Mathematical Linguistics

Geoffrey K. Pullum and András Kornai

Final version

MATHEMATICAL LINGUISTICS is the study of mathematical structures and methods that are of importance to linguistics. As in other branches of applied mathematics, the influence of the empirical subject matter is somewhat indirect: theorems are often proved more for their inherent mathematical value than for their applicability. Nevertheless, the internal organization of linguistics remains the best guide for understanding the internal subdivisions of mathematical linguistics, and we will survey the field following the traditional division of linguistics into → *Phonetics*, → *Phonology*, → *Morphology*, → *Syntax*, and → *Semantics*, looking at other branches of linguistics such as → *Sociolinguistics* or → *Language Acquisition* only to the extent that these have developed their own mathematical methods.

Phonetics The key structures of both mathematical and phonetic interest are → *Hidden Markov Models* (HMMs). Their importance stems from the way their structure is set up: discrete, psychologically relevant underlying units as hidden states coupled with continuous, physically relevant output. Though phoneticians routinely use the mathematical apparatus of ACOUSTICS ever since the pioneering work of Helmholtz (1859), neither DIFFERENTIAL EQUATIONS nor HARMONIC ANALYSIS are considered part of mathematical linguistics, because they enter the picture only indirectly, as part of the physics of the medium carrying the linguistic signal. HMMs, on the other hand, remain equally applicable if the modality is changed from spoken to written or signed language (see → *Speech Recognition*, → *Optical Character Recognition*, → *Sign Language*).

The HMM idea of discrete structural units (typically → *Phonemes* or → *Words*) coupled with continuous phonetic phenomena inspired the LAFS

(Lexical Access From Spectra) model (Klatt 1980), the first explicit → *Psycholinguistic* model incorporating the modern apparatus of SIGNAL PROCESSING.

Though their structure is well suited for continuous phenomena, NEURAL NET models (→ *Cognitive Science*) generally shy away from any attempt at detail phonetic modeling: the influential TRACE model (Rumelhart and McClelland 1986) is typical in this respect. The mathematical reason for this is to be found in the fundamental difference between the way temporal succession is handled in the two models. Without the additional expense of adding recurrence (Jordan 1986) neural nets can only deal with inputs and outputs of a fixed dimension, and once recurrence is added, neural net training becomes extremely complex. HMMs, on the other hand, assume a Markovian underlying structure, which is, for the most part, ideally suited for modeling the succession of linguistic units, having been developed by Markov (1913) for this very purpose.

Phonology and Morphology Starting with Bloomfield's (1926) postulates, the basic conceptual apparatus of mathematical linguistics — in particular, the idea of hierarchical structures composed of relatively stable recurrent items — was developed primarily on the basis of phonological and morphological phenomena. Chomsky (1956, 1959) formulated three theoretical models for the description of linguistic structure, one based on → *Finite-State Automata* (FSA), one based on → *Context-Free Grammars* (CFGs), and one on context-sensitive grammars (CSGs) and/or the even more powerful Unrestricted Rewriting Systems (URSs). The relation between these is investigated under the heading → *Generative Capacity*, and was the basis of much further work on formal language theory within computer science.

Regarding mathematical work on phonology, there were some logicians and linguists (including Tadeusz Batóg, F. H. H. Kortlandt, Jan Mulder, and Anders Wedberg) who worked on phonemic theory from a set-theoretic standpoint in the 1960s and 1970s, but such work had little impact on linguistic practice. The definitive formalization of theoretical phonology and morphology was that proposed by Chomsky and Halle (1968), using URSs. By that time, it was well known that ordered sets of CSG or URS rules provide a good mathematical reconstruction of Panini's (morpho)phonological rules (Cardona 1965), and are superior to the neogrammarian SOUND LAWS both in descriptive detail and in predictive power (Kiparsky 1965). It therefore came as something of a surprise that a system of considerably weaker generative capacity, finite state transducers (FSTs), can apparently describe the same phenomena (Johnson 1970). This observation was not fully assimilated in models of $\rightarrow$ *Computational Morphology* until well over a decade later with the introduction of two-level phonology and morphology (Koskenniemi 1983, Kaplan and Kay 1994).

