Ian Joo*

Phonosemantic biases found in Leipzig-Jakarta lists of 66 languages

https://doi.org/10.1515/lingty-2019-0030
Received 08 March 2018; revised 07 March 2019

Abstract: Based on the vocabulary of 66 genealogically distinct languages, this study reveals the biased association between phonological features and the 100 lexical meanings of the Leipzig-Jakarta List. Morphemes whose meanings are related to round shapes ('egg', 'navel', 'neck', and 'knee') tend to contain phonemes that bear the [+round] feature. Also observable is the positive association between buccal actions and the phonological features they resemble ('to blow' with [+labial] and 'to suck' with [+delayed release]). Grammatical morphemes related to proximity ('this', 'in', 1SG and 2SG pronoun) are positively associated with [+nasal]. The phonosemantic patterns found in the most basic vocabulary of spoken languages further confirm that the sound-meaning association in natural languages is not completely arbitrary but may be motivated by human cognitive biases.

1 Introduction

A growing body of studies has shown that certain meanings tend to be represented by lexical items bearing certain sounds. So far, studies have confirmed that 1st and 2nd person pronouns tend to contain nasal sounds (Gordon 1995; Nichols & Peterson 1996); proximal pronouns prefer vowels with higher F2, whereas distal pronouns prefer those with lower F2 (Tanz 1971; Woodworth 1991; Johansson & Zlatev 2013); and words for 'lips' and 'nose' tend to have bilabial stops and nasal sounds, respectively (Urban 2011). Blasi et al. (2016) examined how the basic vocabulary of thousands of languages shows phonosemantic biases. The study, based on the 6,447 Swadesh Lists (Swadesh 1955) of 4,298 languages, found that 30 out of the 100 Swadesh List terms show preference or dispreference for certain sounds in their phonological representations. For example, lexemes for 'tongue' tend to contain /e/, /ɛ/, or /l/ and not contain /u/ or /k/.

As much as Blasi et al.'s study provides us valuable insight on lexical phonosemantic biases, a similar typological study based on a different wordlist

***Corresponding author: Ian Joo [tɕʰù ì.án],** Eurasia3angle, Max Planck Institute for the Science of Human History, Jena, Germany, E-mail: joo@shh.mpg.de

could also be helpful. Swadesh deliberately excluded from his list three lexical terms that he deemed to have "sound imitative tendencies" (Swadesh 1955: 126): 'to blow (air)', 'to cry/weep', and 'to laugh'. For example, he judged that words meaning 'to blow' tend to have labial consonants and/or sibilants, which are iconically associated with the buccal action of blowing air. This was because this iconic tendency could lead to the false impression that two genealogically unrelated words are cognates. Investigating a different wordlist containing different lexical terms, including those excluded by Swadesh, would further broaden our view on lexical phonosemantics.

Moreover, Blasi et al.'s **segmental analysis** calls for the need of a **featural analysis** of a similar database. For example, Blasi et al. have found that /tʃ/ appears frequently in the items for 'small'. Since this segment is composed of diverse phonological features such as [–voice, +coronal, +distributed, +delayed release], we are left relatively uncertain regarding which of these features motivate the association between /tʃ/ and smallness. Thus, examining what features (rather than segments) are associated with each meaning may provide us a clearer view on the cognitive motivation behind those associations.

2 Research questions

This study attempts to answer two questions inspired by the study of Blasi et al.: One, what are the phonosemantic patterns observable in lexical items not contained in the Swadesh List? Two, what are the phonological features statistically associated with each meaning?

3 Methodology

3.1 Wordlist

Instead of the Swadesh List, the present study uses the Leipzig-Jakarta (LJ) List (Tadmor 2009). This list consists of 100 basic terms, 62 of them overlapping with the Swadesh List items, although some overlapping items differ in detail ('hand' in Swadesh List v. 'arm/hand' in the LJ List). The list is shown in Table 1, in alphabetical order.

The 100 terms were empirically selected by Tadmor (2009) out of the World Loanword Database (Haspelmath & Tadmor 2009), a lexical database consisting of 41 languages' vocabularies (the contributors' names are available at https://wold.clld.org/contributor). The four criteria of the selection were:

Table 1: The LJ list.

