Information flow, artificial phonology and typology

Adamantios Gafos
University of Potsdam
gafos@uni-potsdam.de

1 Introduction

In the context of Artificial Grammar Learning (AGL) experiments, it is possible to quantify how effectively a stimulus has conveyed information and specifically the information the experimenter thinks it was designed to convey. At the most basic level, this can be done if one has access to the response variability of independent responses to the same stimulus (or subparts of the stimulus). The variability of these responses serves as an index of the amount of information that flows from the source of the stimulus to the perceiver. Quantifying information flow in this way, it is shown that under conditions where participants learn a 'natural' but not an 'unnatural' rule there are asymmetries in entropic quantities under the different conditions.

2 Information flow

In AGL, the experimenter exposes participants to patterns that may or may not reflect systematicities attested in natural languages. I exemplify with Wilson (2003) where two rules are involved. Rule 1 was a consonant harmony-like rule: /-na/ appears as the final syllable of a stem if the stem's final consonant is one of /m, n/, else /-la/ appears. Thus, stem /dume/ combines with /-na/ to give /dumena/, but /tuko/ combines with /-la/ to give /tukola/ (and so on, e.g., /binu/, /binuna/, /dige/, /digela/, /dabu/, /dabula/). Likes of this rule are attested in some languages (Rose and Walker, 2011). Rule 2 was a 'random' rule, not attested in any language: /-na/ if the stem's final consonant is one of /k, g/, else /-la/: thus, /dume/, /dumela/, /tuko/, /tukona/, /suto/, /sutola/, /binu/, /binula/, /dige/, /digena/, and so on. For both rules, the exposure phase consisted in a mere twenty stem-suffix presentations, repeated twice. Wilson's results provided evidence that rule 1 was learned (in a test phase, participants responded correctly with 'yes' to new items that conform to the rule significantly more than to new items that do not conform to the rule) but rule 2 was not. A basis of such results has so far remained unclear (for valuable discussion, see Greenwood, 2016; Moreton and Pater, 2012a,b). What is the nature of the bias favoring rule 1 over 2?

I begin by considering how well the acoustics of the stimuli used in the experiment above specify the intended phonemes. Producing and, most relevant to AGL studies, perceiving words are complicated events. Any stimulus presented aurally in an AGL experiment does not exist, in and of itself, outside of the context of perception-production cycles. How well any given sequence of symbols, for instance /dumena/ as intended by the experimenter, has conveyed the information it was designed to convey can be empirically and quantitatively assessed. To preview the analysis: hearing nasalization specifies exactly the class of phonemes /m n/, that is, constrains or reduces the alternatives to just /m n/ (I justify why and how this can be said to be true in the forthcoming). Hearing an oral stop closure as in /k g/, on the other hand, specifies at first a broader class: /p b t d k g/; further choices are needed to home in on /k g/. In a processing model, one would go on to specify the further steps needed to home in on /k g/ with perceptually salient features such as nasality said to be detected first, followed by weaker features such as place of articulation. However, the approach I adopt and its relevant quantities are invariant with respect to processing assumptions in a profound sense which need not be elaborated on here as it does not affect the validity of the ensuing demonstration.

To obtain a (much needed in artificial phonology) quantitative handle, I move to the go-to source for how well the acoustics specifies classes of consonants. This is the classic Miller and Nicely (1955) study, henceforth MN55,

which offers confusion matrices for (English) consonants under different signal-to-noise ratio and filtered speech conditions. Examination of the MN55 tables indicates that, across all signal-to-noise ratios (SNRs), including those where noise is negligible, the set of alternative responses to /k g/ is more populated and their frequencies are amplified compared to (alternative responses to) /m n/. To wit, consider MN55 table II; stimulus /ka/ is heard as /ka/ 62 times and as /ga/ 1 time out of a total of 236 /ka/ stimulus presentations; /ga/ is heard as /ka/ 1 time and as /ga/ 29 times out of 240 /ga/ presentations. Much of the time, then, /k g/ were heard as other consonants. Now, for /m n/, stimulus /ma/ is heard as /ma/ 109 times and as /na/ 60 times out of 212 /ma/ presentations; /na/ is heard as /ma/ 84 times and as /na/ 145 times out of 260 /na/ presentations; the nasals are heard predominately as nasals. In other words, the set of alternative responses to /k g/ is far more populated and their frequencies are amplified compared to /m n/. In more formal terms, the question which class of consonants (from the /m n/-based versus /k g/-based rules above) do listeners most reliably map to the intended (by the experimenter) set of consonants can be expressed as: which of the two classes, /m n/ versus /k g/, has higher information flow, I(X|Y), from source to listener. For two random variables X, Y, information flow (or mutual information) is defined as the original (unconditional) uncertainty of X, when we know nothing about Y, minus the conditional uncertainty of X given Y. Formally, I(X|Y) = H(X) – H(X|Y), where X is the perceptual category cashed in by the participant in the AGL study, Y is the stimulus, H(X) is the entropy of X (Shannon, 1948), and H(X|Y) is the conditional entropy of X (what is perceived) given the stimulus Y. The higher the I(X|Y), the more information flows from source to listener – a measure of the reduction of alternatives that the stimulus imposes on what listeners perceived.

