Andras Kornai

Mathematical Linguistics

Springer

András Kornai
MetaCarta Inc.
350 Massachusetts Ave.
Cambridge, MA 02139
USA

ISBN: 978-1-84628-985-9 e-ISBN: 978-1-84628-986-6
DOI: 10.1007/978-1-84628-986-6

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2007940401

Printed on acid-free paper

9 8 7 6 5 4 3 2 1

Springer Science+Business Media
Springer.com

Contents

1

Introduction

1.1 The subject matter

What is *mathematical linguistics*? A classic book on the subject, (Jakobson 1961), contains papers on a variety of subjects, including a categorial grammar (Lambek 1961), formal syntax (Chomsky 1961, Hiż 1961), logical semantics (Quine 1961, Curry 1961), phonetics and phonology (Peterson and Harary 1961, Halle 1961), Markov models (Mandelbrot 1961b), handwriting (Chao 1961, Eden 1961), parsing (Oettinger 1961, Yngve 1961), glottochronology (Gleason 1961), and the philosophy of language (Putnam 1961), as well as a number of papers that are harder to fit into our current system of scientific subfields, perhaps because there is a void now where once there was cybernetics and systems theory (see Heims 1991).

A good way to understand how these seemingly so disparate fields cohere is to proceed by analogy to mathematical physics. Hamiltonians receive a great deal more mathematical attention than, say, the study of generalized incomplete Gamma functions, because of their relevance to mechanics, not because the subject is, from a purely mathematical perspective, necessarily more interesting. Many parts of mathematical physics find a natural home in the study of differential equations, but other parts fit much better in algebra, statistics, and elsewhere. As we shall see, the situation in mathematical linguistics is quite similar: many parts of the subject would fit nicely in algebra and logic, but there are many others for which methods belonging to other fields of mathematics are more appropriate. Ultimately the coherence of the field, such as it is, depends on the coherence of linguistics.

Because of the enormous impact that the works of Noam Chomsky and Richard Montague had on the postwar development of the discipline, there is a strong tendency, observable both in introductory texts such as Partee et al. (1990) and in research monographs such as Kracht (2003), to simply equate mathematical linguistics with formal syntax and semantics. Here we take a broader view, assigning syntax (Chapter 5) and semantics (Chapter 6) no greater scope than they would receive in any book that covers linguistics as a whole, and devoting a considerable amount of space to phonology (Chapter 2), morphology (Chapter 3), phonetics (Chapters 8 and 9), and other areas of traditional linguistics. In particular, we make sure that the reader will learn (in Chapter 7) the central mathematical ideas of information theory and algorithmic complexity that provide the foundations of much of the contemporary work in mathematical linguistics.

This does not mean, of course, that mathematical linguistics is a discipline entirely without boundaries. Since almost all social activity ultimately rests on linguistic communication, there is a great deal of temptation to reduce problems from other fields of inquiry to purely linguistic problems. Instead of understanding schizoid behavior, perhaps we should first ponder what the phrase *multiple personality* means. Mathematics already provides a reasonable notion of 'multiple', but what is 'personality', and how can there be more than one per person? Can a proper understanding of the suffixes *-al* and *-ity* be the key? This line of inquiry, predating the Schoolmen and going back at least to the *cheng ming* (rectification of names) doctrine of Confucius, has a clear and convincing rationale (*The Analects* 13.3, D.C. Lau transl.):

> When names are not correct, what is said will not sound reasonable; when what is said does not sound reasonable, affairs will not culminate in success; when affairs do not culminate in success, rites and music will not flourish; when rites and music do not flourish, punishments will not fit the crimes; when punishments do not fit the crimes, the common people will not know where to put hand and foot. Thus when the gentleman names something, the name is sure to be usable in speech, and when he says something this is sure to be practicable. The thing about the gentleman is that he is anything but casual where speech is concerned.

In reality, linguistics lacks the resolving power to serve as the ultimate arbiter of truth in the social sciences, just as physics lacks the resolving power to explain the accidents of biological evolution that made us human. By applying mathematical techniques we can at least gain some understanding of the limitations of the enterprise, and this is what this book sets out to do.

1.2 Cumulative knowledge

It is hard to find any aspect of linguistics that is entirely uncontroversial, and to the mathematician less steeped in the broad tradition of the humanities it may appear that linguistic controversies are often settled on purely rhetorical grounds. Thus it may seem advisable, and only fair, to give both sides the full opportunity to express their views and let the reader be the judge. But such a book would run to thousands of pages and would be of far more interest to historians of science than to those actually intending to learn mathematical linguistics. Therefore we will not necessarily accord equal space to both sides of such controversies; indeed often we will present a single view and will proceed without even attempting to discuss alternative ways of looking at the matter.

