Comparing cross-language phonological profiles

George Moroz

Linguistic Convergence Laboratory (HSE University)

November 9, 2021

presentation is available here: tinyurl.com/yj2tacek

Why did I decide to give this talk?

- During the talk about “Geography and language divergence: the case of Andic languages” at our Lab

Jeff Good: How did you come up with the idea of calculating phonological distances? Is it some established procedure?

Me: No, we thought that this would be the most obvious step…

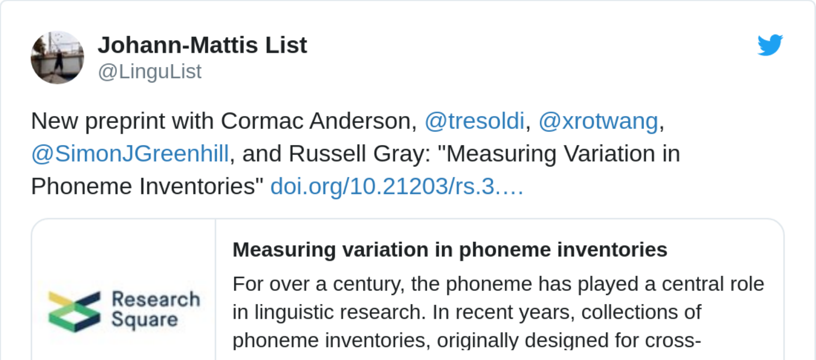
Why did I decide to give this talk?

- During the talk about “Geography and language divergence: the case of Andic languages” at our Lab

Jeff Good: How did you come up with the idea of calculating phonological distances? Is it some established procedure?

Me: No, we thought that this would be the most obvious step…

- The second reason:



Why did I decide to give this talk?

The main reason for this talk is that I performed calculations of phonological distances (or supervised other people making calculations) for Circassian languages [Мороз 2021] and languages of the Andic branch of the East Caucasian family [Moroz 2020; Давиденко 2021; Tsyzova and Zhang 2021], in order to get a simple distance between different idioms, that is less connected to language phylogeny.

Why did I decide to give this talk?

The main reason for this talk is that I performed calculations of phonological distances (or supervised other people making calculations) for Circassian languages [Мороз 2021] and languages of the Andic branch of the East Caucasian family [Moroz 2020; Давиденко 2021; Tsyzova and Zhang 2021], in order to get a simple distance between different idioms, that is less connected to language phylogeny.

But does this measure make any sense? How can we compare phonological profiles of languages?

Why did I decide to give this talk?

The main reason for this talk is that I performed calculations of phonological distances (or supervised other people making calculations) for Circassian languages [Мороз 2021] and languages of the Andic branch of the East Caucasian family [Moroz 2020; Давиденко 2021; Tsyzova and Zhang 2021], in order to get a simple distance between different idioms, that is less connected to language phylogeny.

But does this measure make any sense? How can we compare phonological profiles of languages?

Unlike lexicostatistical distance, phonological distance can be evidence for language contact, since languages can adopt some phonological feature or property from a contact language (see [Andersson et al. 2017]). This can possibly be explained by the **Perceptual Magnet framework** [Blevins 2017]. In most cases phonological change in unrelated languages is more salient to linguists, however it is worth mentioning that there is at least one study about how to catch contact-induced phonological change in

Materials for the analysis

Materials for the calculation of phonological distance can be different:

- segment[1] inventory (and grammar, if you are lucky);
- dictionaries;
- parallel dictionaries;
- corpora;
- parallel corpora.

[1] Let's leave the phonology vs. phonetics debate aside.

Overview

Criticism by [Simpson 1999]

Complexity-based approaches

Distance-based approaches

Criticism by [Simpson 1999]

[Simpson 1999] criticizes UPSID[1]-like researches:

- phonemes mask allophones
  - Standard High German /ç/ stands for [ç], [x] and [χ];
  - "The allophone no longer represents the phoneme, it *replaces* it";
- the phonological relations between segments are lost
  - it is impossible to get information about e.g. vowel harmony by comparing just vowel inventories;
- there is no non-arbitrary way of assigning phonological features (e.g. SPE [Chomsky and Halle 1968]) to segments.

[1] UPSID stands for UCLA Phonological Segment Inventory Database [Maddieson and Abramson 1987], which consists of the phonemic systems of a representative sample of 451 (this number changes from publication to publication) of the world's languages in machine-readable form. UPSID can be accessed via the PHOIBLE database [Moran and McCloy 2019].