Figure 1 quantifies information flow on the basis of the MN55 datasets for the /m n/- versus /k g/-based rules. This quantification is based on 24000 datapoints (all six MN55 tables, 4000 datapoints per table). Figure 1 shows that information flow for /m n/ is consistently higher than for /k g/: class /m n/ is more strongly associated with participants' perceiving /m n/ than class /k g/ is associated with participants' perceiving /k g/. There is thus a robust asymmetry between the assimilation and the random rule throughout all MN55 conditions.

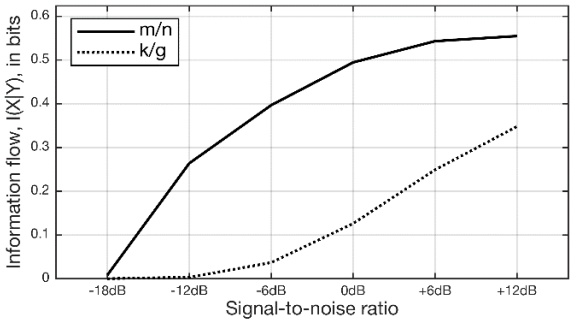

Figure 1: Information flow, I(X|Y) = H(X) – H(X|Y), for two rules based on two different classes of sounds, /m n/ and /k g/ (see text for details).

3 Some implications and other measures

In answer to the question of what may be a basis for the results obtained in AGL studies on phonological patterns, I have proposed that one quantifiable basis is information flow.

More broadly, there are at least two preconditions on rules. First, rules must be learnable by the child, that is, adapted to the cognitive skills (and limitations) of the individual. Second, the patterns encoded in rules must be transmittable or reproducible. In principle, two rules may both be learnable by individuals under sufficient input, but one may not be as reproducible as the other in the sense shown in the preceding. That is, the transmittability of sound patterns, e.g., how well the intended sets /m n/ or /k g/ reduce the choices among alternatives at the perceiver's side, reflects their replicability and thus whether rules with these patterns are likely to be attested in languages.

Practitioners of the AGL paradigm will likely consider an account along the lines given in Section 2 as a 'channel' account. This is partly correct. Any AGL stimulus must be encoded in some form and this encoding, whatever its details turn out to be, is subject to short term and longer term effects at nested time scales including the very short time scale of the current stimulus, the longer time time scale of the exposure phase, and the still longer time time scale of lexical statistics. Thus 'early perception' of any given stimulus includes effects from all these time scales. A related matter concerns the space of hypotheses entertained by the learner. During exposure, participants in the experiment reviewed in Section 2 listen to /dumena/, /digela/, /binuna/, /sutola/ and so on. With each stimulus presentation, certain syntagmatic intra-stimulus relations are strengthened more than others because they ride on the presentation of (almost) each stimulus: the

constraint ‘a nasal is followed by a nasal’ is strengthened more than ‘a coronal is followed by a coronal’ as in /sutola/ or ‘two back round vowels are followed by /a/’ as in /sutola/ and /tukola/ (but not /binuna/) which in turn is strengthened still more than ‘/dumena/ is a word’. Stimulus recurrence adds crucial detail: don’t care what consonant starts a word, don’t care what vowel follows the first vowel, and others. At issue is the number of such constraints entertained by the learner, that is, the size of the hypothesis space. Foundational results in computational learning theory (Valiant, 1984) tell us that the accuracy in learning is a function of the (log of the) cardinality of the hypothesis space as well as the number of examples. A larger hypothesis space results in worse learning outcomes (a worse upper bound on the so-called generalization error on unseen data) assuming the same number of training examples (more examples improves the error). Note how perception of /m/ or /n/ as /m/ or /n/ (in either order) but not as other consonant(s) reduces the hypothesis space. There is an interplay between perception and learning mechanisms and, to my knowledge, next to no systematic studies addressing this issue in AGL exist (but see Cristia et al., 2013). This seems to be an important consideration for future research. See also Wilson (2006) and White (2017) on how perception may play out in models of the learner.

I turn next to clarify some formal aspects of the main notion implicated in Section 2, information flow. This notion is a special instance of another, ultimately also useful, notion of information gain. Let $p(x)$ be the distribution of a pronounced symbol (this can be an intended phoneme or an intended feature of a stimulus) and $q(x)$ that of one of its contrasting alternatives. We think of symbols (in the context of Section 2, symbols are consonants) as distributions, because every intended symbol is cashed in as a distribution of potential outcomes on the side of the perceiver. This is in fact the data a confusion matrix provides us with: any row in such a matrix is a probability distribution of one category, say, /ba/, being perceived as one of several alternatives (/ba/, /pa/, /ma/, /da/ and so on in the columns of the matrix). I have effectively proposed in Section 2 that an appropriate measure of quantifying how much information participants gain in the exposure phase is the quantity known as information gain:

$$D[q(x)||p(x)] = \sum_x q(x) \log \left[\frac{q(x)}{p(x)}\right] \quad (1)$$

Information gain quantifies the expected amount of surprise or distortion when perceiving $q(x)$ while intending to convey $p(x)$. Kullback-Leibler divergence is also used for the same quantity (hence the D in $D[q(x)||p(x)]$). Unlike information flow, which is symmetric, I(X|Y) = I(Y|X), information gain is asymmetric.