Since part of our goal is to orient the reader not familiar with linguistics, typically we will present the majority view in detail and describe the minority view only tersely. For example, Chapter 4 introduces the reader to morphology and will rely heavily on the notion of the morpheme – the excellent book by Anderson (1992) denying the utility, if not the very existence, of morphemes, will be relegated to footnotes. In some cases, when we feel that the minority view is the correct one, the emphasis will be inverted: for example, Chapter 6, dealing with semantics, is more informed by the 'surface compositional' than the 'logical form' view. In other cases, particularly in Chapter 5, dealing with syntax, we felt that such a bewildering variety of frameworks is available that the reader is better served by an impartial analysis that tries to bring out the common core than by in-depth formalization of any particular strand of research.

In general, our goal is to present linguistics as a cumulative body of knowledge. In order to find a consistent set of definitions that offer a rational reconstruction of the main ideas and techniques developed over the course of millennia, it will often be necessary to take sides in various controversies. There is no pretense here that mathematical formulation will necessarily endow a particular set of ideas with greater verity, and often the opposing view could be formalized just as well. This is particularly evident in those cases where theories diametrically opposed in their means actually share a common goal such as describing all and only the well-formed structures (e.g. syllables, words, or sentences) of languages. As a result, we will see discussions of many 'minority' theories, such as case grammar or generative semantics, which are generally believed to have less formal content than their 'majority' counterparts.

1.3 Definitions

For the mathematician, definitions are nearly synonymous with abbreviations: we say 'triangle' instead of describing the peculiar arrangement of points and lines that define it, 'polynomial' instead of going into a long discussion about terms, addition, monomials, multiplication, or the underlying ring of coefficients, and so forth. The only sanity check required is to exhibit an instance, typically an explicit set-theoretic construction, to demonstrate that the defined object indeed exists. Quite often, counterfactual objects such as the smallest group K not meeting some description, or objects whose existence is not known, such as the smallest nontrivial root of ζ not on the critical line, will play an important role in (indirect) proofs, and occasionally we find cases, such as *motivic cohomology*, where the whole conceptual apparatus is in doubt. In linguistics, there is rarely any serious doubt about the existence of the objects of inquiry. When we strive to define 'word', we give a mathematical formulation not so much to demonstrate that words exist, for we know perfectly well that we use words both in spoken and written language, but rather to handle the odd and unexpected cases. The reader is invited to construct a definition now and to write it down for comparison with the eventual definition that will emerge only after a rather complex discussion in Chapter 4.

In this respect, mathematical linguistics is very much like the empirical sciences, where formulating a definition involves at least three distinct steps: an *ostensive*

1.5 Foundations

For the purposes of mathematical linguistics, the classical foundations of mathematics are quite satisfactory: all objects of interest are sets, typically finite or, rarely, denumerably infinite. This is not to say that nonclassical metamathematical tools such as Heyting algebras find no use in mathematical linguistics but simply to assert that the fundamental issues of this field are not foundational but definitional.

Given the finitistic nature of the subject matter, we will in general use the terms set, class, and collection interchangeably, drawing explicit cardinality distinctions only in the rare cases where we step out of the finite domain. Much of the classical linguistic literature of course predates Cantor, and even the modern literature typically conceives of infinity in the Gaussian manner of a potential, as opposed to actual, Cantorian infinity. Because of immediate empirical concerns, denumerable generalizations of finite objects such as ω-words and Büchi automata are rarely used,[1] and in fact even the trivial step of generalizing from a fixed constant to arbitrary n is often viewed with great suspicion.

Aside from the tradition of Indian logic, the study of languages had very little impact on the foundations of mathematics. Rather, mathematicians realized early on that natural language is a complex and in many ways unreliable construct and created their own simplified language of formulas and the mathematical techniques to investigate it. As we shall see, some of these techniques are general enough to cover essential facets of natural languages, while others scale much more poorly.

There is an interesting residue of foundational work in the Berry, Richard, Liar, and other paradoxes, which are often viewed as diagnostic of the vagueness, ambiguity, or even ‘paradoxical nature’ of natural language. Since the goal is to develop a mathematical theory of language, sooner or later we must define English in a formal system. Once this is done, the buck stops there, and questions like “what is the smallest integer not nameable in ten words?” need to be addressed anew.

We shall begin with the seemingly simpler issue of the first number not nameable in *one* word. Since it appears to be one hundred and one, a number already requiring *four* words to name, we should systematically investigate the number of words in number names. There are two main issues to consider: what is a word? (see Chapter 4); and what is a name? (see Chapter 6). Another formulation of the Berry paradox invokes the notion of syllables; these are also discussed in Chapter 4. Eventually we will deal with the paradoxes in Chapter 6, but our treatment concentrates on the linguistic, rather than the foundational, issues.

1.6 Mesoscopy

Physicists speak of mesoscopic systems when these contain, say, fifty atoms, too large to be given a microscopic quantum-mechanical description but too small for the classical macroscopic properties to dominate the behavior of the system. Linguistic

[1] For a contrary view, see Langendoen and Postal (1984).

systems are mesoscopic in the same broad sense: they have thousands of rules and axioms compared with the handful of axioms used in most branches of mathematics. Group theory explores the implications of five axioms, arithmetic and set theory get along with five and twelve axioms respectively (not counting members of axiom schemes separately), and the most complex axiom system in common use, that of geometry, has less than thirty axioms.