1SG pronoun	2SG pronoun	3SG pronoun	ant	arm/hand
ash	back	big	bird	to bite
bitter	black	blood	to blow	bone
breast	to burn (intr.)	to carry	child (kin term)	to come
to crush/grind	to cry/weep	to do/make	dog	to drink
ear	to eat	egg	eye	to fall
far	fire	fish	flesh/meat	fly
to give	to go	good	hair	hard
to hear	heavy	to hide	to hit/beat	horn
house	in	knee	to know	to laugh
leaf	leg/foot	liver	long	louse
mouth	name	navel	neck	new
night	nose	not	old	one
rain	red	root	rope	to run
salt	sand	to say	to see	shade/shadow
skin/hide	small	smoke	soil	to stand
star	stone/rock	to suck	sweet	tail
to take	thick	thigh	this	to tie
tongue	tooth	water	what?	who?
wide	wind	wing	wood	yesterday

1. **Resistance to borrowability**. Out of thousands of lexical items of the 41 languages, the researchers of the languages judged which are loanwords and which are not. And the lexical concepts that were less likely loanwords were given more credit to be selected as part of the wordlist.
2. **Non-analyzability**. The LJ List was to exclude meanings that are likely to be represented by compounds rather than individual morphemes. Hence, there are 'who?' and 'what?' in the list but not other interrogative pronouns, such as 'where?' or 'why?' which, according to Tadmor (2009), are likely to be polymorphemic (as e.g. 'what-place?' 'for-what?').
3. **Universality**. The LJ List needed to be a list of concepts that are found in all (or most) human societies. Thus, animals that are found in all human societies, such as 'dog', 'fish', 'fly', 'louse', and 'ant' are up on the list, but not the animals that are (or were) absent in some societies, such as 'cow', 'pig', or 'cat'.
4. **Stability**. Lexical items that have existed in a language for a longer time were given more credit to be chosen as part of the list than those which have only existed for a shorter time.

The usage of this list thus minimizes the etymological relatedness between the morphemes in the database. The list's non-analyzability is also suitable for

compiling a database containing only single morphemes and no polymorphemic words. Lastly, its universality assures that most of the sample languages, which are typologically diverse, will not lack a morpheme for a given meaning because the speakers of a language do not lexically represent that meaning.

The four factors are gradual, not absolute. The LJ List's non-borrowability does not mean that there are no loan-morphemes in my data. I have not excluded loan-morphemes because it is impossible for me to detect all the loan-morphemes, especially within the lesser-studied languages, and it is inconsistent to pick out loan-morphemes only from the languages I have diachronic knowledge of. Additionally, of the 66 languages studied in this paper, not all have a non-analyzable morpheme for every single meaning of the LJ List: in some cases, a language has only analyzable compounds for a given meaning.

3.2 Sample languages

In order to avoid the risk of morphemes being phonetically similar because they are cognates, it was necessary that I select languages from different language families. I made the list of the 66 largest language families (based on the number of speakers), based on the data from the 20th Edition of Ethnologue (Simons & Fennig 2017), excluding pidgins, creoles, mixed languages, unclassified languages, sign languages, and the constructed language Esperanto. Each language isolate was counted as a language family with a single language member. I have selected the largest language of each language family (based on the number of native speakers), i.e. Spanish selected out of Indo-European, Mandarin selected out of Sino-Tibetan, and so on.

Why pick the largest language from each family? Since larger languages usually have more bibliographic resources available, selecting the largest language makes the data easier to collect and more credible. Moreover, relying on a factor such as native speaker population, which is largely irrelevant to the nature of the language per se, avoids making the sample overrepresent languages with a specific language character (e.g. tonal languages). Blasi et al. also attested that population size of each language was not a significantly relevant factor for their results (p. 4). Having a systematic criterion - arbitrary, but systematic - also helps to avoid the risk of cherry-picking languages for the sake of proving my hypothesis.

One disadvantage of this sampling is that areal distribution wasn't taken into consideration. Languages that are areally close to each other may influence each other's lexicon. Among my sample languages, several languages are areally clustered in Mexico and Papua New Guinea. In other areas, however, the

Figure 1: Distribution map of sample languages. Longitude and latitude of each language were retrieved from Glottolog 3.0 (Hammarström et al. 2017). Map created with Microsoft Excel 2016.

languages are fairly well spread out from each other. The distribution is shown in Figure 1.