I illustrate information gain with one example. Infants look longer at the picture of the object referred to by a word when a labial-initial word is misspoken with a coronal than when a coronal-initial word is misspoken with a labial, e.g., /poes/ → /toes/ or /bal/ → /dal/ versus /teen/ → /peen/ or /duif/ → /buif/ (van der Feest, 2007: 109-110). A coronal to labial change results in a different response than a labial to coronal change. Using information gain, it can be shown that a p, b → t, d change has higher expected surprise than a t, d → p, b change. However, in keeping with AGL, the example I will use to demonstrate information gain derives from White (2014) who shows that adult speakers of English exposed to a /t/ → [ð] alternation innovate this to a /d/ → [ð] and a /θ/ → [ð] alternation during test. A more specific result was that participants trained with /t/ → [ð] innovated to a /d/ → [ð] more than they did to a /θ/ → [ð] alternation. White (2014) proposes that innovation rates call on implicit knowledge of how perceptually similar the sounds in the innovated alternation are. To index similarity, White (2014) uses mutual confusability, defined as the average of the proportion of times two phonemes are confused with each other. Mutual confusability (MC) of two phonemes ‘a’ and ‘b’ is a symmetric quantity, that is, MC(a,b) = MC(b,a). Information gain is asymmetric. White (2014) extracts MC values from the perceptual confusion tables of Wang and Bilger (1973) which align well with the results of his AGL experiment, i.e., innovation percent for /a/ → [b] scales with MC(a,b). However, MC values derive from averages across SNRs (Wang and Bilger unfortunately do not give per SNR confusion matrices). Averaging across SNRs uniformly is not optimal as noise at different SNRs affects spectral and temporal cues (involved in the alternation pairs in this AGL study) differently (Jiang et al., 2006). A more stringent test of White’s proposal is to use information gain with a per SNR analysis. The predictions are that the divergence for /t/ → [ð] should be higher than for /θ/ → [ð] which in turn should be higher than for /d/ → [ð]: $D[ð||t] > D[ð||\theta] > D[ð||d]$. Figure 2 verifies these inequalities with the MN55 datasets. The asymmetries are present throughout

the different SNRs and expectedly weaken at the most favorable listening condition (+12dB).

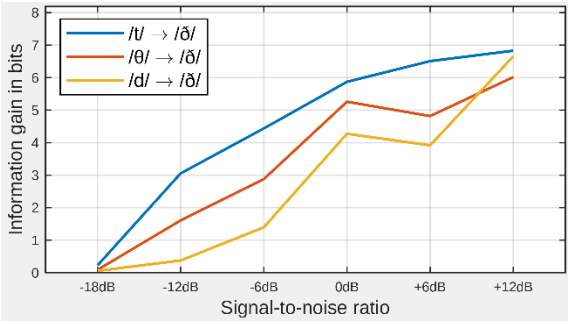

Figure 2: Divergence for three alternations: /t/ → [ð], /θ/ → [ð] and /d/ → [ð]. See text for details.

In sum, a more stringent test of the proposal in White (2014) confirms that proposal. The test is more stringent because the results are based on a per SNR analysis with information gain. Furthermore, this metric is applicable to this case as well to cases of asymmetric directional sensitivities (as in labial to coronal versus coronal to labial, which I cannot demonstrate here) whereas MC is applicable only in the former case.

4 Relation to other approaches

In the context of AGL, Pothos (2010) first used a notion of entropy to quantify the degree of compatibility between a test stimulus and a set of training stimuli. The approach requires 'dividing the [test: AG] item into parts' and quantifying the uncertainty of continuations between these parts given the statistics of the training stimuli. Two reasons make this approach not applicable to our domain. First, the proposed metric of compatibility is silent in the domain of asymmetries obtained in artificial phonology rule learning. Take, for instance, the stimuli in the experiment discussed in Section 2. These are not amenable to the same analysis as in Pothos (2010). The metric of compatibility in Pothos (2010), namely, the 'entropies of the test items' do not differ between the two rules (if we are to use phonemes or features as the correspondents to the symbols of the approach promoted in Pothos). I use quotes here because the concept (within the quotes) is not endemic to Shannon's theory. Entropy is a global property of a set of events or stimuli (or distributions over stimuli properties). It is not a notion that applies to individual test items (surprise is such a notion).

The second reason is more important. The tasks wherein the approach of Pothos has shown considerable success involve grammars defined over arbitrarily–chosen and arbitrarily–combined features such as visual stimuli of lines or shapes or strings of letters mixed with numbers and so on. Issues of 'stimulus format' are largely external to the paradigm (Pothos, 2010: 7). When it comes to spoken words and the rules of natural phonologies, such issues become primary. Linguistic percepts are not linear combinations of immutable symbols. Crucially, the places where immutability breaks down (most notably, coarticulation and misperception thereof) happen to be the breeding grounds of natural phonologies (Ohala, 1981).

Yet Pothos (2010) remains an important contribution to the AGL paradigm outside of the speech domain and has served as an inspiration for new theoretical developments on language acquisition that employ notions of entropy to account for other results or propose novel experiments that sharpen ideas (see especially Radulescu et al., 2019).

Finally, notions of information and entropy are being explored in all aspects of linguistic inquiry, and the reader is encouraged to consult, among others, Hale (2016) for a pedagogic exposition with a focus on sentence parsing, as well as as Aylett and Turk (2004), Currie-Hall (2009), Cohen-Priva (2015), Culbertson et al. (2020), Graff (2012), Hume et al. (2011), Jaeger (2010), Keller (2004), Levy (2008), Martin and Peperkamp (2017), Milin et al. (2009), Piantadosi et al. (2011, 2012), Radulescu et al. (2019), Seyfarth (2014), and Shaw and Kawahara (2019).

