SPATIAL SOUND SYMBOLISM: A CROSS-LINGUISTIC STUDY

Alexandra Wikström, Lari Vainio & Martti Vainio

University of Helsinki, Finland
firstname.lastname@helsinki.fi

ABSTRACT

Recent behavioral evidence shows that the spatial concepts of *front* and *back* are sound symbolically tied to vowels that are pronounced either at the front or the back of the mouth, respectively. With such effects found from behavioral experiments, we examined whether such phenomenon could be found cross-linguistically from words pertaining to the concepts of front and back. The frequency counts of vowels of such words from 266 languages were extracted and analyzed. The analysis showed statistically significant differences in frequency counts of vowels between the *back* and *front* concepts. These findings suggest that the phenomenon of vowel articulation connecting to word meaning does have some cross-linguistic basis and hints at a common motor-sensory foundation for spatial sound symbolic effects.

Keywords: sound symbolism, spatial sound symbolism, vowel space, cross-linguistic research

1. INTRODUCTION

Sound symbolism assumes a non-arbitrary connection between the phonetic form and conceptual meaning of a word when it comes to certain concepts. Empirical studies have shown the effects of sound symbolism in multiple aspects, such as round and pointy shapes and large and small objects connecting to differing speech sounds ([1], [2]). Recent experimental data has also expanded this to the modality of the spatial concepts of *front* and *back* in addition to the more dynamic concepts of *forward* and *backward*. For example, [3] discovered the connection between front vowels and front concepts and back vowels and back concepts. Earlier evidence comes from Tanz's work [4] who studied the occurrence of front and back vowels in word forms indicating both spatial and temporal distances in different languages. Most commonly sound symbolism has been behaviorally researched with forced-choice tests ([1], [5]) as well as with cross-linguistic methods focusing on the universality of the phenomenon, usually by studying the lexicons of the world ([4], [6]).

With the empirical and earlier cross-linguistic research suggesting sound symbolic effects in vowels in the modality of spatial positions, we set out to discover if the vocabularies of the world would show preference of front or back vowels when it comes to the words presenting the spatial concepts of *front* and *back*, respectively.

2. METHODS

The phonetic representations of words pertaining to the concepts of *front* and *back* were extracted from the Database of Cross-Linguistic Colexifications (CLICS) [7] which includes word lists from varying datasets. All the words contained in the database have been connected to particular, catalogued, meanings. More notes on the utilizing of this database can be found in section 5.1.

Data was gathered from 266 languages, with the distribution of data points within language families shown in Table 1 and the language-typological distribution of the data in Figure 1. The dataset ended up consisting of words with the meanings of *front*, *in front of* and *forward* in the *front* category and *back*, *behind* and *backward* in the *back* category. Altogether 713 words were included, with 355 words connected to the *front* concepts and 358 words connected to the *back* concepts. Multiple words from a single language connected to a concept were included in the analysis taken that the words for a particular concept were different from each other (e.g., *behind* vs. *backward* pertaining to the concept of *back* in English), and would thus offer more varied phonetic representations, while still representing the two separate spatial concepts.

3. ANALYSIS

Out of the gathered words, the frequency counts of all vowels were extracted, and each vowel was given an (x,y) coordinate based on the location of said vowel in the IPA vowel diagram (x values 0-8, y values 0-6). This helped to create a vowel space heat map to visualize the distribution of the vowels in the contrasting concepts (Figure 2). Out of all the words, 1821 vowels were extracted, 890 from the

Figure 1: A map of the languages included in the study, locations based on Glottolog language coordinates [8].

back concepts and 931 from the *front* concepts.

Statistical analysis of the vowel frequency counts was carried out using RStudio [9]. Considering the variation in how many data points were observed due to the nature of certain vowels being more common than others, for analysis purposes the frequency counts of vowels containing zero samples in either of the front or back categories were excluded from the analysis. A Fisher's Exact Test for count data was carried out to determine the statistical significance of the differences in vowel frequency counts of the samplings of *front* and *back* concepts altogether. To analyze the groupings in more detail, a Pearson's Goodness-of-Fit Test was computed on the accumulated frequency counts of vowel groups of front vowels and back vowels. Lastly, to better study the effects of frontness/backness in individual vowels that show seemingly robust effects, individual Pearson's Goodness-of-Fit Tests were computed on the frequency counts of individual vowels that had frequency counts higher than 5 in both the *front* and the *back* condition.