Criticism by [Simpson 1999]

[Simpson 1999] criticizes UPSID[1]-like researches:

- phonemes mask allophones
  - Standard High German /ç/ stands for [ç], [x] and [χ];
  - “The allophone no longer represents the phoneme, it *replaces* it”;
- the phonological relations between segments are lost
  - it is impossible to get information about e.g. vowel harmony by comparing just vowel inventories;
- there is no non-arbitrary way of assigning phonological features (e.g. SPE [Chomsky and Halle 1968]) to segments.

My metaphor: omelets and pancakes share the same ingredients, but they are significantly different meals.

[1] UPSID stands for UCLA Phonological Segment Inventory Database [Maddieson and Abramson 1987], which consists of the phonemic systems of a representative sample of 451 (this number changes from publication to publication) of the world’s languages in machine-readable form. UPSID can be accessed via the PHOIBLE database [Moran and McCloy 2019].

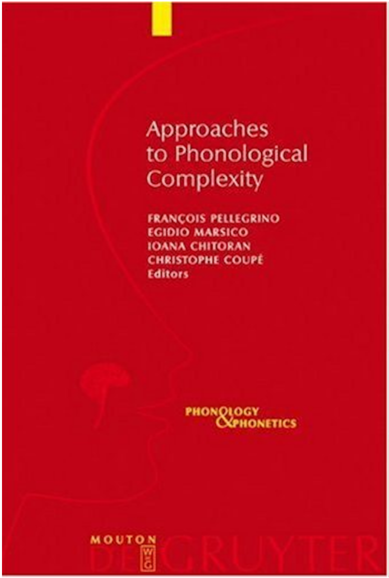
Overview

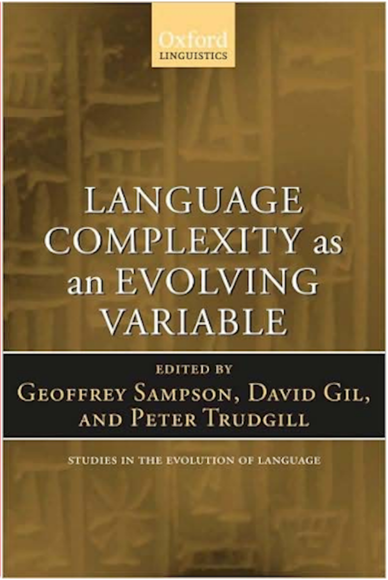
Criticism by [Simpson 1999]

Complexity-based approaches

Distance-based approaches

[Pellegrino et al. 2009] and [Sampson et al. 2009]





[Pellegrino et al. 2009] and [Sampson et al. 2009]

- [Pellegrino et al. 2009]
  - [Ohala 2009]
  - [Maddieson 2009]
  - [Coupé et al. 2009]
- [Sampson et al. 2009]
  - [Nichols 2009]
  - [Deutscher 2009]

[Nichols 2009]

The main goal of this paper was to calculate the overall complexity for a typological sample of languages based on phonology, synthesis, classification (gender, numeral classifiers), syntax, and lexicon. The main goal was to prove:

- that all languages **are not** equal in complexity;
- that different parts of grammar **do not** compensate for complexity in other parts of grammar.

[Nichols 2009]

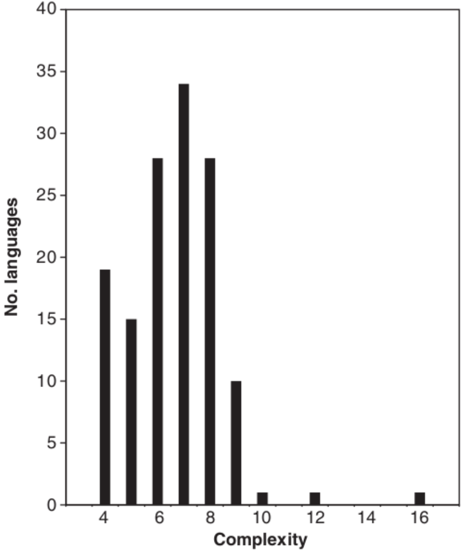
The main goal of this paper was to calculate the overall complexity for a typological sample of languages based on phonology, synthesis, classification (gender, numeral classifiers), syntax, and lexicon. The main goal was to prove:

- that all languages **are not** equal in complexity;
- that different parts of grammar **do not** compensate for complexity in other parts of grammar.