5 Conclusion

Languages and their speakers are systems of many degrees of freedom and strong interactions among their components. We currently lack the tools to analyze them at this level of description. Yet there are properties of these systems that are so fundamental, linguists can feel them in their bones; for example, the fact that languages show macroscopic simplicities in terms of the form of the rules they exhibit. These are properties that we cannot compute directly by taking into account all interactions playing out in the development of a language's phonology. It is here where entropic measures come to the rescue. For large enough datasets (e.g., MN55), such measures and their attendant theory (Shannon, 1948) offer ways via which one can see with tractable calculations how these phenomena take place.

Acknowledgments

This work benefited from the feedback of three anonymous reviewers. Special thanks go to Barbara Höhle, Tom Fritzsche, Sara Finley, Stephan Kuberski, Natalie Boll-Avetisyan, Shihao Du, and James White. Preparation of this manuscript was supported during its early stages by ERC Advanced Grant 249440, which provided a focused time period for research, and in its later stages by a German Research Foundation (DFG) grant, Project ID 317633480, C03.

References

Enes Avcu and Arild Hestvik. 2020. Unlearnable phonotactics. *Glossa*, 5(1):1-22. https://doi.org/10.5334/gjgl.892.

Matthew Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47(1):31-56. https://doi.org/10.1177/00238309040470010201.

Uriel Cohen Priva. 2015. Informativity affects consonant duration and deletion rates. *Laboratory Phonology*, 6(2):243-278. https://doi.org/10.1515/lp-2015-0008.

Kathleen Currie Hall. 2009. *A probabilistic model of phonological relationships: from contrast to allophony*. Doctoral dissertation, The Ohio State University.

Jennifer Culbertson, Marieke Schouwstra, and Simon Kirby. 2020. From the world to word order: deriving biases in noun phrase order from statistical properties of the world. *Language*, 96(3). https://doi.org/ 10.1353/lan.2020.0045.

Alejandrina Cristia, Jeff Mielke, Robert Daland, and Sharon Peperkamp. 2013. Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, 4(2):259-285. https://doi.org/10.1515/lp-2013-0010.

Gary S. Dell. 1984. Representation of linear order in speech: evidence for the repeated phoneme effect in speech errors. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10:222-233. https://doi.org/10.1037//0278-7393.10.2.222.

Gary S. Dell. 1986. A spreading activation theory of retrieval in sentence production. *Psychological Review* 93:283-321. https://doi.org/10.1037/0033-295X.93.3.283.

William K. Estes. 1972. An associative basis for coding and organization in memory. In A. W. Melton & E. Martin (eds.), *Coding processes in human memory*. Washington, DC: Winston, pages 161-190.

Alejandrina Cristia, Jeff Mielke, Robert Daland, and Sharon Peperkamp. 2013. Similarity in the generalization of implicitly learned sound patterns. *Laboratory Phonology*, 4(2):259-285. https://doi.org/10.1515/lp-2013-0010.

Sara Finley. 2011. The privileged status of locality in consonant harmony. *Journal of Memory and Language*, 65:74-83. https://doi.org/10.1016/j.jml.2011.02.006

Victoria A. Fromkin. 1971. The non-anomalous nature of anomalous utterances. *Language*, 47(1):27-52.

Adamantios Gafos. 1996a. Correspondence in Temiar: No need for long distance spreading here. In W. de Reuse and S. Chelliah (eds.), *Papers from the Fifth Annual Meeting of the South East Asian Linguistics Society*, Tucson, AZ: Arizona State University, pages 30-47.

Adamantios Gafos. 1996b[1999]. *The articulatory basis of locality in phonology*. Doctoral dissertation, Johns Hopkins University. [Published 1999, Outstanding Dissertations in Linguistics, Routledge Publishers.]

Diamandis Gafos. 1998. Eliminating long-distance consonantal spreading. *Natural Language and Linguistic Theory*, 16(2):223-278. https://doi.org/10.1023/A:1005968600965.

Adamantios Gafos. 2003. Greenberg's asymmetry in Arabic: a consequence of stems in paradigms. *Language*, 79(2):317-357. https://doi.org/10.1353/lan.2003.0116.

Adamantios Gafos. 2021. Consonant harmony, disharmony, memory and time scales. *Proceedings of the Society for Computation in Linguistics*: Vol. 4, pages 188-205.

Wendell R. Garner. 1974. *The processing of information and structure*. Psychology Press.

Peter Graff. 2012. *Communicative efficiency in the lexicon*. Doctoral dissertation, MIT.

Anna Greenwood. 2016. *An experimental investigation of phonetic naturalness*. Doctoral dissertation, University of California, Santa Cruz.

John Hale. 2016. Information-theoretical complexity metrics. *Language and Linguistics Compass*, 10(9): 397-412. https://doi.org/10.1111/lnc3.12196.

Gunnar Hansson. 2001. *Theoretical and typological issues in consonant harmony*. Doctoral Dissertation, University of California, Berkeley.

Jonathan Harrington, Felicitas Kleber, Ulrich Reubold, Florian Schiel, Mary Stevens. 2018. Linking cognitive and social aspects of sound change using

agent-based modeling. *Topics in Cognitive Science*, 10(4):707-728. https://doi.org/ 10.1111/tops.12329.

Jeff Heinz. 2010. Learning long-distance phonotactics. *Linguistic Inquiry*, 41(4):623-661. https://doi.org/10.1162/LING_a_00015.