4. RESULTS

The frequency count distribution among vowels is shown in Table 2. The representation of different vowels is quite varied in the data, as is their appearance in the world's languages themselves: the vowels [ɜ], [ɞ] and [ɶ] do not appear in the data at all and other rarer vowels such as [y], [ɘ] or [œ] appear only a handful of times, whereas the more common vowels, e.g. [i], [a] and [u] are heavily present [10]. It is useful to note that these more common vowels also represent clear articulatory and acoustic contrasts and are at the edges of the articulatory vowel space, and are thus an opportune place for such sound-spatial effects to emerge that are of interest to this research.

Language family	# of data points
Indo-European	174
Tungusic	16
Turkic	32
Basque	6
Uralic	117
Sino-Tibetan	95
Yeniseian	6
Ainu	4
Mongolic	15
Dravidian	16
Afro-Asiatic	17
Chukotko-Kamchatkan	8
Nakh-Daghestanian	25
Eskimo-Aleut	9
Yukaghir	8
Burushaski	2
Abkhaz-Adyge	8
Koreanic	4
Kartvelian	3
Japonic	4
Nivkh	1
Nuclear Trans New Guinea	18
Austronesian	8
Kehu	1
Anim	6
Lakes Plain	13
Geelvink Bay	1
Kaure-Narau	1
Tor-Orya	2
Atlantic-Congo	10
Cariban	5
Mosetén-Chimané	2
Zuni	1
Tsimshian	4
Nubian	2
Hmong-Mien	45
Pama-Nyungan	2
Saharan	2
Karok	2
Otomanguean	1
Mayan	2
Uru-Chipaya	1
Zamucoan	1
Nuclear-Macro-Je	1
Chonan	2
Araucanian	2
Kunza	3
Tupian	5

Table 1: Distribution of data points (words) per language family.

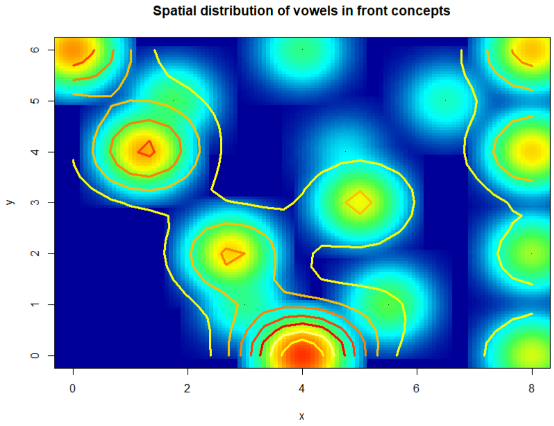
From the heat maps in Figure 2 some differences in data distribution can be observed, also visible in the frequency count data of 2. Most prominently the vowel coordinates of i/y, e/ø, ɯ/u and ɤ/o show more densely in the front concepts than in the back concepts, whereas the vowel coordinate of ɑ/ɒ is more prominent in the back concepts. Most of the data points in both back and front concepts are located in the a/ɶ coordinate.

From the statistical tests we see some significant differences between the frequency counts. With the Fisher's Exact Test a p-value of 0.009 was observed (Monte Carlo simulation, 2000 rounds). This indicates that the altogether uneven distribution of frequency counts between the two categories is