Phonological features in the

- number of contrastive manners of articulation in stops;
- number of vowel quality distinctions;
- tone system (none/simple/complex, after [Maddieson 2013b]);
- syllable structure (after [Maddieson 2013a]).

[Nichols 2009: 116]: results



Phonological complexity (N = 137)

[Ohala 2009]

'Secondary distinctive features' are important for phonologization:

- nasals in French: saint [sɛ̃] < Latin sanctus 'holy';
- average F0 contour of vowels following English stops is falling after voiceless and rising after voiced.

They are not captured by the segmental inventories.

Allophones, like English /t/: [tʰ] vs [t] vs [ɾ] (cf [Simpson 1999]).

[Maddieson 2009]

- Merged measure for consonants, vowels, tones and syllable structure;

Indonesian (Austronesian)	Birom (Niger-Congo; Nigeria)	Kiowa (Kiowa-Tanoan; USA)
p t k	p t k kp	p t̪ k ʔ
		p^h t̪^{h} k^h
b d g	b d g gb	b d̪ g
		p’ t̪’ k’
tʃ dʒ	tʃ dʒ	ts ts’
f s ʃ x h	f s h	s h
z	v z	z
m n ɲ ŋ	m n ŋ	m n̪
l r	l r	d̪ɮ
w j	w j	j

[Maddieson 2009]

- Merged measure for consonants, vowels, tones and syllable structure;

Indonesian = 24	Birom = 27	Kiowa = 32
1 1 1	1 1 1 2	1 1 1 1
		2 2 2
1 1 1	1 1 1 2	1 1 1
		2 2 2
1 2	1 2	1 2
1 1 1 1 1	1 1 1	1 1
2	2 2	2
1 1 1 1	1 1 1	1 1
1 1	1 1	3
1 1	1 1	1

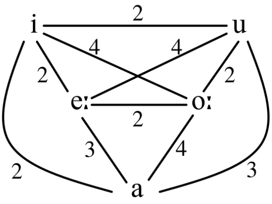
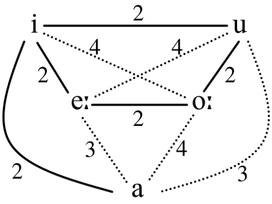
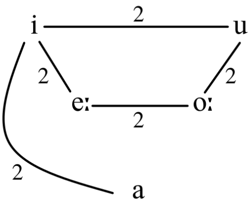
[Maddieson 2009]

- Merged measure for consonants, vowels, tones and syllable structure;
- The number of possible distinct syllables allowed by the language (cf [Shosted 2006]);
- Frequency measures based on lexicon or texts (cf [Давиденко 2021] for Andic):

"To compare these data, it is useful to calculate some kind of index. There are a number of ways this might be done. One possibility is to calculate a summed frequency × complexity score over the top ten segments, in which each segment contributes decreasingly according to its rank, and increasingly according to its complexity". [Maddieson 2009: 97]

[Coupé et al. 2009]

- In this work the authors use phonological features as distances between segments and then use graphs with segments as the nodes and distances as the edges:

STEP 1	STEP 2	STEP 3
We compute the direct phonetic distance for each phonemes pair.	Identification of pairs of phonemes for which an indirect path requires smaller "jumps" than the direct one.	Suppression of costly direct paths.
		

[Coupé et al. 2009]

- In this work the authors use phonological features as distances between segments and then use graphs with segments as the nodes and distances as the edges:
- Then, the authors use the *off-diagonal complexity* proposed by [Claussen 2007],[1] which makes it possible to disassociate from linguistics and phonology and rely purely on graph structure.

---

[1]Their motivation for this choice, was that this measure

- does not explicitly take into account graph size;
- is sensitive to the presence of hierarchical sub-structures in the network;
- is minimal for regular graphs and maximal for free-scale graphs.

Unfortunately, off-diagonal complexity can not be calculated for valued graphs, so authors were ought to drop phonologcial distance values from their graphs.

[Deutscher 2009]

- 'All Languages are Equally Complex' — is a myth (actually, a lot of papers from [Sampson et al. 2009] state the same).
- Complexity is a polysemous notion: some scholars focus on the multipartite nature of language, others on the complicated relations within the system.
- Overall complexity is better to present as a vector of values rather than one value.