It comes as no surprise that with such a large number of axioms, linguistic systems are never pursued microscopically to yield implications in the same depth as group theory or even less well-developed branches of mathematics. What is perhaps more surprising is that we can get reasonable approximations of the behavior at the macroscopic level using the statistical techniques pioneered by A. A. Markov (see Chapters 7 and 8).

Statistical mechanics owes its success largely to the fact that in thermodynamics only a handful of phenomenological parameters are of interest, and these are relatively easy to link to averages of mechanical quantities. In mathematical linguistics the averages that matter (e.g. the percentage of words correctly recognized or correctly translated) are linked only very indirectly to the measurable parameters, of which there is such a bewildering variety that it requires special techniques to decide which ones to employ and which ones to leave unmodeled.

Macroscopic techniques, by their very nature, can yield only approximations for mesoscopic systems. Microscopic techniques, though in principle easy to extend to the mesoscopic domain, are in practice also prone to all kinds of bugs, ranging from plain errors of fact (which are hard to avoid once we deal with thousands of axioms) to more subtle, and often systematic, errors and omissions. Readers may at this point feel very uncomfortable with the idea that a given system is only 70%, 95%, or even 99.99% correct. After all, isn't a single contradiction or empirically false prediction enough to render a theory invalid? Since we need a whole book to develop the tools needed to address this question, the full answer will have to wait until Chapter 10.

What is clear from the outset is that natural languages offer an unparalleled variety of complex algebraic structures. The closest examples we can think of are in crystallographic topology, but the internal complexity of the groups studied there is a product of pure mathematics, while the internal complexity of the syntactic semigroups associated to natural languages is more attractive to the applied mathematician, as it is something found in vivo. Perhaps the most captivating aspect of mathematical linguistics is not just the existence of discrete mesoscopic structures but the fact that these come embedded, in ways we do not fully understand, in continuous signals (see Chapter 9).

1.7 Further reading

The first works that can, from a modern standpoint, be called mathematical linguistics are Markov's (1912) extension of the weak law of large numbers (see Theorem 8.2.2) and Thue's (1914) introduction of string manipulation (see Chapter 2), but pride of place must go to Pāṇini, whose inventions include not just grammatical rules but also a formal metalanguage to describe the rules and a set of principles governing their interaction. Although the *Ashṭādhyāyī* is available on the web in its entirety, the reader will be at a loss without the modern commentary literature starting with Böhtlingk (1887, reprinted 1964). For modern accounts of various aspects of the system see Staal (1962, 1967) Cardona (1965, 1969, 1970, 1976, 1988), and Kiparsky (1979, 1982a, 2002). Needless to say, Pāṇini did not work in isolation. Much like Euclid, he built on the inventions of his predecessors, but his work was so comprehensive that it effectively drove the earlier material out of circulation. While much of linguistics has aspired to formal rigor throughout the ages (for the Masoretic tradition, see Aronoff 1985, for medieval syntax see Covington 1984), the continuous line of development that culminates in contemporary formal grammar begins with Bloomfield's (1926) Postulates (see Section 3.1), with the most important milestones being Harris (1951) and Chomsky (1956, 1959).

Another important line of research, only briefly alluded to above, could be called mathematical antilinguistics, its goal being the elimination, rather than the explanation, of the peculiarities of natural language from the system. The early history of the subject is discussed in depth in Eco (1995); the modern mathematical developments begin with Frege's (1879) system of *Concept Writing* (Begriffsschrift), generally considered the founding paper of mathematical logic. There is no doubt that many great mathematicians from Leibniz to Russell were extremely critical of natural language, using it more for counterexamples and cautionary tales than as a part of objective reality worthy of formal study, but this critical attitude has all but disappeared with the work of Montague (1970a, 1970b, 1973). Contemporary developments in model-theoretic semantics or 'Montague grammar' are discussed in Chapter 6.

Major summaries of the state of the art in mathematical linguistics include Jakobson (1961), Levelt (1974), Manaster-Ramer (1987), and the subsequent Mathematics of Language (MOL) conference volumes. We will have many occasions to cite Kracht's (2003) indispensable monograph *The Mathematics of Language*.

The volumes above are generally more suitable for the researcher or advanced graduate student than for those approaching the subject as undergraduates. To some extent, the mathematical prerequisites can be learned from the ground up from classic introductory textbooks such as Gross (1972) or Salomaa (1973). Gruska (1997) offers a more modern and, from the theoretical computer science perspective, far more comprehensive introduction. The best elementary introduction to the logical prerequisites is Gamut (1991). The discrete side of the standard "mathematics for linguists" curriculum is conveniently summarized by Partee et al. (1990), and the statistical approach is clearly introduced by Manning and Schütze (1999). The standard introduction to pattern recognition is Duda et al. (2000). Variable rules were introduced in Cedergren and Sankoff (1974) and soon became the standard modeling method in sociolinguistics – we shall discuss them in Chapter 5.

2

The elements

A primary concern of mathematical linguistics is to effectively enumerate those sets of words, sentences, etc., that play some important linguistic role. Typically, this is done by means of *generating* the set in question, a definitional method that we introduce in Section 2.1 by means of examples and counterexamples that show the similarities and the differences between the standard mathematical use of the term 'generate' and the way it is employed in linguistics.