3.3 Database

Unlike the database of Blasi et al., which contains polymorphemic words, my database consists only of morphemes. A language may have more than one morpheme to express one meaning. For each meaning per language, I have listed up to three morphemes (or none, if a language has no corresponding morpheme for a meaning). I have only selected morphemes that are used in the neutral context, avoiding literary or vulgar words.

Some of the specifications I have set for certain LJ List terms are shown in Table 2.

The grammatical category of the listed morpheme may not always match with that of the English equivalent. For example, Yoruba has no verbal morpheme (to my knowledge) that means 'to laugh' but only the noun /rĩ/ 'laughter,' which I listed as 'to laugh'.

Due to data availability, I could not find all the morphemes available. There are 66 languages and 100 meanings, thus 6,600 "slots" to fill (or leave empty when a language does not have an appropriate morpheme for a meaning). Out

Table 2: Specifications for some of the LJ List terms.

Term	Specification
1SG, 2SG, and 3SG pronouns	Unmarked and/or unbound pronouns.
Back	The body part, not the direction.
Breast	What primarily refers to the breastfeeding organ, rather than what primarily refers to the gender-neutral body part (i.e. 'chest').
Child (kin term)	Can be substituted by 'son' and 'daughter' only if a language has no commonly used corresponding gender-neutral term.
In	The affix or adposition meaning "in a location or a container."
To know	To have the knowledge, not to be acquainted with a person.
Old	Both the age of an object and of a person.
Thick	The cylindrical thickness (as in "thick stick") and the surface thickness (as in "thick surface"), but not density (as in "thick hair").

of the 6,600 slots, 94 were left empty due to lack of data. The database is thus approximately 99% complete.[1]

3.4 Statistical analysis

I conducted binomial tests to examine whether phonemes in the morphemes for a given meaning tend to contain a certain [+feature] ([+f]) more frequently than what the null hypothesis would predict or less so. The null hypothesis is that the sound-meaning association is completely arbitrary, i.e. that the frequency of [+f] in the morphemes for any given meaning (M) would not be significantly different from the mean frequency of [+f] in the morphemes for all 100 meanings.

I did not conduct binomial tests on the frequency of [−f]s (e.g. [−back]) for two reasons. One, the high or low frequency of [+f] predicts (to some degree) the frequency of [−f], in reverse direction. For example, if the results show that [+back] is abnormally frequent in M, then it is likely that [−back] is abnormally infrequent in M. This predictable redundancy reduces the reasons to double the multiplicity of the statistical tests by also analyzing the negative equivalents of the positive features analyzed. Two, a [+f] is usually more descriptive about the nature of a phoneme than a [−f] is. For example, the [+labial] feature of /p/ tells us that it is

1 The lexical database, along with the list of sample languages, the bibliography of the database's sources, and the R script used for conducting the statistical analysis, is available at https://github.com/ianjoo/LJ-List.

articulated with the lips, whereas the [–labial] feature of /t/ only tells us that it is **not** articulated with the lips. In describing any concept, it is usually more helpful to know what it **is** than to know what it is **not**. This relative deficiency of descriptive power in negative features makes them less interesting for the present study to analyze.

The **frequency** of [+f] in M is defined as follows. (|| = token number)

$$Frequency_{[+f]M} = \frac{|phonemes\ with\ [+f]\ in\ morphemes\ for\ M|}{|phonemes\ in\ morphemes\ for\ M|} \quad (1)$$

The **mean frequency** of [+f] is defined as follows. (*Mn* = Meaning number *n*)

$$\overline{Frequency_{[+f]}} = \frac{\sum_{n=1}^{100} Frequency_{[+f]Mn}}{100} \quad (2)$$

Whether a phoneme has [+f] or not was judged based on the the phonological feature database PanPhon (Mortensen et al. 2016) (retrieved from https://github.com/dmort27/panphon/blob/master/panphon/data/ipa_all.csv). Not all [+f]s are common throughout the 66 languages. I did not conduct binomial tests on [+f]s that are present in less than 50 out of 66 sample languages, e.g. [+constricted glottis].