Elizabeth Hume, Kathleen Currie Hall, Andrew Wedel, Adam Ussishkin, Martine Adda-Dekker, and Cédric Gendrot. 2011. Anti-markedness patterns in French epenthesis: An information-theoretic approach. In *Annual Meeting of the Berkeley Linguistics Society*, vol. 37, no. 1, pp. 104-123. https://doi.org/10.3765/bls.v37i1.3196.

Florian T. Jaeger. 2010. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, 61(1):23-62. https://doi.org/ 10.1016/j.cogpsych.2010.02.002.

Jiang Jingao, Marcia Chen and Albeer Alwan. 2006. On the perception of voicing in syllable-initial plosives in noise. *The Journal of the Acoustical Society of America*, 119(2): 1092-105. https://doi.org/ 10.1121/1.2149841.

John Kingston and Randy L. Diehl. 1995. Intermediate properties in the perception of distinctive feature values. In Bruce Connell and Amalia Arvaniti (eds.), *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, Cambridge U.P., Cambridge, England, pages 7-27.

Regina Lai. 2015. Learnable vs. unlearnable harmony patterns. *Linguistic Inquiry*, 46(3): 425-451. https://doi.org/10.1162/LING_a_00188.

Catherine L. Lee and Estes K. William. 1977. Order and position in primary memory for letter strings. *Journal of Memory and Language*, 16(4):395-418. https://doi.org/10.1016/S0022-5371(77)80036-4.

Roger Levy. 2008. A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the 13th conference on empirical methods in natural language processing*, Waikiki, Honolulu, pages 234-243.

Alexander Martin and Sharon Peperkamp. 2017. Assessing the distinctiveness of phonological features in word recognition: Prelexical and lexical influences. *Journal of Phonetics*, 62:1-11. https://doi.org/10.1016/j.wocn.2017.01.007.

Robert S. McLean and Lee W. Gregg. 1967. Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology*, 74(4):455-459. https://doi.org/10.1037/h0024785.

Laura McPherson and Bruce Hayes. 2016. Relating application frequency to morphological structure: the case of Tommo So vowel harmony. *Phonology*, 33(1), 125-167. https://doi.org/10.1017/S0952675716000051.

Jeff Mielke. 2004[2008]. *The emergence of distinctive features*. Doctoral dissertation, The Ohio State University. [Published 2008, Oxford University Press].

Petar Milin, Victor Kuperman, Aleksandar Kostic, and Harald R. Baayen. 2009. Paradigms bit by bit: An information theoretic approach to the processing of paradigmatic structure in inflection and derivation. *Analogy in grammar: Form and acquisition*, pages 214-252.

George A. Miller and Patricia E. Nicely. 1958. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27:38-352. https://doi.org/10.1121/1.1907526.

Elliot Moreton. 2008. Analytic bias and phonological typology. *Phonology*, 25(1):83-127. https://doi.org/10.1017/S0952675708001413.

Elliot Moreton and Joseph Pater. 2012a. Structure and substance in artificial phonology learning, Part I: Structure. *Language and Linguistics Compass*, 6(11):686-701. https://doi.org/10.1002/lnc3.363.

Elliot Moreton and Joseph Pater. 2012b. Structure and substance in artificial phonology learning, Part II: Substance. *Language and Linguistics Compass*, 6(11):702-718. https://doi.org/10.1002/lnc3.366.

James S. Nairne. 1991. Positional uncertainty in long-term memory. *Memory and Cognition*, 19(4):332-340. https://doi.org/10.3758/BF03197136.

Terrance M. Nearey. 1995. A double-weak view of trading relations: comments on Kingston and Diehl. In Bruce Connell and Amalia Arvaniti (eds.), *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, Cambridge U.P., Cambridge, England, pages 28-40.

Ian Neath and Aimée M. Surprenant. 2003. *Human memory: An introduction to research, data, and theory*. Second edition. Belmont, CA: Wadsworth.

John J. Ohala. 1981. The listener as a source of sound change. In C. S. Masek, R. A. Hendrick, and M. F. Miller (eds.), *Papers from the Parasession on Language and Behavior*. Chicago, Chicago Linguistic Society: pages 178-203. o:.

John J. Ohala. 1995. The perceptual basis of some sound patterns. In Bruce Connell and Amalia Arvaniti (eds.), *Phonology and Phonetic Evidence: Papers in Laboratory Phonology IV*, Cambridge U.P., Cambridge, England, pages 87-92.

Athanasios Papoulis. 1984. *Probability, random variables and stochastic processes*. 3rd Edition, McGraw-Hill: New York.

Steven T. Piantadosi, Harry Tily, Edward Gibson. 2011. Word lengths are optimized for efficient

communication. *Proceedings of the National Academy of Sciences of the United States of America*, 108(9):3526-3529. https://doi.org/10.1073/pnas.1012551108.

Steven T. Piantadosi and Joshua B. Tenenbaum. 2012. Modeling the acquisition of quantifier semantics: A case study in function word learnability. Manuscript, Rochester University, MIT, and Stanford University.

Emmanuel Pothos, 2010. An entropy model for artificial grammar learning. *Frontiers in Psychology*, 1:16. https://doi.org/0.3389/fpsyg.2010.00016.

Silvia Radulescu, Frank Wijnen and Sergey Avrutin. 2020. Patterns bit by bit. An entropy model for rule induction. *Language Learning and Development*, 16(2):109-140. https://doi.org/10.1080/15475441.2019.1695620.