Until the mid-1970s, the internal structure of phonological representations, based on $\rightarrow$ *Distinctive Features*, could be formalized by embedding it in an n-dimensional cube (Cherry, Halle, Jakobson 1953; Cherry 1956). With the advent of $\rightarrow$ *Autosegmental* and $\rightarrow$ *Metrical* phonology, a considerably more involved formalism became necessary (Kornai 1991). While the representations retained this additional complexity, in the 1990s the whole notion of rules operating on such representations in sequence was abandoned in favor of $\rightarrow$ *Optimality Theory* which describes the relationship between underlying and surface units in terms of rank-ordered constraint systems. A number of mathematical linguists (including Jason Eisner, Robert Frank, Markus Hiller, Lauri Karttunen, Giorgio Satta) have shown that this mode of description need not imply an increase in generative capacity, inasmuch as FSTs, under various sets of assumptions, have sufficient power to model the interaction of systems of ranked constraints.

Markov's pioneering work is a contribution both to phonology and to $\rightarrow$ *Statistical Linguistics*, given the near-phonemic nature of Russian orthography. Historically, the development of Markov models took place largely in isolation from mainstream phonology and morphology, largely because these offer a rich storehouse of LONG DISTANCE and NON-CONCATENATIVE phenomena, which in a segmental framework appear as violations of the Markovian assumption. The autosegmental framework, by resolving these violations, helped to usher in a more mature understanding of the key technical issues, and it is fair to say that today the mathematical apparatus of phonology and morphology is centered on the study of deterministic, nondeterministic, and probabilistic FSTs.

Syntax Chomsky's first significant technical contribution to linguistics was his formalization of IMMEDIATE CONSTITUENT ANALYSIS (Wells 1947, Harris 1951) by means of $\rightarrow$ *Context Free Grammars* (Chomsky 1956, 1959). Though in the definition of CFGs he sacrificed some of the detail of the earlier work (in particular, his system did not provide for DISCONTINUOUS CON-

STITUENTS or for BAR LEVEL SUPERSCRIPTS), from a mathematical perspective CFGs hit on a particularly sweet spot: just as FSA correspond to the rationals, CFGs correspond to algebraic numbers (see Eilenberg 1974).

CFGs found an immediate application in the design of PROGRAMMING LANGUAGES, where they retain a central position to this day, in spite of the fact that it can be shown that a number of widely used programming languages go beyond the context-free in some respects. For example, if it is considered a syntactic requirement that each variable used be declared at the start of the program, that aspect of the syntax is likely not to be CFG-describable; see Harrison 1978, 219–221). In fact, much of the early work in mathematical linguistics concerned with efficient methods of parsing eventually found a better home in COMPILER DESIGN → *Parsing*.

The key idea of CFGs was to replace the symmetrical (equational) notation used in earlier formulations by the asymmetrical notion of string rewriting that had, up to that point, been applied only by logicians, and only in settings of considerably broader generality, recursively enumerable or recursive (Thue 1914, Post 1936, 1943; for a modern discussion see Salomaa 1973). Though distributional equivalence, which was the basis for the equational notation, remained an important technical tool (Nerode 1958), the focus shifted from mutual to unidirectional substitutability, which helped to clarify the effect of CONTEXT. In a Markovian world, only context linearly to the left matters, and even that, only within a limited window: it is this limitation which makes it possible to state Markov's original ideas in the contemporary framework of (probabilistic) FSA. In a CFG, only hierarchical context (parent node in a tree) matters: string context, including immediate neighbors, is immaterial. The resulting theory can be expressed in terms of finite TREE AUTOMATA (Thatcher 1971). In a CSG, both hierarchical and linear context plays a role, and the resulting theory turns out to be equivalent to Turing machines with workspace linear in the size of the input. → *Automata Theory*.

In syntax, the appropriate choice of → *Grammar Formalism* is more of a contentious issue than in phonology/morphology. Unsettled issues include whether COMPETENCE has any probabilistic aspects, and what → *Generative Capacity* is necessary and sufficient for the range of actual and potential natural languages. Though probabilistic CFGs are widely used in → *Computational Linguistics,* the theoretical necessity of a probabilistic component has been broadly accepted only in → *Sociolinguistics*, but even there, the dominant statistical model (LOGISTIC REGRESSION, see Sankoff 1987) has its detractors (Kay and McDaniel 1979). The issue of generative capacity played a key role in GENERALIZED PHRASE STRUCTURE GRAMMAR (GPSG, see Gazdar et al. 1985), which developed the idea that CFGs are sufficient (Pullum and Gazdar 1982). By providing a critical assessment of the earlier literature GPSG paved the way for subsequent work (Huybregts 1985, Shieber 1985, Culy 1985) that resulted in the current near consensus that some power beyond that of CFGs is required. This development led to renewed interest in → *Mildly Context Sensitive* languages which are equivalently definable (Vijay-Shanker and Weir 1994) by at least four distinct grammar formalisms: TREE-ADJOINING GRAMMARS, HEAD GRAMMARS, COMBINATORY CATEGORIAL GRAMMARS, and LINEAR INDEXED GRAMMARS.