There were 17 [+f]s present in at least 50 languages. For each pair of [+f] and M, I conducted a two-way binomial test with the following parameters:

Number of trials:	Token number of phonemes in morphemes for M
Number of successes:	Token number of phonemes with [+f] in morphemes for M
Hypothesized probability of success:	Mean frequency of [+f]

There were 17*100 = 1700 binomial tests. The 1700 tests were corrected for multiple comparison by the Benjamini-Hochberg Procedure (Benjamini & Hochberg 1995) at the False-Discovery Rate (FDR) of 10%.

One weakness of this analysis is that the token number of phonemes in the morpheme(s) for M was not controlled. Thus, a language that has more tokens of phonemes in its morpheme(s) for M has more effect on the binomial test than a language that has fewer phoneme tokens in its morpheme(s) for M. For example, if the only morpheme for M in language A is /aaaaaaaaaa/ whereas the only morpheme for M in language B is /ii/, this makes the positive association between [+low] and M more likely to appear in the results than the positive association between [+high] and M, even though the percentage of phoneme tokens of /aaaaaaaaaa/ that are [+low] (100%) is equal to the percentage of phoneme tokens of /ii/ that

are [+high]. This makes languages that have longer, more numerous (max. 3) morphemes for a given meaning play a bigger role in the statistical analysis than languages that have smaller, fewer morphemes for the same meaning. The token number of phonemes in the morpheme(s) for each meaning in each language is min = 1, max = 24, mean ≈ 5.07, sd ≈ 2.99.

4 Results and discussion

Figure 2 shows the lexical meanings that show positive or negative associations with any of the phonological features at the FDR of 10%. Among the 17 features

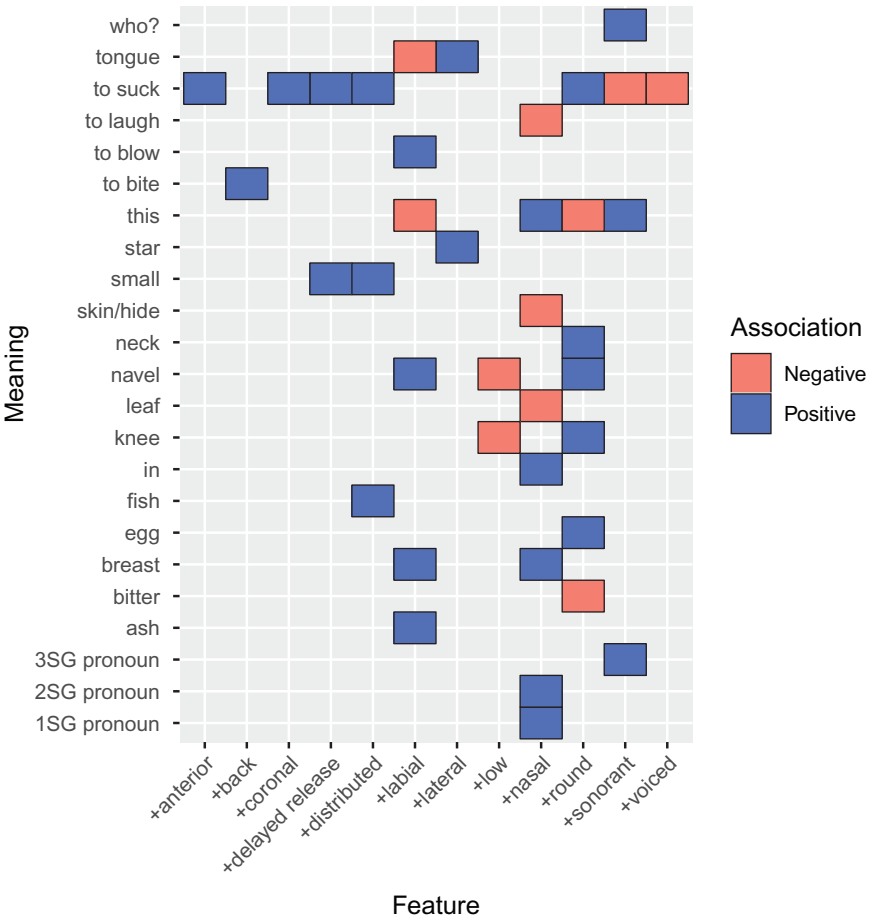

Figure 2: Biases between features and meanings

that were present in at least 50 languages, five did not show any association and are not represented in the figure: [+syllabic], [+consonantal], [+continuant], [+high], and [+tense].

