University of Pennsylvania Working Papers in Linguistics

Volume 7
Issue 1 *Proceedings of the 24th Annual Penn Linguistics Colloquium*

Article 4

1-1-2000

Ramifications of Phonology-Syntax Interactions for Phonological Models

Benjamin K. Bergen

Ramifications of Phonology-Syntax Interactions for Phonological Models

Ramifications of Phonology-Syntax Interactions for Phonological Models[*]

Benjamin K. Bergen

1 Syntax-Phonology Relations are not Arbitrary

This paper presents evidence that, contrary to the typically assumed *arbitrariness of the sign*, probabilistic correlations exist between syntactic and phonological properties of lexical items. Moreover, language users make use of these correlations during language processing. Deterministic linguistic models cannot account for this behavior, but the processing properties emerge naturally in linguistic models which allow the assignment of probabilities of application to linguistic generalizations. This paper presents a Belief Net model in which probabilistic asymmetries in processing arise from the representation of probabilistic distributions of English phonosyntactic generalizations. Such a model has the desirable properties of being neurally-plausible and cleanly learnable at the connectionist level.

1.1 Arbitrariness

The *arbitrariness of the sign* (Saussure 1916) is a doctrine which implicitly underlies most linguistic theories. It holds that the form of linguistic units, for example, words, is completely arbitrary; there is no deterministic relationship between what a word means and what phonological form it takes. Clear examples of lexical arbitrariness can be found by simply comparing monomorphemic words signifying similar concepts across unrelated languages. Aside from observing infrequent and controversial sound-symbolic lexical properties, doing so suggests that knowing a word's meaning does not permit us any insight into its form or vice versa. On the basis of the arbitrariness of the sign, most linguistic models conclude that there need be no direct relation between phonological and semantic properties of words.

The *phonosyntactic arbitrariness of the sign* is a related tenet, which holds that phonological properties of lexical items are arbitrary relative to their syntactic properties. However, contrary to arbitrariness assumptions, lexica display regular correlations between syntactic properties and

[*] Thanks are due to Mark Paskin for his technical assistance and to Steve Chang for helpful commentary. All errors and omissions are my own.

phonological ones that do not belong to the domain of predictable or rule-governed morphology, and yet are not entirely arbitrary either. At least three types of such *phonosyntactic* generalizations can be described.

- **'Strict' grammatical category restrictions on the distribution of phonological elements.** For example, word-initial [ð] is claimed to be restricted in English to function words.[1]
- **Sub-morphemic elements correlated with morphosyntactic category.** Certain English past-tense and past-participle strong verbs seem to be best analyzed as category-specific schemata, rather than as derivations from (heterogeneous) base forms (Bybee and Moder 1982).
- **Statistical asymmetries in the distribution of phonological elements in grammatical categories.** Phonological properties like stress (Davis and Kelly 1997) and vowel quality (Sereno 1994) are distributed in unequal proportions in English verbs and nouns. The present article addresses this type of phonosyntactic generalization.

1.2 The Phonology of Syntactic Classes

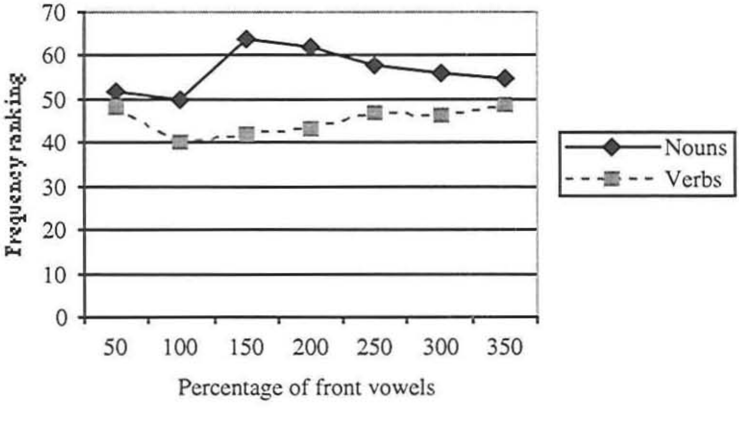


Figure 1: Front vowels in frequent English words (Sereno & Jongman 1990)