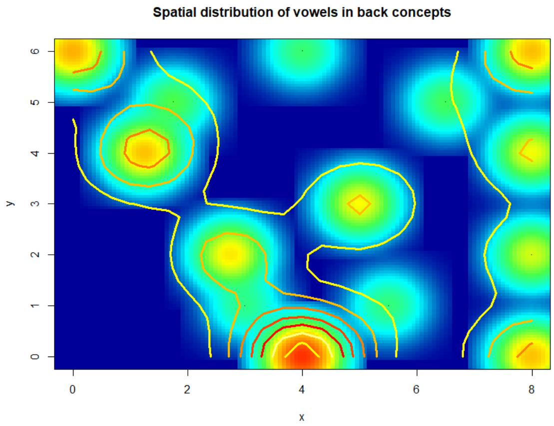
Vowel	Back	Front	Total	p-value
i	96	120	216	0.103
y	4	4	8	NA
ɨ	15	12	27	0.564
ʉ	1	2	3	NA
ɯ	23	10	33	0.024 *
u	68	82	150	0.253
ɪ	19	19	38	1
ʏ	3	0	3	NA
ʊ	19	8	27	0.034 *
e	85	108	193	0.098 .
ø	4	3	7	NA
ɘ	0	1	1	NA
ɵ	0	2	2	NA
ɤ	0	8	8	NA
o	56	72	128	0.157
ə	58	58	116	1
ɛ	66	78	144	0.317
œ	1	5	6	NA
ɜ	0	0	0	NA
ɞ	0	0	0	NA
ʌ	5	4	9	NA
ɔ	42	33	75	0.299
æ	13	13	26	1
ɐ	15	24	39	0.150
a	208	216	424	0.698
ɶ	0	0	0	NA
ɑ	83	45	128	0.0008 ***
ɒ	6	4	10	NA

Table 2: Frequency counts of vowels in the front and back concepts and the p-values of Pearson's Goodness-of-Fit Tests on the significance of frequency count differences between the concepts. Significant values are marked with *, vowels with too little data are marked NA.

statistically significant. With a Pearson's Goodness-of-Fit Test on the vowel groupings of Front vowels (N=499 back concepts, N=599 front concepts) and Back vowels (N=302 back concepts, N=266 front concepts), we obtain a p-value of 0.04 for the Front vowels group and a p-value of 0.131 for the Back vowels group. This is to say that the front vowels are significantly more frequent in the *front* concepts, and that the back vowels are more frequent in the *back* concepts, but not significantly. A Goodness-of-Fit Test was also computed for the groupings of rounded and unrounded vowels to identify any effect that the feature of roundness might have on the frequency counts. No significant differences in frequency between the groups was observed (Unrounded vowels N=628 back concepts, N=658 front concepts, p=0.403; Round vowels N=204 back concepts, N=215 front concepts, p=0.591).



(a) Vowels in front concept words



(b) Vowels in back concept words

Figure 2: Distribution of vowels in a) front concepts and b) back concepts presented in the vowel space. The rounded and unrounded versions of a vowel are represented by the same coordinate.

Looking at Table 2 with the p-values acquired from the Pearson's Goodness-of-Fit Tests on individual vowels we see that the only significant differences in frequency counts of individual vowels between the two concepts occur with the back vowels [ɑ], [ʊ] and [ɯ] which show more prominently in the *back* concepts, although the front vowel [e] comes close with a value of p=0.098 and showing prominence in the *front* concepts.

5. DISCUSSION

The results revealed both expected and unexpected outcomes. The *back* concept was mostly connected to the back vowels [ɑ], [ʊ] and [ɯ]. The *front* concept seemed to have vowels connected to it more prominently ([i, e, o u]), both front and back, but of which the back vowels were rounded. One explanation for the emergence of [o] and [u] in the *front* words could be that lip protrusion can be seen as a way of pointing forward to an object [11]. In addition, for example the unrounded counterpart of [u], [ɯ], was significantly more frequent in the *back* concepts than in the *front* concepts. When comparing the groups of front and back vowels only the connection of front vowels to *front* concepts was significant, although there were more back vowels in the *back* concepts as well. This is in accordance with the fact that the back vowels [u] and [o] appeared more strongly, although not significantly so, in the *front* concepts and thus would not contribute to the *back* concepts as much.

Conducting cross-linguistic experiments in a way that quantifies the vowel space in the way that this study did still needs to be improved upon. Although this study concentrated on the dimensions of *front* and *back*, future analysis will also be done on other spatial concepts such as *high* and *low*, and *up* and *down*, as well as on the distribution of consonants between the two categories.