'To blow' is positively associated with [+labial], as predicted by Swadesh. 'To laugh', on the other hand, is negatively associated with [+nasal], similar to actual human laughter, which usually does not involve any nasal sound (although it frequently involves the nareal fricative [n̊], the friction of airflow within the nostrils (Urbain & Dutoit 2011)). Some of the associations of 'to suck' are iconically related to actual sucking, i.e. protruding the lips ([+round]) and causing affricate-like friction ([+delayed release]). The friction is arguably more affricate-like than fricative-like, since during the articulation of an affricate, the passive and the active articulators come into contact and then are released with friction, similar to actual sucking where the sucked object inevitably touches the oral cavity, whereas during the articulation of a fricative, there is no contact between the articulators, but only narrowing. But since there is no single feature among those in PanPhon that represents a fricative, we remain uncertain regarding whether fricatives are associated with 'to suck.'

Among other meanings positively associated with [+round] are those related to round shapes, be it circular ('navel'), spherical ('egg'), hemispheric ('knee'), or cylindrical ('neck'), which is in line with previous experimental findings of the perceptual association between rounded vowels and round shapes (Ozturk et al. 2013; Fort et al. 2015). Blasi et al. also found that /u/ was positively associated with 'breast' and /o/ and /u/ with 'knee'. The positive association between 'small' and the two features [+delayed release] and [+distributed] is also in line with the tendency of the world's languages to convert non-palatal and/or non-affricate consonants into palatal affricates to express smallness and childishness (aka "expressive palatalization"), e.g. Japanese [toko-toko] 'trotting' → [tʃoko-tʃoko] 'moving like a small child' (Kochetov & Alderete 2011).

Nasals are positively associated with 1SG and 2SG pronouns. The motivation for pronominal nasality may be explained by the pre-linguistic infant's behavior to use sounds with /m/ to request for objects (Carter 1975) and seek for the attention of the caretaker (Goldman 2001). In other words, in the early years of acquisition, the (bilabial) nasal sound may be related to the concept of selfhood (1SG) and the addressee (2SG). Interestingly, Carter (1975) observed that an English-learning infant's vocal gesture /m/, used for object request, gradually evolved into the English words *more* and *mine* as the child grew and acquired English.

Another possible explanation for pronominal nasality is hinted at by the nasal preference of 'this' and 'in' shown in the results. Grammatical morphemes related to proximity ('this', 'in', 1SG and 2SG pronoun) all prefer [+nasal]. Although

some may question the proximity of 2SG, I would argue that since the hearer is usually spatially close to the speaker, 2SG is also related to proximity. Although no empirical study (to my knowledge) has demonstrated the cross-modal correspondence between nasality and proximity, it calls for future empirical studies that may verify or falsify it.

The positive association between 'star' and [+lateral] suggests the cross-modal correspondence between brightness and lateral sounds. In the perceptual experiment of Greenberg & Jenkins (1966), where 61 subjects were asked whether each of ten English consonants sounded dark or bright and to what degree, the subjects rated *L* as the most bright-sounding (5.1 on a 1 to 7 scale).

Table 3 shows the overlapping results between the present study and Blasi et al. (2016). By overlapping results I refer to cases where a segment Blasi et al. found to be positively or negatively associated with a meaning has at least one feature that I found to be associated in the same direction with the same meaning (notwithstanding that some of the "same" meanings differ slightly in detail, e.g. 'skin' in the Swadesh List used by Blasi et al. v. 'skin/hide' in the Leipzig Jakarta List used in the present study).

Table 3: Overlapping associations found in this study and Blasi et al. (2016).

Meaning	Present study		Blasi et al. (2016)	
	Positive	**Negative**	**Positive**	**Negative**
1SG pronoun	+nasal		ɲ	
Ash	+labial		u	
To bite	+back		k	
Breast	+labial +nasal		u m	
Knee	+round		o u	
Skin (/hide)		+nasal		m n
Small	+distributed +delayed release		tʃ	
Tongue	+lateral	+labial	l	u

5 Conclusion

This study reached its two research goals by (1) discovering phonosemantic associations in lexical terms not included in the Swadesh List (e.g. 'to blow' and [+labial]) and (2) providing a clearer view on sound-meaning association by showing what features are associated with each meaning (e.g. 'small' and [+delayed release]).