The English lexicon shows subtle but significant asymmetries in the distribution of phonological features across grammatical categories. For ex-

[1] However, I know of at least one attested use of voiced [ð] word-initially in an open-class word, *this* used as a verb, in its sense as a piece of Java jargon.

ample, in a survey of the Brown Corpus (Francis and Kucera 1982), Sereno and Jongman (1990) found frequent English verbs to more often have front vowels than not, while they found the reverse for frequent nouns (Figure 1).

This distributional asymmetry is of little interest unless it is shown to be part of linguistic knowledge. A small set of psycholinguistic studies have recently demonstrated that language users use knowledge of asymmetrical phonosyntactic generalizations during perception (Sereno and Jongman 1990, Kelly 1994, Sereno 1994, and Davis and Kelly 1997). For example, Sereno's (1994) work with the English lexicon yielded the following observations:

- Nouns with back vowels (716 ms) are categorized significantly faster than nouns with front vowels (777 ms).
- Verbs with front vowels (776 ms) are categorized significantly faster than verbs with back vowels (783 ms).

Importantly, this perceptual advantage holds not only for frequent words, but for words of all frequencies.

2 Existing Solutions

2.1 Ramifications

These findings suggest that detailed (morpho-)syntactic information is directly related to phonological information in generalizations over lexical forms. Additionally, since neither the distributions nor the processing properties are categorical in nature, these generalizations must have probabilistic properties. These two ramifications stand in direct opposition to the normal assumptions of generative phonology: that syntactic properties are irrelevant for phonological generalizations and vice versa, and that phonological generalizations are categorical, not probabilistic.

What would a model of phonological knowledge look like if it is to display the behavior described above? It would extract probabilistic correlations between information from different domains from the phonological signal, and also adapt its production to multi-modal factors impacting phonology. At the service of these functions, it would encode probabilistic, multi-modal knowledge. One such model is embodied by Variable Rules (Labov 1972).

2.2 Variable Rules

Variable Rule analysis (Labov 1972) treats variation by adding quantitative weightings correlated with social factors to SPE-style generative rules. The best-known case study treats English word-final t/d deletion (e.g. Guy 1991, i.a.). In the variable rule below, weighted phonological contexts are marked with angled brackets.

- t, d → <Ø> / <-stress> <+cons> [+cons] _ <+son>

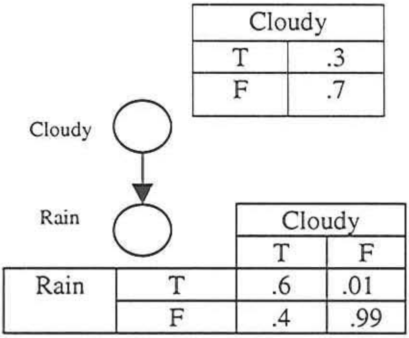
Unfortunately, Variable Rules are inappropriate for probabilistic phonosyntactic generalizations, however, since they deal with sociolinguistic and not syntactic correlates of variable phonological behavior (Fasold 1996). Moreover, extending their domain of application to syntactic variables opposes the fundamental assumption that different values of a given variable all convey the same meaning. Finally, Variable Rules disallow the interaction of constraints, but Sereno and Jongman (1990) found frequency to interact with the processing correlation between grammatical class and vocalic frontness in certain test conditions. In their study, the more frequent a word was, the more likely it was to be processed along the lines predicted by the lexical asymmetry.

Aside from Variable Rule analysis, other existing phonological frameworks have no way of capturing the probabilistic correlations described above since they assume that both linguistic representations and their combination are discrete and deterministic. The next section introduces a mechanism that can account for the properties described above through the use of probabilistic representations and interactions of phonological and syntactic knowledge.