Because the techniques used in defining sets, functions, relations, etc., are not always directly useful for evaluating them at a given point, an equally important concern is to solve the membership problem for the sets, functions, relations, and other structures of interest. In Section 2.2 we therefore introduce a variety of *grammars* that can be used to, among other things, create *certificates* that a particular element is indeed a member of the set, gets mapped to a particular value, stands in a prescribed relation to other elements and so on, and compare generative systems to logical calculi.

Since *generative grammar* is most familiar to mathematicians and computer scientists as a set of rather loosely collected string-rewriting techniques, in Section 2.3 we give a brief overview of this domain. We put the emphasis on context-sensitive grammars both because they play an important role in phonology (see Chapter 3) and morphology (see Chapter 4) and because they provide an essential line of defense against undecidability in syntax (see Chapter 5).

2.1 Generation

To define a collection of objects, it is often expedient to begin with a fixed set of primitive elements E and a fixed collection of *rules* (we use this term in a broad sense that does not imply strict procedurality) R that describe permissible arrangements of the primitive elements as well as of more complex objects. If x, y, z are objects *satisfying* a (binary) rule $z = r(x, y)$, we say that z **directly generates** x and y (in this order) and use the notation $z \rightarrow_r xy$. The smallest collection of objects closed under direct generation by any $r \in R$ and containing all elements of E is called the set **generated** from E by R.

Very often the simplest or most natural definition yields a superset of the real objects of interest, which is therefore supplemented by some additional conditions to narrow it down. In textbooks on algebra, the symmetric group is invariably introduced before the alternating group, and the latter is presented simply as a subgroup of the former. In logic, closed formulas are typically introduced as a special class of well-formed formulas. In context-free grammars, the sentential forms produced by the grammar are kept only if they contain no nonterminals (see Section 2.3), and we will see many similar examples (e.g. in the handling of agreement; see Section 5.2.3).

Generative definitions need to be supported by some notion of *equality* among the defined objects. Typically, the notion of equality we wish to employ will abstract away from the derivational history of the object, but in some cases we will need a stronger definition of identity that defines two objects to be the same only if they were generated the same way. Of particular interest in this regard are derivational *strata.* A specific intermediary stage of a derivation (e.g. when a group or rules have been exhausted or when some well-formedness criterion is met) is often called a **stratum** and is endowed with theoretically significant properties, such as availability for interfacing with other modules of grammar. Theories that recognize strata are called *multistratal*, and those that do not are called *monostratal* – we shall see examples of both in Chapter 5.

In mathematical linguistics, the objects of interest are the collection of words in a language, the collection of sentences, the collection of meanings, etc. Even the most tame and obviously finite collections of this kind present great definitional difficulties. Consider, for example, the set of characters (graphemes) used in written English. Are uppercase and lowercase forms to be kept distinct? How about punctuation, digits, or Zapf dingbats? If there is a new character for the euro currency unit, as there is a special character for dollar and pound sterling, shall it be included on account of Ireland having already joined the euro zone or shall we wait until England follows suit? Before proceeding to words, meanings, and other more subtle objects of inquiry, we will therefore first refine the notion of a generative definition on some familiar mathematical objects.

Example 2.1.1 Wang tilings. Let C be a finite set of colors and S be a finite set of square tiles, each colored on the edges according to some function $e : S \rightarrow C^4$. We assume that for each coloring *type* we have an infinite supply of *tokens* colored with that pattern: these make up the set of primitive elements E. The goal is to tile the whole plane (or just the first quadrant) laying down the tiles so that their colors match at the edges. To express this restriction more precisely, we use a rule system R with four rules n, s, e, w as follows. Let $\mathbb{Z}$ be the set of integers, $'$ be the successor function "add one" and ' be its inverse "subtract one". For any $i, j \in \mathbb{Z}$, we say that the tile u whose bottom left corner is at (i, j) has a correct neighbor to the north if the third component of $e(u)$ is the same as the first component of $e(v)$ where v is the tile at (i, j'). Denoting the ith projection by π_i, we can write $\pi_3(e(u)) = \pi_1(e(v))$ for v at (i, j'). Similarly, the west rule requires $\pi_4(e(u)) = \pi_2(e(v))$ for v at (i', j), the east rule requires $\pi_2(e(u)) = \pi_4(e(v))$ for v at (i', j), and the south rule requires

As a matter of fact, it is often tempting to replace natural languages, the true object of inquiry, by some well-regimented semiformal or fully formal construct used in mathematics. Certainly, there is nothing wrong with a bit of idealization, especially with ignoring factors best classified as noise. But a discussion about the English word *triangle* cannot rely too much on the geometrical object by this name since this would create problems where there aren't any; for example, it is evident that a hunter *circling* around a clearing does not require that her path keep the exact same distance from the center at all times. To say that this amounts to fuzzy definitions or sloppy language use is to put the cart before the horse: the fact to be explained is not how a cleaned-up language *could be* used for communication but how real language *is* used.