Semantics Early efforts to address linguistic semantics within generative grammar assumed that the meaning of expressions was to be given by providing them with translations into expressions in a representational system of some sort. Philosophers have objected that no such representation in any vocabulary can amount to a specification of meaning. Since the work of Montague (1974) became known to linguists, attention has shifted to providing natural language expressions with actual model-theoretic interpretations, exactly as is done with formal languages in logic. Much of Partee et al. (1990) is geared toward providing enough mathematical background to understand developments in model-theoretic semantics. For reasons that have much to do with the still controversial AUTONOMY OF SYNTAX thesis, formal semantics is very often done in conjunction with nontransformational theories of syntax such as → *Categorial Grammar*. For recent developments, see Jacobson (1999, 2000).

On the whole, approaches to semantics based on → *Information Theory* are still largely restricted to LEXICAL SEMANTICS, though many tasks such as → *Machine Translation* that were originally believed to require sophisticated semantic analysis are now often performed by purely statistical models.

Model-theoretic syntax One recent line of research connects model theory to syntax by means of a logical theory that has well-formed structures in the language as its models. Rogers (1998) devises a monadic second-order logic that characterizes the sets of trees generable by a CFG. The statements that context-free grammars make about sets of trees are made directly, without phrase structure rules. Blackburn and Meyer-Viol (1997) explore similar ideas using MODAL LOGIC on trees, and Rogers (2001) extends these ideas to structures more complex than trees.

Both in phonology/morphology and in syntax/semantics the choice of linguistic formalism is to some extent influenced by considerations that go beyond the primary issue of DESCRIPTIVE ADEQUACY. One important issue is → *Recognition Complexity.* This concerns the complexity of the decision problem for membership in a language: it is assumed that a grammatical theory should have the property of guaranteeing that there is some reasonably rapid (polynomial in the lenght of the input) computation that will answer the question of whether a given sequence of words is a grammatical expression according to a given grammar. Human beings certainly do much more than this when they listen to an utterance and figure out the meaning of what was said, so a grammatical theory that cannot even guarantee reasonably rapid confirmation of well-formedness is probably not psycholinguistically realistic. Another one is → *Learnability*, which concerns what sorts of mathematically definable procedures could in principle correctly guess the grammars for languages.

References Blackburn, Patrick and Wilfried Meyer-Viol. 1997. Modal logic and model-theoretic syntax. In M. de Rijke (Ed.), *Advances in Intensional Logic*. Dordrecht: Kluwer Academic.

Bloomfield, Leonard 1926. A set of postulates for the science of language *Language* **2** 153-164. Reprinted in Martin Joos (ed) *Readings in linguistics* ACLS, 1958

Cardona, G. 1965. On translating and formalizing Paninian rules. *Journal of the Oriental Institute* (Baroda) **14** 306-314

Cherry, Colin. 1956 Roman Jakobson's distinctive features as the normal coordinates of a language. In Morris Halle (ed.) *For Roman Jakobson* The Hague: Mouton.

Cherry, Colin, Morris Halle, and Roman Jakobson 1953. Toward the logical description of languages in their phonemic aspect *Language* **29** 34-46

Chomsky, Noam. 1956. Three models for the description of language. *IRE Transactions on Information Theory* **IT-2**

Chomsky, Noam. 1959. On certain formal properties of grammars. *Information and Control* 2, 137-167

Chomsky, Noam and Morris Halle 1968. *The Sound Pattern of English* New York: Harper & Row

Culy, Christopher 1985. The complexity of the vocabulary of Bambara. *Linguistics and Philosophy* **8** 345-351

Eilenberg, Samuel 1974. *Automata, languages, and machines* New York: Academic Press

Gazdar, Gerald, Ewan Klein, Geoffrey Pullum, and Ivan Sag 1985. *Generalized Phrase structure grammar* Blackwell, Cambridge MA

Harris, Zellig 1951. *Methods in structural linguistics* Chicago: University of Chicago Press

Harrison, Michael A. 1978. *Introduction to Formal Language Theory* Reading MA: Addison-Wesley

von Helmholtz, Hermann 1859. Transl. as *On the sensations of tone as a physiological basis for the theory of music* 1912, Longmans, Green, and Co., London.