3 Belief Nets: Aspects of the Representational Architecture

Belief Networks (BNs; Jensen 1996) are a concise and powerful computational representation of uncertain knowledge in a propositional network. They are made up of nodes with probabilities assigned to their values. Nodes are connected through causal links, and each node specifies the dependent probabilities of its values given its parents. Such a network calculates the probabilities of the values for a node, given observed values of its relatives.

In a simple example, two propositions, each with multiple possible values, stand in a causal relation (Figure 2). Cloudiness and raininess are represented as Cloudy and Rain nodes in a network, and each can be in one of two states: true or false. In more complicated cases, values will be more numerous and states of a node are continuous, rather than discrete. The unidirectional causal relation between the two propositions is represented as an arrowed link. In this network, each node has a prior probability for each of its values (which add up to 1). For example, a prior probability of cloudiness might be 0.3, thus Cloudy(true) will have a prior of 0.3 and Cloudy(false) one of 0.7. Let's imagine that the prior probability of Rain(true) is 0.1. The causal relationship between the two propositions is encoded in a probability distribution for the downstream node which captures the probability of each Rain state given each Cloudy state. If we know that Cloudy(true), let's say we have observed there is a 0.6 probability of rain, and if Cloudy(false), the probability of rain is 0.01.



Cloudy	
T	.3
F	.7

		Cloudy	
		T	F
Rain	T	.6	.01
	F	.4	.99

Figure 2: BN relating cloudiness and rain

		Cloudy	Rain
a.	True	0.3	0.19
	False	0.7	0.81
b.	True	**1**	0.6
	False	**0**	0.4
c.	True	0.96	**1**
	False	0.04	**0**

Figure 3: a. Unconditional probabilities; b. Causal; and c. Diagnostic inference.

BNs would be entirely innocuous, however, if they were not equipped with a means for performing inference on the basis of their correlative representations. Various inferencing mechanisms exist for. BNs, and all perform essentially the same function; given observed states of some subset of the nodes in a network, predictions are made about the probabilities of all other node values. In the example above, in the absence of any observation, the unconditional (prior) probabilities of the two nodes are as displayed in Figure 3a. above. If Cloudy is observed to be true (in bold), the network concludes the probability of Rain(true) to increase through causal (forward) reasoning (Figure 3b.). If Rain is observed to be true (in bold), the probability of Cloudy(true) increases through diagnostic (backwards) inference.

4 Belief Nets for Phonosyntactic Generalizations

Belief Nets are shown in this section to be appropriate for modeling the kinds of interactions responsible for the processing asymmetries described above. For example, in the network in Figure 4, one node represents the set of words known by the speaker, another the grammatical classes of those words, and a third a schematized phonological feature representing front/backness. This model assumes that in production, expressive desires evoke lexical representations, which subsequently give rise to grammatical and phonological properties. Thus, forward causative relations hold between lexical identity and grammatical or vocalic properties. Conversely, in recognition, phonological information (and some grammatical class information) is directly extracted from the speech signal, and the lexical information most likely to have caused those properties is induced.

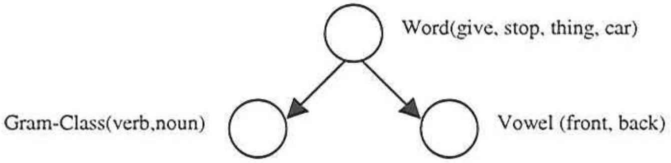


Figure 4: BN for phonological and syntactic properties in words

For such a network, the only statistically relevant distributions we will find on the basis of a data source like the Brown corpus will be over relative frequencies of the words. That is, because of the asymmetric relation between vowel quality and grammatical class, words like *give* will be on average more frequent than words like *stop* and words like *stay* more frequent than words like *thing*. Thus, if the values of the node Word are taken as representative of entire classes, then the relative probabilities of the different word types can be reflected in the probabilities of the values of Word.[2]

[2] This simplification is made for the purpose of not representing an entire lexicon

From the data in Figure 1, we can schematize the relative ratio of verbs with front vowels to verbs with back vowels at approximately 3:2, and about the same distribution for nouns with back vowels versus nouns with front vowels. Assuming that nouns and verbs are equally likely, this means that the priors of front vowel verbs and back vowel nouns are 0.3, and the others 0.2.