Exercise 2.1 The Fibonacci numbers are defined by $f_0 = 0, f_1 = 1, f_{n+1} = f_n + f_{n-1}$. Is this a generative definition? Why?

2.2 Axioms, rules, and constraints

There is an unbroken tradition of argumentation running from the Greek sophists to the Oxford Union, and the axiomatic method has its historic roots in the efforts to regulate the permissible methods of debate. As in many other fields of human activity, ranging from ritual to game playing, regulation will lay bare some essential features of the activity and thereby make it more enjoyable for those who choose to participate. Since it is the general experience that almost all statements are debatable, to manage argumentation one first needs to postulate a small set of primitive statements on which the parties agree – those who will not agree are simply excluded from the debate. As there is remarkable agreement about the validity of certain kinds of inference, the stage is set for a fully formal, even automatic, method of verifying whether a given argument indeed leads to the desired conclusion from the agreed upon premises.

There is an equally venerable tradition of protecting the full meaning and exact form of sacred texts, both to make sure that mispronunciations and other errors that may creep in over the centuries do not render them ineffectual and that misinterpretations do not confuse those whose task is to utter them on the right occasion. Even if we ignore the phonetic issues related to 'proper' pronunciation (see Chapter 8), writing down the texts is far from sufficient for the broader goals of preservation. With any material of great antiquity, we rarely have a single fully preserved and widely accepted version – rather, we have several imperfect variants and fragments. What is needed is not just a frozen description of some texts, say the Vedas, but also a grammar that defines what constitutes a proper Vedic text. The philological ability to determine the age of a section and undo subsequent modifications is especially important because the words of earlier sages are typically accorded greater weight.

In defining the language of a text, a period, or a speech community, we can propagate *grammaticality* the same way we propagate truth in an axiomatic system, by choosing an initial set of grammatical expressions and defining some permissible combinatorical operations that are guaranteed to preserve grammaticality. Quite often, such operations are conceptualized as being composed of a purely combinatorical step (typically concatenation) followed by some tidying up; e.g., adding a third-person suffix to the verb when it follows a third-person subject: compare *I see* to *He sees*. In logic, we mark the operators overtly by affixing them to the sequence of the operands – prefix (Polish), interfix (standard), and postfix (reverse Polish) notations are all in wide use – and tend not to put a great deal of emphasis on tidying up (omission of parentheses is typical). In linguistics, there is generally only one operation considered, concatenation, so no overt marking is necessary, but the tidying up is viewed as central to the enterprise of obtaining all and only the attested forms.

The same goal of characterizing all and only the grammatical forms can be accomplished by more indirect means. Rather than starting from a set of fully grammatical forms, we can begin with some more abstract inventory, such as the set of words W, elements of which need not in and of themselves be grammatical, and rather than propagating grammaticality from the parts to the whole, we perform some computation along the way to keep score.

Example 2.2.1 Balanced parentheses. We have two atomic expressions, the left and the right paren, and we assign the values $+1$ to '(' and -1 to ')'. We can successively add new paren symbols on the right as long as the score (overall sum of $+1$ and -1 values) does not dip below zero: the well-formed (balanced) expressions are simply those where this WFC is met and the overall score is zero.

Discussion The example is atypical for two reasons: first because linguistic theories are noncounting (they do not rely on the full power of arithmetic) and second because it is generally not necessary for a WFC to be met at every stage of the derivation. Instead of computing the score in $\mathbb{Z}$, a better choice is some finite structure G with well-understood rules of combination, and instead of assigning a single value to each atomic expression, it gives us much-needed flexibility to make the assignment disjunctive (taking any one of a set of values). Thus we have a mapping $c : W \rightarrow 2^G$ and consider grammatical only those sequences of words for which the rules of combination yield a desirable result. Demonstrating that the assigned elements of G indeed combine in the desired manner constitutes a **certificate** of membership according to the grammar defined by c.

Example 2.2.2 Categorial grammar. If G behaves like a free group except that formal inverses of generators do not cancel from both sides ($g \cdot g^{-1} = e$ is assumed but $g^{-1} \cdot g = e$ is not) and we consider only those word sequences $w_1.w_2 \ldots w_n$ for which there is at least one h_i in each $c(w_i)$ such that $h_1 \cdot \ldots \cdot h_n = g_0$ (i.e. the group-theoretical product of the h_i yields a distinguished generator g_0), we obtain a version of *bidirectional categorial grammar* (Bar-Hillel 1953, Lambek 1958). If we take G as the free Abelian group, we obtain *unidirectional categorial grammar* (Ajdukiewitz 1935). These notions will be developed further in Chapter 5.2.

Example 2.2.3 Unification grammar. By choosing G to be the set of rooted directed acyclic node-labeled graphs, where the labels are first order variables and constants, and considering only those word sequences for which the assigned graphs will unify, we obtain a class of *unification grammars*.

Example 2.2.4 Link grammar. By choosing G to satisfy a generalized version of the (horizontal) tiling rules of Example 2.1.1, we obtain the *link grammars* of Sleator and Temperley (1993).

