. (2020). NorthEuraLex: a wide-coverage lexical database of Northern Eurasia. Language Resources and Evaluation, 54(1):273–301.
- Dockum, Rikker and Bowern, Claire (2019). Swadesh lists are not long enough: Drawing phonological generalizations from limited data. Language Documentation and Description, 16:35–54.
- Goldsmith, John, and Jason Riggle. “Information theoretic approaches to phonological structure: The case of Finnish vowel harmony.” Natural Language & Linguistic Theory 30, no. 3 (August 2012): 859–96. https://doi.org/10.1007/s11049-012-9169-1.
- Mayer, Thomas, Christian Rohrdantz, Miriam Butt, Frans Plank, and Daniel Keim. “Visualizing vowel harmony.” Linguistic Issues in Language Technology; Vol 4, January 1, 2010.

References

- Pimentel, T., Cotterell, R., and Roark, B. (2021). Disambiguatory signals are stronger in word-initial positions. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 31–41.

Appendix

Results: Finnish

Results: Hungarian

Results: Khalkha Mongolian

Results: Manchu

Results: Turkish