The values of Vowel given its parent should be straightforward; given *stop* or *car*, Vowel(back) will approach a probability of 1, while for *give* and *thing*, it will approach 0. The values of Gram-Class follow along the same lines: given *give* or *stop*, Gram-Class(verb) approaches 1, while for *thing* or *car*, it approaches 0.

	Word				Vowel		Gram-Class	
	Give	Stop	thing	car	front	Back	Verb	Noun
a.	0.3	0.2	0.2	0.3	0.5	0.5	0.5	0.5
b.	0.59	0	0.4	0.01	**1**	**0**	0.6	0.4
c.	0.01	0.4	0	0.59	**0**	**1**	0.4	0.6
d.	0.01	0	0.4	0.59	0.4	0.6	**0**	**1**
e.	0.59	0.4	0	0.01	0.6	0.4	**1**	**0**

Figure 5: A BN for phonosyntactics: a. prior probabilities; b. front vowel observed; c. back vowel observed; d. noun observed; and e. verb observed.

The network just described, representing only frequency information and correlations between domains of knowledge, demonstrates a graded bias for verbs when presented with front vowels and for nouns when presented with back vowels. This is demonstrated in Figure 5, where a. shows the prior probabilities of all values, b. an observed front vowel and c. an observed back vowel.[3] Relevant are the relative probabilities of *give* versus *thing* in Figure 5b. and *stop* versus *car* in Figure 5b.

in the little space available here. At scale, a single node representing the entire lexicon would become unwieldy as it interacts with other nodes, since the conditional distributions would need to take account of the entire lexicon. Thus, at scale, separate nodes for the lexicon and the four word classes would be needed.

[3] Such a network additionally allows us to make an empirically testable prediction; that there is also an advantage in speed of production for front verbs and back nouns, as shown in Figures 5d and e. To my knowledge, no study exists that could confirm or deny this prediction.

5 Properties of this Solution

It remains to be demonstrated, however, how the probabilistic asymmetries shown in the previous section can be related to processing speed differences.

The best way to think about this problem is in the neural grounding of a language representation system. Two competing models for the neural representation of mental constructs have very similar properties in terms of speed of recognition. Local representational schemes posit groups of neurons realizing mental representations, while distributed models posit different states of networks representing different mental representation (Feldman 1988). In the first model, identifying a mental construct involves the attainment of a relatively or absolutely high level of activation on the part of the appropriate group of neurons. In the second, a single network settles on a state representing that mental construct to the detriment of other states. In both, increased speed can arise from stronger default activations of certain nodes/states or from stronger or more numerous connections impinging on those nodes or leading to that state. A neural translation of the BN in Figure 4 would settle into a state of high activation more quickly the higher its probability, if probability is interpreted neurally as degree of activation.

It is relatively obvious how such a network would learn the asymmetric distributions we see in Figure 1. Since all that needs to be extracted is the probability of each class, a simple algorithm could increase the relative probability of a value each time it was observed. By the same token, in a neural implementation of such a network, Hebbian learning suffices for learning these probabilities (Wendelken and Shastri, in preparation).[4] Abducing the structure of BNs is a more complex exercise, but various methods have met with significant success, including entropy methods, score metrics, simulated annealing, and genetic algorithms (Jordan 1998).