We will investigate a variety of such systems in detail in Chapters 5 and 6, but here we concentrate on the major differences between truth and grammaticality. First, note that systems such as those above are naturally set up to define not only one distinguished set of strings but its cosets as well. For example, in a categorial grammar, we may inquire not only about those strings of words for which multiplication of the associated categories yields the distinguished generator but also about those for which the yield contains another generator or any specific word of G. This corresponds to the fact that e.g. *the house of the seven gables* is grammatical but only as a noun phrase and not as a sentence, while *the house had seven gables* is a grammatical sentence but not a grammatical noun phrase. It could be tempting to treat the cosets in analogy with n-valued logics, but this does not work well since the various stringsets defined by a grammar may overlap (and will in fact irreducibly overlap in every case where a primitive element is assigned more than one disjunct by c), while truth values are always uniquely assigned in n-valued logic.

Second, the various calculi for propagating truth values by specific rules of inference can be supported by an appropriately constructed theory of model structures. In logic, a model will be unique only in degenerate cases: as soon as there is an infinite model, by the Löwenheim-Skolem theorems we have at least as many non-isomorphic models as there are cardinalities. In grammar, the opposite holds: as soon as we fix the period, dialect, style, and possibly other parameters determining grammaticality, the model is essentially unique.

The fact that up to isomorphism there is only one model structure M gives rise to two notions peculiar to mathematical linguistics: *overgeneration* and *undergeneration*. If there is some string $w_1.w_2 \dots w_n \not\in M$ that appears in the yield of c, we say that c **overgenerates** (with respect to M), and if there is a $w_1.w_2 \dots w_n \in M$ that does not appear in the yield of c, we say that c **undergenerates**. It is quite possible, indeed typical, for working grammars to have both kinds of errors at the same time. We will develop quantitative methods to compare the errors of different grammars in Section 5.4, and note here that neither undergeneration nor overgeneration is a definitive diagnostic of some fatal problem with the system. In many cases, overgeneration is benign in the sense that the usefulness of a system that e.g. translates English sentences to French is not at all impaired by the fact that it is also capable of translating an input that lies outside the confines of fully grammatical English. In other cases, the aim of the system may be to shed light only on a particular range of phenomena, say on the system of intransitive verbs, to the exclusion of transitive, ditransitive, etc., verbs. In the tradition of Montague grammar (see Section 6.2), such systems are explicitly called *fragments*. Constraint-based theories, which view the task of characterizing all and only the well-formed structures as one of (rank-prioritized) intersection of WFCs (see Section 4.2) can have the same under- and overgeneration problems as rule-based systems, as long as they have too many (too few) constraints.

In spite of these major differences, the practice of logic and that of grammar have a great deal in common. First, both require a systematic ability to analyze sentences in component parts so that generalizations involving only some part can be stated and the ability to construct new sentences from ones already seen. Chapter 5 will discuss such *syntactic* abilities in detail. We note here that the practice of logic is largely *normative* in the sense that constructions outside those explicitly permitted by its syntax are declared ill-formed, while the practice of linguistics is largely *descriptive* in the sense that it takes the range of existing constructions as given and strives to adjust the grammar so as to match this range.

Second, both logic and grammar are largely driven by an overall consideration of economy. As the reader will have no doubt noticed, having a separate WFC for the northern, southern, eastern, and western edges of a tile in Example 2.1.1 is quite unnecessary: any two orthogonal directions would suffice to narrow down the range of well-formed tilings. Similarly, in context-free grammars, we often find it sufficient to deal only with rules that yield only two elements on the right-hand side (Chomsky normal form), and there has to be some strong reason for departing from the simplest binary branching structure (see Chapter 5).

From the perspective of linguistics, logical calculi are generation devices, with the important caveat that in logic the rules of deduction are typically viewed as possibly having more than one premiss, while in linguistics such rules would generally be viewed as having only one premiss, namely the conjunction of the logically distinct premisses, and axiom systems would be viewed as containing a single starting point (the conjunction of the axioms). The deduction of theorems from the axiom by brute force enumeration of all proofs is what linguists would call **free generation**. The use of a single conjunct premiss instead of multiple premisses may look like a distinction without a difference, but it has the effect of making generative systems *invertible:* for each such system with rules $r_1, \ldots, r_k$ we can construct an inverted system with rules $r_1^{-1}, \ldots, r_k^{-1}$ that is now an **accepting**, rather than generating, device. This is very useful in all those cases where we are interested in characterizing both production (synthesis, generation) and perception (analysis, parsing) processes because the simplest hypothesis is that these are governed by the same set of abstract rules.

Clearly, definition by generation differs from deduction by a strict algorithmic procedure only in that the choice of the next algorithmic step is generally viewed as being completely determined by the current step, while in generation the next step is freely drawn from the set of generative rules. The all-important boundary between recursive and recursively enumerable (r.e.) is drawn the same way by certificates (derivation structures), but the systems of interest congregate on different sides of this boundary. In logic, proving the negation of a statement requires the same kind of certificate (a proof object rooted in the axioms and terminating in the desired conclusion) as proving the statement itself – the difficulty is that most calculi are r.e. but not recursive (decidable). In grammar, proving the ungrammaticality of a form requires an apparatus very different from proving its grammaticality: for the latter purpose an ordinary derivation suffices, while for the former we typically need to exhaustively survey all forms of similar and lesser complexity, which can be difficult, even though most grammars are not only r.e. but in fact recursive.