Huybregts, Riny (1985). The weak inadequacy of CFPSGs. In: G. de Haan, M. Trommelen and W. Zonneveld (eds.), *Van Periferie naar Kern* Dordrecht: Foris Publications 81–99

Jacobson, Pauline 1999. Towards a variable-free semantics *Linguistics and Philosophy* **22** 117-184

Jacobson, Pauline 2000. Paycheck pronouns, Bach-Peters sentences, and variable free semantics *Natural Language Semantics* **8** 77-155

Johnson, C. Douglas 1970. *Formal aspects of phonological representation* PhD Thesis UC Berkeley. Published by Mouton, The Hague 1974

Jordan, Michael 1986. Serial order: A parallel distributed processing approach. Institute for Cognitive Science, University of California at San Diego, La Jolla, Technical Report 8604.

Kaplan, Ronald M. and Martin Kay. Regular models of phonological rule systems *Computational Linguistics* **20** 331–378.

Kay, Paul and Chad K. McDaniel. 1979. On the logic of variable rules. *Language in Society* **8** 151-187

Kiparsky, Paul. 1965. *Phonological Change* PhD dissertation. Massachusetts Institute of Technology

Klatt, D. H. (1980) SCRIBER and LAFS: Two approaches to speech analysis. In W. A. Lea (Ed.) *Trends in speech recognition* Englewood Cliffs NJ: Prentice Hall

Kornai, András 1991. *Formal Phonology* PhD dissertation, Stanford, published by Garland 1994

Koskenniemi, Kimmo 1983 *Two-level Morphology: a general computational model for word-form recognition and production* Helsinki Department of General Linguistics, University of Helsinki Publication **11**

Markoff, A.A. 1913. Essai d'une recherche statistique sur le texte du roman 'Eugene Onegin.' *Bull. Acad. Imper. Sci. St. Petersbourg* **7**.

Montague, Richard. 1974. *Formal philosophy.* Edited by Richmond Thomason. New Haven: Yale University Press.

Nerode, Anil 1958. Linear automaton transformations. *Proc. Amer. Math. Soc.* **9** 541-544.

Partee, Barbara Hall, Alice ter Meulen, and Robert E. Wall. 1990. *Mathematical Methods in Linguistics.* Dordrecht: Kluwer Academic.

Post, Emil 1936. Finite combinatory processes. Formulation 1. *Journal of Symbolic Logic* **1** 103-105.

Post, Emil 1943 Formal reduction of the general combinatorial decision problem. *American Journal of Mathematics* **6** 197–215

Pullum, Geoffrey K. and Gerald Gazdar 1982. Natural languages and context free languages *Linguistics and Philosophy* **4** 471-504

Rogers, James. 1998. *A Descriptive Approach to Language-Theoretic Complexity.* Stanford, CA: CSLI Publications.

Rogers, James. 2001. wMSO Theories as Grammar Formalisms *Theoretical Computer Science*, To appear.

Rumelhart, D. and J. McClelland 1986. On learning the past tenses of English verbs In D. Rumelhart and J. McClelland (eds.) *Parallel distributed processing: Explorations in the microstructure of cognition* Cambridge MA: MIT Press

Salomaa, Arto 1973. *Formal Languages* New York: Academic Press

Sankoff, David 1987. Variable Rules. In Ammon, Dittmar & Mattheier (eds) *Sociolinguistics: An international handbook of the science of language and society*, **I** 984-997

Shieber, Stuart 1985. Evidence against the context-freeness of natural language. *Linguistics and Philosophy* **8** 333-343

Thatcher, James W. 1971. *Tree automata: an informal survey* In: Alfred Aho (ed): *Current trends in the theory of computing* 143-172

Thue, Axel (1914) Probleme uber Veranderungen von Zeichenreihen nach gegeben Regeln *Skr. Vid. Kristianaia I. Mat. Naturv. Klasse* 10/34

Vijay-Shanker, K. and David J. Weir. 1994. The equivalence of four extensions of context-free grammars. *Mathematical Systems Theory* **27** 511–546

Wells, Roulon S. 1947. Immediate constituents *Language* **23** 321-343. Reprinted in Martin Joos (ed) *Readings in linguistics* ACLS, 1958