Sharon Rose and Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language*, 80:475-531. https://doi.org/10.1353/lan.2004.0144.

Sharon Rose and Rachel Walker. 2011. Harmony systems. In John Goldsmith, Jason Riggle, and Alan Yu (eds.), *Handbook of Phonological Theory*, Second Edition, Oxford: Wiley-Blackwell, pages 240-290.

Scott Seyfarth. 2014. Word informativity influences acoustic duration: Effects of contextual predictability on lexical representation. *Cognition*, 133(1):140-155. https://doi.org/10.1016/j.cognition.2014.06.013.

Claude E. Shannon. 1948. A mathematical theory of communication (part 1). *Bell Systems Technical Journal*, 27:379-423.

Jason A. Shaw and Shigeto Kawahara. 2019. Effects of surprisal and entropy on vowel duration in Japanese. *Language and Speech*, 62(1):80-114. https://doi.org/10.1177/0023830917737331.

Kenneth N. Stevens. 1980. Discussion during symposium on phonetic universals in phonological systems and their explanation. In *Proceedings of the 9th International Congress of Phonetic Sciences*, vol. 3, pages 181-194.

Kenneth N. Stevens and Samuel J. Keyser. 1989. Primary features and their enhancement in consonants. *Language*, 65(1):81-106. https://doi.org/10.2307/414843.

Simon Todd, Janet B. Pierrehumbert, and Jennifer Hay. 2019. Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model. *Cognition*, 185:1-20. https://doi.org/10.1016/j.cognition.2019.01.004.

Rebecca Treiman and Catalina Danis. 1988. Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(1): 145-152. https://doi.org/10.1037/0278-7393.14.1.145.

Less G. Valiant. 1984. A theory of the learnable. *Communications of the ACM*, 27(11): 1134-1142. https://dl.acm.org/doi/10.1145/1968.1972.

Suzanne V. H. van der Feest. 2007. *Building a phonological lexicon: the acquisition of the Dutch voicing contrast in perception and production*. Ph.D. Dissertation, Radboud University, Nijmegen.

Rachel Walker. 2000. Long-distance consonantal identity effects. In *Proceedings of WCCFL (19).* Somerville, MA, Cascadilla Press, pages 532-545.

Marilyn D. Wang and Robert C. Bilger. 1973. Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, 54(5): 1248-1266. https://doi.org/10.1121/1.1914417.

Adam Wayment. 2009. *Assimilation as attraction: Computing distance, similarity, and locality in phonology*. Doctoral dissertation, Johns Hopkins U.

Andrew B. Wedel. 2006. Exemplar models, evolution and language change. *The Linguistic Review*, 23(3):247-274. https://doi.org/10.1515/TLR.2006.010.

James White. 2014. Evidence for a learning bias against saltatory phonological alternations. *Cognition*, 130(1): 96-115. https://doi.org/10.1016/j.cognition.2013.09.008.

James White. 2017. Accounting for the learnability of saltation in phonological theory: A maximum entropy model with a P-map bias. *Language*, 93(1): 1-36. https://doi.org/10.1353/lan.2017.0013.

Colin Wilson. 2003. Experimental investigation on phonological naturalness. In G. Garding and M. Tsujimura (eds.), *WCCFL 2*2, Somerville, MA: Cascadilla Press, pages 533-546.

Colin Wilson. 2006. Learning phonology with a substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30:945-82. https://doi.org/10.1207/s15516709cog0000_89.

Alan C. L. Yu. 2011. On measuring phonetic precursor robustness: a response to Moreton. *Phonology*, 28:491-518. https://doi.org/10.1017/S0952675711000236.

Jesse A. Zymet. 2014. Distance-based decay in long-distance phonological processes. In *Proceedings of the 32nd West Coast Conference on Formal Linguistics*, pages 72-81.

A Appendix: Imperfect memory

A reviewer's comment offers an opportunity to bring up the additional consideration of memory as a largely neglected factor in accounting for results in AGL experiments and in assessing the import of such results for natural phonologies.

Lai (2015) demonstrates that participants fail to learn an agreement pattern involving (only) the first and last sibilant segments in trisyllabic words: thus, /ʃVsVCVʃ/ or /sVsVCVs/ conform to the pattern but /sVCVCVʃ/ or /ʃVCVCVs/ do not because the first and last sibilants disagree in [±anterior].[1] In contrast, participants succeed in learning an agreement pattern in which all sibilants are required to agree in [±anterior]. These results mirror phonological typology and are consistent with a hypothesis from Heinz (2010) on the complexity of natural language phonotactics which Lai (2015) aimed at assessing via an AGL study. The proposed interpretation of the learning asymmetry from the AGL results was that learning biases narrow the range of hypotheses entertained by learners.

An understanding of the issues surrounding such results and their potential interpretations requires examination of certain aspects of the relation between the learning scenario in the lab and harmonies in natural phonologies. Sibilant harmonies represent the most dominant (in terms of attestation frequency) example of long distance consonantal identity phenomena and a conspiracy of three distinct but convergent factors seem to explain this dominance in the realm of natural phonologies (Gafos, 2021): the propensity of the tip-blade to coarticulate (strictly locally) through vowels and neutralize the [±anterior] contrast between (pre-harmony stage) /sVʃ/-/ʃVʃ/ lexical pairs, the auditory saliency of repeated values of [±anterior] in sibilants (that is, the fact that the coarticulated output of /sVʃ/ → [ʃVʃ] is salient for listeners due to the repetition of the same value of [±anterior]), and perhaps also the propensity of planning errors in such sequences of sibilants.