2.3 String rewriting

Given a set of atomic symbols Σ called the **alphabet**, the simplest imaginable operation is that of **concatenation**, whereby a complex symbol xy is formed from x and y by writing them in succession. Applying this operation recursively, we obtain **strings** of arbitrary *length*. Whenever such a distinction is necessary, the operation will be denoted by . (dot). The result of the dot operation is viewed as having no internal punctuation: $u.v = uv$ both for atomic symbols and for more complex strings, corresponding to the fact that concatenation is by definition associative. To forestall confusion, we mention here that in later chapters the . will also be used in *glosses* to connect a word stem to the complex of morphosyntactic (inflectional) features the word form carries: for example *geese* = *goose.PL* (the plural form of *goose* is *geese*) or Hungarian *házammal* = *house.POSS1SG.INS* ‘with my house’, where *POSS1SG* refers to the suffix that signifies possession by a first-person singular entity and *INS* refers to the instrumental case ending roughly analogous to English *with*. (The reader should be forewarned that translation across languages rarely proceeds as smoothly on a morpheme by morpheme basis as the example may suggest: in many cases morphologically expressed concepts of the source language have no exact equivalent in the language used for glossing.)

Of special interest is the **empty string** λ, which serves as a two-sided multiplicative unit of concatenation: $\lambda.u = u.\lambda = u$. The whole set of strings generated from Σ by concatenation is denoted by Σ^+ (λ-**free Kleene closure**) or, if the empty string is included, by Σ^* (**Kleene closure**). If $u.v = w$, we say that u (v) is a **left (right) factor** of w. If we define the **length** $l(x)$ of a string x as the number of symbols in x, counted with multiplicity (the empty word has length 0), l is a homomorphism from Σ^* to the additive semigroup of nonnegative integers. In particular, the semigroup of nonnegative integers (with ordinary addition) is isomorphic to the Kleene closure of a one-symbol alphabet (with concatenation): the latter may be called integers in **base one** notation.

Subsets of Σ^* are called **stringsets**, **formal languages**, or just **languages**. In addition to the standard Boolean operations, we can define the **concatenation** of strings and languages U and V as $UV = \{uv|u \in U, v \in V\}$, suppressing the distinction between a string and a one-member language, writing xU instead of $\{x\}U$, etc. The (λ-free) Kleene closure of strings and languages is defined analogously to the closure of alphabets. For a string w and a language U, we say $u \in L$ is a **prefix** of w if u is a left factor of w and no smaller left factor of w is in U.

Finite languages have the same distinguished status among all stringsets that the natural numbers $\mathbb{N}$ have among all numbers: they are, after all, all that can be directly listed without relying on any additional interpretative mechanism. And as in arithmetic, where the simplest natural superset of the integers includes not only finite decimal fractions but some infinite ones as well, the simplest natural to Type 0, Chomsky (1956) demonstrated that each type is properly contained in the next lower one. These proofs, together with examples of context-free but not regular, context-sensitive but not context-free, and recursive but not context-sensitive languages, are omitted here, as they are discussed in many excellent textbooks of formal language theory such as Salomaa (1973) or Harrison (1978). To get a better feel for CSLs, we note the following results:

Theorem 2.3.3 (Karp 1972) The membership problem for CSLs is PSPACE-complete.

Theorem 2.3.4 (Szelepcsényi 1987, Immerman 1988) The complement of a CSL is a CSL.

Exercise 2.5 Construct three CSGs that generate the language F of Fibonacci numbers in base one, the language F_2 of Fibonacci numbers in base two, and the language F_{10} of Fibonacci numbers in base ten. Solve the membership problem for 117467.

Exercise 2.6 Call a set of natural numbers k-regular if their base k representations are a regular language over the alphabet of k digits. It is easy to see that a 1-regular language is 2-regular (3-regular) and that the converse is not true. Prove that a set that is both 2-regular and 3-regular is also 1-regular.

2.4 Further reading

Given that induction is as old as mathematics itself (the key idea going back at least to Euclid's proof that there are infinitely many primes) and that recursion can be traced back at least to Fibonacci's (1202) *Liber Abaci*, it is somewhat surprising that the closely related notion of generation is far more recent: the first systematic use is in von Dyck (1882) for free groups. See Chandler and Magnus (1982 Ch. I.7) for some fascinating speculation why the notion did not arise earlier within group theory. The kernel membership problem is known as the *word problem* in this setting (Dehn 1912). The use of freely generated pure formula models in logic was pioneered by Herbrand (1930); Wang tilings were introduced by Wang (1960). Theorem 2.1.1 was proven by Berger (1966), who demonstrated the undecidability by encoding the halting problem in tiles. For a discussion, see Gruska (1997 Sec. 6.4.3). The notion that linguistic structures are noncounting goes back at least to Chomsky (1965:55).