Conclusions

Despite the criticism that a language's phonological system is a complex system that can not be reduced to the set of its elements [Simpson 1999; Ohala 2009; Coupé et al. 2009; Deutscher 2009], I think that any phonological complexity measure can be used in order to compare different languages. The sophistication and granularity of this measure will influence the possible effect size gathered by this measure.

Overview

Criticism by [Simpson 1999]

Complexity-based approaches

Distance-based approaches

Distance-based approaches

- [Hoppenbrouwers and Hoppenbrouwers 2001] (after [Heeringa 2004])
- [Nerbonne and Heeringa 2001] (after [Heeringa 2004])
- [Heeringa 2004]
- [Eden 2018]
- [Anderson et al. 2021]

[Anderson et al. 2021]

In this paper the authors apply the **Jaccard similarity** between two phoneme inventories, that is a ratio of similar segments in two languages out of all the possible segments in two languages.

[Anderson et al. 2021]

In this paper the authors apply the **Jaccard similarity** between two phoneme inventories, that is a ratio of similar segments in two languages out of all the possible segments in two languages.

The reason why authors do this, is because their goal is to compare different inventories of the **same** languages across four databases of phonological inventories (UPSID [Maddieson and Abramson 1987], LAPSyD [Maddieson et al. 2013], Core PHOIBLE [Moran and McCloy 2019], JIPA [Baird et al. 2021]). The results are unfavorable: researchers found a high degree of variation across datasets.

[Hoppenbrouwers and Hoppenbrouwers 2001] after [Heeringa 2004]

- Extract frequencies of units (these can be segments, syllables or phonological features) from corpora or dictionaries.
- The distance between two languages is the sum of the differences between the corresponding unit frequencies.

[Nerbonne and Heeringa 2001] after [Heeringa 2004]

The authors applied the same stratagy as [Hoppenbrouwers and Hoppenbrouwers 2001], but used words as corpora. So the idiom distance is calculated as an average word distance.

[Heeringa 2004]

Since the methods of [Hoppenbrouwers and Hoppenbrouwers 2001] and [Nerbonne and Heeringa 2001] do not account for unit order, Heeringa decided to use the Levenstein distance [Левенштейн 1965]. The Levenstein distance is the minimum number of unit edits (insertions, deletions or substitutions) that should be applied to the unit string in order to get another:

- the distance between *pancake* and *omelet* is 7
- the distance between *pancake* and *cake* is 3
- the distance between *sing* and *sign* is 1

Shortcoming:

- diphthong vs. vowel + consonant combination (/au/ or /aw/?);
- suprasegmental features;
- sequence length: the longer the sequences, the greater the chance of differences between them.

[Heeringa 2004]

To address the sequence length problem [Heeringa 2004] uses normalization by the length of the alignment:

$$\begin{array}{ccc} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ \hline 1 & 1 & 1 \end{array} \qquad \begin{array}{ccc} a_1 & a_2 & \emptyset \\ b_1 & b_2 & b_3 \\ \hline 1 & 1 & 1 \end{array} \qquad \begin{array}{ccc} a_1 & a_2 & \emptyset \\ \emptyset & \emptyset & b_3 \\ \hline 1 & 1 & 1 \end{array} \qquad \begin{array}{ccc} a_1 & a_2 & \emptyset \\ \emptyset & b_2 & b_3 \\ \hline 1 & 1 & 1 \end{array}$$

$$(1) \qquad\qquad (2) \qquad\qquad (3) \qquad\qquad (4)$$

All four examples are normalized by the value 3.

[Heeringa 2004]: interlanguage stimuli mismatch

- It is possible that for one of the pair of idioms one lacks stimuli, in which case the effect of this stimulus is discounted.
- In case of multiple transcriptions they are matched according the minimum distance:
    - L1: [hys]; L2: [hys] and [hus]
    - L1: [hys] and [hus]; L2: [hys] and [hus]

[Eden 2018]

- Parametric typology:
  - annotate languages according to 27 syllable structure parameters, and 29 inventory parameters;
  - apply Hamming distance,[1]
- Cross-entropy of the transcribed example texts

[1]This is just an ordered version of the Jaccard distance: the ratio of similar units in two languages out of all possible units in two languages.