The convergence of these three factors may be seen to characterize the early stages of the development of long distance identity. At later stages, processes of extension of the short-range CVC(V) context must necessarily take effect, so that the pattern ultimately ends up holding also within larger spans, as in /sVpVʃ/ → [ʃVpVʃ], wherein the trigger and the target sites are separated by more than a single vowel. The factors implicated during that transition appear to draw on the auditory saliency of repeated sibilants. That is, sequences of repeated [s] versus repeated [ʃ] present the listener-learner with a salient dichotomy in spectral energy plateaux. The wider and somewhat more retracted channel of [ʃ] results in a turbulence of lower ('dull') frequencies compared to that of higher ('sharp') frequencies [s].

To return to the AGL setting, studies of sibilant harmony in the lab lift the pattern from its natural setting by excising the first and third convergent factors discussed above (no overt production in the AGL setting) and by collapsing the different time scales over which these factors play out. We are thus left with the second factor as the locus of intersection between learning of sibilant harmony in the natural setting and in the lab. In the latter, given that participants do learn certain sibilant agreement patterns invites asking whether listeners latch on to a generalization in terms of frequency plateau (that is, a division of the stimuli into two classes along the single dimension of spectral energy, 'dull' versus 'sharp' sibilants) and whether the extent to which this may be so should (not) be equated with specifically phonological learning mechanisms. It is unclear whether such questions are decidable. In part, this is because it is unclear whether it is possible to loosen the already evolved functional couplings between specifically auditory and specifically linguistic cognition. For now, such questions can be put aside, not because they may be difficult but because there are other more pressing questions that should be asked first.

The appeal of the results from the AGL paradigm, remarkable as they may be, should be considered in the context of the challenges the paradigm is heir to. First and foremost among

[1] The reviewer also points to Avcu and Hestvik (2020) who demonstrate that, when using a more sensitive test, participants exposed to the same rule, which does not conform to the formal complexity hypothesis of Heinz (2010) about natural language phonotactics, do show positive d-prime scores, indicating that participants can learn the distinction between rule conforming versus non-conforming stimuli. Whether this is taken as evidence against Heinz (2010)'s complexity hypothesis is a matter of interpretation (as the authors indicate; Avcu and Hestvik, 2020: 17) and the matter is furthermore complicated by calling on other domain-general mechanisms (Avcu and Hestvik, 2020: 18) implicated in the subtleties of the results. Most likely, what is observed here is a trade-off between what is referred to in learning theory as sample complexity of the input (how much input is needed to learn the pattern) and accuracy of learning (Valiant, 1984; et seq.). However, there are more pressing questions to be asked (see text).

these is addressing the problem of specifying the dimensions of the space where the stimuli live in the participants' perceptual and memory systems. A second challenge, anticipated in the preceding, can be referred to here as time scale conflation. The neutralization of the lexical contrast between (pre-harmony stage) lexical pairs /sVʃ/-/ʃVʃ/ to post-harmony stage /ʃVʃ/ has its own intrinsic time scale, which is different (much slower) from the time scales of the other two factors (Gafos, 2021). Finally, memory considerations, seem to be involved. I turn to this last issue of memory in the remainder of this Appendix.

In a thoughtfully articulated application of AGL to phonological typology, Moreton (2008) shows that participants exposed to CVCV stimuli learn vowel-to-vowel height (both vowels high or both vowels non-high; henceforth HH) but not vowel height, consonant voicing (high vowel with voiced medial C or low vowel with voicelesss medial C; henceforth HV) restrictions. HH conforming stimuli were forms as in /CiCu/ (both vowels are high) or /CæCɔ/ (both vowels are non-high). HV conforming stimuli were forms as in /CidV/ (a high vowel co-occurs with a voiced consonant) or /CætV/ (a low vowel co-occurs with a voiceless consonant).

Moreton (2008) follows a long line of fruitful work where the factors responsible for sound change are perception and production (Ohala, 1981). Memory has not been considered in any systematic way as a source of selection forces in sound change. To clarify, memory does play a role in exemplar approaches wherein 'rich' memory, an all-encompassing storage of phonetic details, in concert with lexical frequency considerations, is argued to play out in the course of sound change (Wedel, 2006; Harrington et al., 2018; Todd et al., 2019, among others). Here, I mean not the rich but the fallible memory in the same way Ohala emphasized the fallible parsing of coarticulation by perception, as well as the memory that imposes structure or 'chunking' (McLean and Gregg, 1967 et seq.) on an otherwise linear order of segmental sequences.

To return to the task at hand, when properties of memory which target coherent storage chunks (e.g., syllable onsets or rhymes but not VC chunks in a CVCV as the latter straddle syllables) and classes of similar sounds (e.g., the vowels or the consonants in CVCV; Dell, 1984, 1986; Wayment, 2009; among others) are taken into consideration, both as a basis of forming generalizations but also as a basis for interference effects, there are reasons to doubt that the HH, HV patterns were equally supported in an otherwise impeccably designed set of stimuli. I only address the latter interference aspect here.