23 out of 100 LJ List meanings showed at least one phonosemantic association, which suggests that a significant portion of vocabulary of spoken languages has at least some phonosemantic bias. Thus, sound-meaning association in human languages is not completely arbitrary and conventional but is also significantly motivated by human cognition biases.

Funding: The research leading to these results has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 646612) granted to Martine Robbeets.

References

Benjamini, Yoav & Yosef Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57(1). 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Blasi, Damián E, Søren Wichmann, Harald Hammarström, Peter F Stadler & Morten H Christiansen. 2016. Sound-meaning association biases evidenced across thousands of languages. *Proceedings of the National Academy of Sciences of the United States of America* 113(39). 10818–23. doi: 10.1073/pnas.1605782113.

Carter, Anne L. 1975. The transformation of sensorimotor morphemes into words: A case study of the development of 'more' and 'mine'. *Journal of Child Language* 2(2). 233–250.

Fort, M, A Martin & S Peperkamp. 2015. Consonants are more important than vowels in the bouba-kiki effect. *Language and Speech* 58(2). 247–266. doi: 10.1177/0023830914534951.

Goldman, Herbert I. 2001. Parental Reports of 'MAMA' sounds in infants: An exploratory study. *Journal of Child Language* 28(2). 497–506. doi: 10.1017/S030500090100472X.

Gordon, Matthew J. 1995. The phonological composition of personal pronouns: Implications for genetic hypotheses. *Annual Meeting of the Berkeley Linguistics Society* 21(1). 117–128.

Greenberg, Joseph H & James J Jenkins. 1966. Studies in the psychological correlates of the sound system of American English. *Word* 22(1–3). 207–242.

Hammarström, Harald, Robert Forkel & Martin Haspelmath. 2017. *Glottolog 3.0*. Jena: Max Planck Institute for the Science of Human History. http://glottolog.org. Accessed: 2017-09-30.

Haspelmath, Martin & Uri Tadmor (eds.). 2009. *World loanword database*. Leipzig: Max Planck Institute for Evolutionary Anthropology. https://wold.clld.org/. Accessed: 2019-07-02.

Johansson, Niklas & Jordan Zlatev. 2013. Motivations for sound symbolism in spatial deixis: A typological study of 101 languages. *The Public Journal of Semiotics* 5(1). 3–20.

Kochetov, Alexei & John Alderete. 2011. Patterns and scales of expressive palatalization: Typological and experimental evidence. *Canadian Journal of Linguistics* 56. 345–376.

Mortensen, David R, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer & Lori Levin. 2016. Panphon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical papers*, 3475–3484.

Nichols, Johanna & David A Peterson. 1996. The Amerind personal pronouns. *Language* 72(2). 336–371. doi: 10.2307/416653.

Ozturk, Ozge, Madelaine Krehm & Athena Vouloumanos. 2013. Sound symbolism in infancy: Evidence for sound–shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology* 114(2). 173–186. doi: 10.1016/j.jecp.2012.05.004.

Simons, Gary F & Charles D Fennig. 2017. *Ethnologue: Languages of the world, Twentieth edition*. Dallas, Texas: SIL International. http://www.ethnologue.com.

Swadesh, Morris. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21(2). 121–137.

Tadmor, Uri. 2009. Loanwords in the world's languages: Findings and results. In Martin Haspelmath & Uri Tadmor (eds.), *Loanwords in the world's languages: A comparative handbook*, 55–75. The Hague: De Gruyter Mouton.

Tanz, Christine. 1971. Sound symbolism in words relating to proximity and distance. *Language and Speech* 14(3). 266–276.

Urbain, Jérôme & Thierry Dutoit. 2011. A phonetic analysis of natural laughter, for use in automatic laughter processing systems. In Sidney D'Mello, Arthur Graesser, Björn Schuller & Jean-Claude Martin (eds.), *Affective computing and intelligent interaction*, 397–406. Berlin, Heidelberg: Springer.

Urban, Matthias. 2011. Conventional sound symbolism in terms for organs of speech: A cross-linguistic study. *Folia Linguistica* 45(1). 199–214. doi: 10.1515/flin.2011.007.

Woodworth, Nancy L. 1991. Sound symbolism in proximal and distal forms. *Linguistics* 29(2). 273–300.