[Eden 2018]: Shannon information entropy

Entropy is a measure of randomness or uncertainty proposed by [Shannon 1948].

$$H(X) = -\sum_{i=1}^{n} P(x_i) \times \log_2 P(x_i)$$

Possible entropy values are $H(X) \in [0, +\infty]$:

dataset	entropy
A-A-A-A-A	0.00
A-A-A-A-B	0.72
A-A-A-B-B	0.97
A-A-B-B-B	0.97
A-A-B-B-C	1.52
A-B-C-A-B	1.52

[Eden 2018]: Cross-entropy

A cross-entropy measure is used in order to compare the two distributions X and Y:

$$H(X, Y) = -\sum_{i=1}^{n} P(x_i) \times \log_2 P(y_i)$$

first dataset	cross-entropy	second dataset
A-A-A-A-B	0.72	A-A-A-A-B
A-A-A-A-B	0.85	A-A-A-B-B
A-A-A-B-B	1.09	A-A-B-B-B
A-A-A-A-B	1.20	A-A-B-B-B

[Eden 2018]: Kullback-Leibler divergence

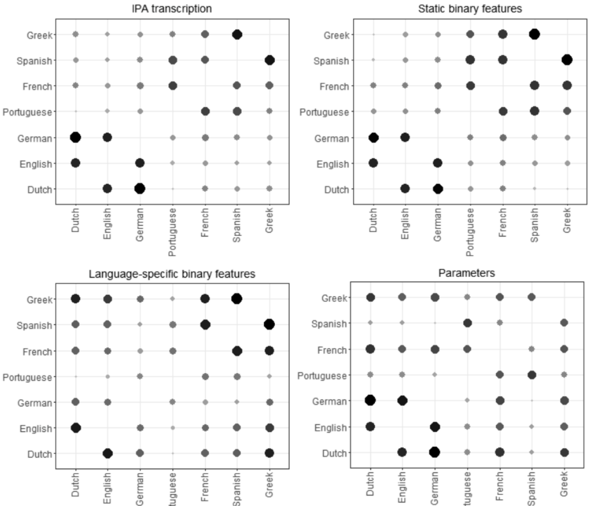
As we have seen, if the distributions X and Y are equal, then the cross-entropy is equal to the entropy of that distribution. So there is a way to normalize it, called the Kullback-Leibler divergence measure:

$$D_{KL}(P \parallel Q) = H(P, Q) - H(P)$$

first dataset	Kullback-Leibler	second dataset
A-A-A-A-B	0.00	A-A-A-A-B
A-A-A-B-B	0.12	A-A-B-B-B
A-A-A-A-B	0.13	A-A-A-B-B
A-A-A-A-B	0.48	A-A-B-B-B

[Eden 2018]: application (greater circles stand for similar languages)

- 7 languages from the corpus of European languages (Portuguese, French, Spanish, German, English, Dutch, Greek)



[Eden 2018]: application

- 7 languages from the corpus of European languages (Portuguese, French, Spanish, German, English, Dutch, Greek)
- Pearson correlation between some metrics:

	Parameter	Entropy: IPA	Entropy: stat
Entropy: IPA	0.67		
Entropy: static binary features	0.55	0.94	
Entropy: ls binary features	0.33	0.46	0.38

Conclusions

In this talk I presented multiple methods of calculating the phonological distances between languages.

Conclusions

In this talk I presented multiple methods of calculating the phonological distances between languages.

It is worth noting that all arguments against using segment inventories as a measure [Simpson 1999; Ohala 2009; Coupé et al. 2009; Deutscher 2009], but dictionaries and corpora are not freely available for any language.

Conclusions

In this talk I presented multiple methods of calculating the phonological distances between languages.

It is worth noting that all arguments against using segment inventories as a measure [Simpson 1999; Ohala 2009; Coupé et al. 2009; Deutscher 2009], but dictionaries and corpora are not freely available for any language.

Both complexity-based and distance-based approaches are valid for language comparison:

- the complexity-based approach probably is better by design, but it depends on a feature set that should be chosen by linguists;
- the distance-based approach is better in capturing all phonotactic details present in the data (as [Maddieson 2009] proposed), but it depends on more or less comparable data from different languages.

Conclusions

In this talk I presented multiple methods of calculating the phonological distances between languages.

It is worth noting that all arguments against using segment inventories as a measure [Simpson 1999; Ohala 2009; Coupé et al. 2009; Deutscher 2009], but dictionaries and corpora are not freely available for any language.

Both complexity-based and distance-based approaches are valid for language comparison:

- the complexity-based approach probably is better by design, but it depends on a feature set that should be chosen by linguists;
- the distance-based approach is better in capturing all phonotactic details present in the data (as [Maddieson 2009] proposed), but it depends on more or less comparable data from different languages.