In one time-honored model of memory, interference applies to the positional encoding of similar elements so that, for instance, the two consonants in a CVCV or the two vowels may exchange their positions (Estes, 1972; Lee and Estes, 1977; Nairne, 1991; Neath and Surprenant, 2003). The crucial observation is that positional swaps affect the strength of the generalization (intended by the experimenter) in the HV but not the HH pattern. Swapping two high or two non-high vowels in a CVCV does not violate the HH pattern; after swapping, the vowels in /CiCu/ (both high) or /CæCɔ/ (both non-high) still agree in height. In contrast, swapping the vowels or the consonants in /CidV/ or /CætV/ may affect height-voicing agreement, because in the training stimuli the voicing of the 'irrelevant' first C was not made to depend on the height of the vowel and the height of the 'irrelevant' second V was not made to depend on the voicing of the medial consonant. The exact extent to which interference weakens the HV generalization is at the mercy of the random choices of the non-controlled C and V in these stimuli. What is clear is that whereas the strength of the evidence for the HV pattern is affected, that for the HH pattern is not. Memory interference mechanisms thus affect the encoding of phonological forms and may contribute to what Moreton and Pater (2012b) refer to as structurally-biased phonology.

We still need to explain why HH patterns are well attested in languages but HV patterns are not, Moreton's underphonologization discovery. Yu (2011) argues that properly assessing the potential of the phonetic pressures behind the HH versus the HV pattern to promote sound change requires perceptual confusability judgments. Such data are extremely valuable but hard to acquire (MN55 tell us that 'tests lasted several months') due to the number of repetitions required to provide representative error rates. Section 2 is a demonstration of what can be expected by an approach along the lines of what Yu advocates when such data are available. In the absence of such data, Yu used production data to estimate parameters of an identification function indexing

the degree of uncertainty imposed by a context on a vowel's identity. When so indexed, the strength of the phonetic pressures is higher for HH than HV. The height-height effect results in more uncertainty in perceptual categorization than the height-voicing effect and thus, arguably, increased likelihood of misperception leading to an HH than an HV pattern as per typology.

Here, I propose a different, non-exclusive consideration that identifies another basis for the typological HH versus HV asymmetry in the perceptual integration potential and temporal span of the cues involved in these phonetic pressures. The phonetic pressure behind the HH pattern is vowel-to-vowel coarticulation. This is a so-called context effect. In contrast, the phonetic pressure behind the HV pattern corresponds to what is known as a trading relation (on the distinction between context effects and trading relations, see especially Repp, 1982: 87-88). In HV, the spectral cue to vowel height is F1. F1 does not constitute a direct cue to voicing perception of the adjacent consonant; rather, F1 is perceptually integrated with another temporal cue (stop closure duration; Nearey, 1995; Kingston and Diehl, 1995) and does not remain audible as a separate phonetic event corresponding to an entire segment. It may be part of a segment, the short-lived span at the end of the vowel, but not the whole segment.

In contrast to HV, for HH the vowel's height cues and in particular its F1 is not perceptually integrated with the next vowel (the target of coarticulation). The vowel and its cues remain audible as a separate segment. This provides for the HH but not for the HV case an ever-present, robust, whole segment source of coarticulation, the key requirement for getting sound change off the ground (Ohala, 1981).[2]

Two sets of factors carry the weight of the explaining done in the above. One plays out in the AGL setting; the other plays out in the setting of phonological rule development in natural languages. The two sets are non-overlapping. This underscores the challenges met by AGL in informing natural phonologies (for further discussion of this issue, see Moreton and Pater, 2012b: 710 ff.).

To return to the finding from Lai (2015), it would seem reasonable to assess alternative and specifically memory-based explanations of the lack of robust learning in that experiment. The crucial sites in the trisyllabic stimuli, such as /ʃVsVCVʃ/, in that experiment are the first and last segments. These sit in non-adjacent syllables, which in turn belong to different feet, and within these structures the segments referred to in the identity relation occupy distinct syllabic roles. Both structure and distance considerations are involved. There are broad sources of converging evidence from psycholinguistics and theoretical phonology on the role of linguistic structure in grammar and processing (Fromkin, 1971; Dell, 1984, 1986; Treiman and Danis, 1988; Wayment, 2009) as well as evidence that a notion of distance is involved in the formal non-local mechanism of effecting identity, namely, the notion of correspondence (McCarthy and Prince, 1995; Gafos, 1996ab, 1998, 2003; Walker, 2000; Hansson, 2001[2010]; Rose and Walker, 2004; Arsenault and Kochetov, 2008). Thus, it seems sensible to examine the extent to which these results may be attributed to memory-based factors (Gafos, 2021) in a learning mechanism which adjusts the strength of the feature co-occurrence restriction *[+anterior]…[−anterior] as a function of the distance between the two sites (see Zymet, 2014 for this latter part). The hypothesis that the learner is equipped with such principles is consistent with findings that AGL participants who acquire a short span agreement pattern, where target and trigger sites are separated by one vowel, do not innovate to agreement at a longer span as robustly as participants who learn a longer span agreement pattern innovate to a shorter span (Finley, 2011).

In sum, there are indications that imperfect memory plays a role in AGL. Incorporating memory principles in models of the learning mechanism would enable careful evaluation of different interpretations of the evidence (about the learner) the AGL paradigm is so effective at providing.

[2] Moreton (2008) reviews production data indicating that the size of the phonetic effect is stronger in (example studies of) the HV than in (example studies of) the HH pattern (using a different approach from Yu 2011). This may very well be true. However, an effect's magnitude is orthogonal to the nature of the effect (trading relation versus context effect).