6 Conclusion

The unmotivated nature of the distributional asymmetries described above means that they most likely exist solely due to historical accident, and as such are unexplainable from a synchronic perspective. But language users unconsciously incorporate this information, as the processing evidence demonstrates. The model presented above gives an account of the processing properties on the basis of a simple probabilistic model of the storage and

[4] Hebbian learning is the simplest and earliest-recognized type of neural learning. It involves the strengthening of connections that fire in association with other, stronger connections, and is responsible for types of associative learning.

relations between linguistic representations. As such it serves as an explanation for the processing data, which can not even be described by deterministic linguistic models.

From a broader perspective, to the extent that all sorts of linguistic knowledge are to be modeled, the particular argument presented above constitutes a piece of evidence for probabilistic and connectionist models of language. A possible objection to the contentions above might be that these observed regularities do not actually constitute facets of the linguistic system or grammar proper, but rather matters of language use. This argument becomes dangerously circular, however, as its definition of *language* or *grammar* as either entirely productive or entirely arbitrary depends on excluding partially productive features, like the ones discussed in this paper, from *language*. If, however, we define *language* to include all knowledge about the relation between sound sequences and the meanings they evoke, then we are unable to overlook these generalizations, since they are empirically shown to be part of the psychological reality of language for speaker-hearers. Other related studies of probabilistic properties of the relations between phonological and semantic (Bergen 2000a), phonological and speaker-specific (Bergen 2000b), and phonological and syntactic knowledge (de Jong 1989) demonstrate the degree to which linguistic knowledge defies the normally accepted determinism assumptions.

References

Bergen. Benjamin K. 2000a. Probabilistic associations between sound and meaning: Belief Networks for modeling phonaesthemes. Paper presented at the Fifth Conference on Conceptual Structure, Discourse, and Language, Santa Barbara.

Bergen, Benjamin K. 2000b. Probability in phonological generalizations: modeling French optional final consonants. *Proceedings of the 26th Annual Meeting of the Berkeley Linguistics Society*, ed. Alan Yu et al. Berkeley: BLS

Bybee, Joan and Carol Moder. 1983. Morphological Classes as Natural Categories. *Language* 59:251-270.

Bybee, Joan and Dan Slobin. 1982. Rules and Schemas in the development and use of the English past tense. *Language* 58:265-289.

Davis, S. and Kelly, Michael 1997. Knowledge of the English noun-verb stress difference by native and nonnative speakers. *Journal of Memory and Language* 36:445-460.

Fasold, Ralph. 1996. The quiet demise of variable rules. *Towards a critical sociolinguistics*, ed. Rajendra Singh, 79-98. Amsterdam: John Benjamins.

Feldman, Jerome. 1988. Connectionist Representation of Concepts. *Connectionist Models and Their Applications*, ed. D. Waltz and J. Feldman. Ablex Publishing Company.

Francis, W. Nelson and Henry Kucera. 1982. *Frequency analysis of English usage : lexicon and grammar*. Boston: Houghton Mifflin.

Jensen, F. V. 1996. *An introduction to Bayesian networks*. New York: Springer.

de Jong, Daan. 1989. A multivariate analysis of French Liaison. *New Methods in Dialectology*, ed. Schouten, M. E. H and Pieter van Reenen, 19-34. Dordrecht: Foris.

Jordan, Michael I. (ed.). 1998. *Learning in Graphical Models*. Kluwer Academic Press.

Kelly, Michael. 1992. Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review* 99(2):349-364.

Labov, William. 1972. *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Saussure, Ferdinand de. 1916. *Cours de linguistique generale*. Paris: Payot.

Sereno, J. A. 1994. Phonosyntactics. *Sound Symbolism*, ed. L. Hinton, J. Nichols, and J. Ohala. Cambridge: Cambridge University Press.

Wendelken, Carter and Lokendra Shastri. In Prep. Probabilistic Inference and Learning in a Connectionist Causal Network.

Department of Linguistics
1203 Dwinelle Hall
University of California at Berkeley
Berkeley, CA 94720
bbergen@socrates.berkeley.edu