Acoustic distances in [Heeringa 2004] and [Eden 2018]!

Thank you for your attention!

References

Anderson, C., Tresoldi, T., Greenhill, S. J., Forkel, R., Gray, R. D., and List, J.-M. (2021). Measuring variation in phoneme inventories (preprint v1). *Research Square*.

Andersson, S., Sayeed, O., and Vaux, B. (2017). The phonology of language contact. *Oxford Handbooks Online*.

Baird, L., Evans, N., and Greenhill, S. J. (2021). Blowing in the wind: Using 'north wind and the sun'texts to sample phoneme inventories. *Journal of the International Phonetic Association*, pages 1–42.

Blevins, J. (2017). Areal sound patterns: From perceptual magnets to stone soup. *The Cambridge handbook of areal linguistics*, 5587.

Bowern, C. (2013). Relatedness as a factor in language contact. *Journal of Language Contact*, 6(2):411–432.

References

Chomsky, N. and Halle, M. (1968). *The sound pattern of English*. Harper and Row.

Claussen, J. C. (2007). Off-diagonal complexity: A computationally quick complexity measure for graphs and networks. *Physica A: Statistical Mechanics and its Applications*, 375(1):365–373.

Coupé, C., Marsico, E., and Pellegrino, F. (2009). Structural complexity of phonological systems. In *Approaches to phonological complexity*, pages 141–170. De Gruyter Mouton.

Deutscher, G. (2009). ”overall complexity”: a wild goose chase? In *Language complexity as an evolving variable*, pages 243–252. Oxford University Press.

Eden, S. E. (2018). *Measuring phonological distance between languages*. PhD thesis, University College London.

References

Heeringa, W. J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. PhD thesis, University Library Groningen.

Hoppenbrouwers, C. A. J. and Hoppenbrouwers, G. A. (2001). *De indeling van de Nederlandse streektalen: dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Uitgeverij Van Gorcum.

Maddieson, I. (2009). Calculating phonological complexity. In *Approaches to phonological complexity*, pages 83–110. De Gruyter Mouton.

Maddieson, I. (2013a). Syllable structure. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

References

Maddieson, I. (2013b). Tone. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Maddieson, I. and Abramson, A. S. (1987). Patterns of Sounds. *The Journal of the Acoustical Society of America*, 82(S1):720–721.

Maddieson, I., Flavier, S., Marsico, E., Coupé, C., and Pellegrino, F. (2013). Lapsyd: Lyon-Albuquerque phonological systems database. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*.

Moran, S. and McCloy, D., editors (2019). *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

Moroz, G. (2020). Comparing phonological systems and syllable structure of Botlikh and Zilo Andi: a data–driven analysis.

References

Nerbonne, J. and Heeringa, W. (2001). Computational comparison and classification of dialects. *Computational Comparison and Classification of Dialects*, 9:69–83.

Nichols, J. (2009). Linguistic complexity: a comprehensive definition and survey. In *Language complexity as an evolving variable*, pages 110–125. Oxford University Press.

Ohala, J. J. (2009). Languages' sound inventories: the devil in the details. In *Approaches to phonological complexity*, pages 47–58. De Gruyter Mouton.

Pellegrino, F., Marsico, E., Chitoran, I., and Coupé, C. (2009). *Approaches to phonological complexity*, volume 16. Walter de Gruyter.

Sampson, G., Gil, D., and Trudgill, P. (2009). *Language complexity as an evolving variable*. Oxford University Press.

References

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Shosted, R. K. (2006). Correlating complexity: A typological approach. *Linguistic Typology*, 10(1):1–40.

Simpson, A. P. (1999). Fundamental problems in comparative phonetics and phonology: does UPSID help to solve them. In *Proceedings of the 14th international congress of phonetic sciences*, volume 1, pages 349–352.

Tsyzova, A. and Zhang, W. (2021). Harmony effects in andic languages. Term paper.

Давиденко, ?. ?. (2021). Сравнение фонологических систем, полученных на основе словарей и корпусов: данные андийских языков. Выпускная курсовая работа.

References

Левенштейн, ?. ?. (1965). Двоичные коды с исправлением выпадений, вставок и замещений символов. In *Доклады Академии наук*, volume 163, pages 845–848. Российская академия наук.

Мороз, ?. ?. (2021). *Некоторые вопросы сегментной и супрасегментной фонологии и фонетики адыгских языков*. PhD thesis, Higher School of Economics, Moscow.