5.1. Notes on utilizing the CLICS database

The CLICS database used in this study, although fairly large, has data of varying phonetic transcription quality. For example, the NorthEuraLex database [12] and the Lalo Regional Varieties database [13] have data with concise and exact phonetic transcriptions, while the data of some other datasets such as the Intercontinental Dictionary series [14] has both phonetic and orthographic forms and had to be separately validated before being added to the data pool. The quality fluctuations of the original datasets also affected the distribution of data points among language families: for some of the languages simply more vocabulary was available, and many of the datasets focused on a particular region, adding to the varying quality of data available for a certain language family.

The CLICS database also lacks linkages of recorded languages to Glottolog [8] or to any other official cataloging of languages, which meant that for mapping purposes, the languages presented in CLICS had to sometimes be linked manually via given name and coordinates to an official name of the language visible in Glottolog.

6. SUMMARY

The emergence of front and back vowels in the word forms of *front* and *back* concepts was examined across the world's languages. The vowel frequency counts gathered from the data showed differences in frequency distributions of the vowels between the two concepts. Most prominently, the vowels [ɑ], [ʊ] and [ɯ] showed significantly more in the *back* concepts. Also, the front vowels [i] and [e] were more frequent in the *front* concepts, although not significantly so. In addition the rounded vowels [o] and [u] showed prominence in the *front* concept data, although not significant either. The group comparison of front and back vowels showed that there are more back vowels in *back* concepts and significantly more front vowels in *front* concepts. These results would suggest that a sound symbolic effect of frontness and backness can be observed in the word inventories of the world, despite all the change that is imposed on languages over time.

Although the data indicate roundness having an effect on how frontal a vowel is perceived as, more research could be done on how roundness affects the emergence of spatial sound symbolic effects. Lip pointing as a conversational gesture is a widespread phenomenon among the world's populations [15] and therefore it would not be surprising for it to be embedded in how frontal a vowel is perceived as depending on its roundness, which could in turn affect the prevalence of rounded vowels in the *front* concepts researched here.

7. ACKNOWLEDGEMENTS

This research was conducted with funding from the Academy of Finland under grant agreement number 342627.

8. REFERENCES

[1] E. Sapir, "A study in phonetic symbolism," *J. Exp. Psychol.*, vol. 12, pp. 225–239, 1929.

[2] V. S. Ramachandran and E. M. Hubbard, "Synaesthesia–a window into perception, thought and language," *J. Conscious. Stud.*, vol. 8, pp. 3–34, 2001.

[3] L. Vainio, M. Kilpeläinen, A. Wikström, and M. Vainio, "Sound-space symbolism: Associating articulatory front and back positions of the tongue with the spatial concepts of forward/front and backward/back," *Journal of Memory and Language*, vol. 130, p. 104414, 2023.

[4] C. Tanz, "Sound symbolism in words relating to proximity and distance," *Lang. Speech*, vol. 14, pp. 266–276, 1971.

[5] M. Imai, S. Kita, M. Nagumo, and H. Okada, "Sound symbolism facilitates early verb learning," *Cognition*, vol. 109, pp. 54–65, 2008.

[6] D. E. Blasi, S. Wichmann, H. Hammarström, P. F. Stadler, and M. H. Christiansen, "Sound-meaning association biases evidenced across thousands of

languages," *PNAS*, vol. 113, no. 39, pp. 10 818–10 823, 2016.
[7] C. Rzymski, T. Tresoldi *et al.*, "The database of cross-linguistic colexifications, reproducible analysis of cross- linguistic polysemies," 2019. [Online]. Available: clics.clld.org
[8] H. Hammarström, R. Forkel, M. Haspelmath, and S. Bank. (2022) Glottolog 4.6.
[9] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/
[10] P. Ladenfoged and I. Maddieson, "Vowels of the world's languages," *J. Phon.*, vol. 18, pp. 93–122, 1990.
[11] A. Wallace, "Review of Tylor's anthropology," *Nature*, vol. 24, pp. 242–245, 1881.
[12] J. Dellert and G. Jäger. Northeuralex (version 0.9). Tübingen: Eberhard-Karls University.
[13] C. Yang. (2011) Lalo regional varieties: Phylogeny, dialectometry and sociolinguistics.
[14] M. R. Key and B. Comrie, Eds., *The Intercontinental Dictionary Series*. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2015.
[15] N. Enfield, "Lip-pointing," *Gesture*, vol. 1, no. 2, pp. 185–212, 2001.