From Pāṇini to the *neogrammarians* of the 19th century, linguists were generally eager to set up the system so as to cover related styles, dialects, and historical stages of the same language by minor variants of the same theory. In our terms this would mean that e.g. British English and American English or Old English and Modern English would come out as models of a single 'abstract English'. This is one point where current practice (starting with de Saussure) differs markedly from the traditional approach. Since grammars are intended as abstract theories of the native speaker's competence, they cannot rely on data that are not observable by the ordinary language learner. In particular, they are restricted to a single temporal slice, called the *synchronic* view by de Saussure, as opposed to a view encompassing different historical stages (called the *diachronic* view). Since the lack of cross-dialectal or historical data is never an impediment in the process of children acquiring their native language (children are capable of constructing their internal grammar without access to such data), by today's standards it would raise serious methodological problems for the grammarian to rely on facts outside the normal range of input available to children. (De Saussure actually arrived at the synchrony/diachrony distinction based on somewhat different considerations.) The neogrammarians amassed a great deal of knowledge about *sound change*, the historical process whereby words change their pronunciation over the centuries, but some of their main tenets, in particular the exceptionlessness of sound change laws, have been found not to hold universally (see in particular Wang 1969, Wang and Cheng 1977, Labov 1981, 1994).

Abstract string manipulation begins with Thue (1914, reprinted in Nagell 1977), who came to the notion from combinatorial group theory. For Thue, rewriting is symmetrical: if AXB can be rewritten as AYB the latter can also be rewritten as the former. This is how Harris (1957) defined transformations. The direct precursors of the modern generative grammars and transformations that were introduced by Chomsky (1956, 1959) are semi-Thue systems, where rewriting need not necessarily work in both directions. The basic facts about regular languages, finite automata, and Kleene's theorem are covered in most textbooks about formal language theory or the foundations of computer science, see e.g. Salomaa (1973) or Gruska (1997). We will develop the connection between these notions and semigroup theory along the lines of Eilenberg (1974) in Chapter 5. Context-free grammars and languages are also well covered in computer science textbooks such as Gruska (1997), for more details on context-sensitivity, see Section 10 of Salomaa (1973). Theorem 2.3.1 was discovered in (McCawley 1968), for a rigorous proof see Peters and Ritchie (1973), and for a modern discussion, see Oehrle (2000).

Some generalizations of the basic finite state notions that are of particular interest to phonologists, namely regular relations, and finite k-automata, will be discussed in Chapter 3. Other generalizations, which are also relevant to syntax, involve weighted (probabilistic) languages, automata, and transducers – these are covered in Sections 5.4 and 5.5. Conspiracies were first pointed out by Kisseberth (1970) – we return to this matter in Section 4.3. The founding papers on categorial grammars are Ajdukiewicz (1935) and Lambek (1958). Unification grammars are discussed in Shieber (1986, 1992).

3

Phonology

The fundamental unit of linguistics is the *sign*, which, as a first approximation, can be defined as a conventional pairing of sound and meaning. By *conventional* we mean both that signs are handed down from generation to generation with little modification and that the pairings are almost entirely arbitrary, just as in bridge, where there is no particular reason for a bid of two clubs in response to one no trump to be construed as an inquiry about the partner's major suits. One of the earliest debates in linguistics, dramatized in Plato's *Cratylus*, concerns the arbitrariness of signs. One school maintained that for every idea there is a true sound that expresses it best, something that makes a great deal of sense for *onomatopoeic* words (describing e.g. the calls of various animals) but is hard to generalize outside this limited domain. Ultimately the other school prevailed (see Lyons 1968 Sec. 1.2 for a discussion) at least as far as the word-level pairing of sound and meaning is concerned.

It is desirable to build up the theory of sounds without reference to the theory of meanings both because the set of atomic units of sound promises to be considerably simpler than the set of atomic units of meanings and because sounds as linguistic units appear to possess clear physical correlates (acoustic waveforms; see Chapter 8), while meanings, for the most part, appear to lack any direct physical embodiment. There is at least one standard system of communication, Morse code, that gets by with only two units, dot (short beep) and dash (long beep) or possibly three, (if we count pause/silence as a separate unit; see Ex. 7.7). To be sure, Morse code is parasitic on written language, which has a considerably larger alphabet, but the enormous success of the alphabetic mode of writing itself indicates clearly that it is possible to analyze speech sounds into a few dozen atomic units, while efforts to do the same with meaning (such as Wilkins 1668) could never claim similar success.

There is no need to postulate the existence of some alphabetic system for transcribing sounds, let alone a meaning decomposition of some given kind. In Section 3.1 we will start with easily observable entities called *utterances*, which are defined as maximal pause-free stretches of speech, and describe the concatenative building blocks of sound structure called *phonemes*. For each natural language L these will act as a convenient set of atomic symbols P_L that can be manipulated by context-sensitive string-rewriting techniques, giving us what is called the *segmental* *phonology* of the language. This is not to say that the set of words W_L, viewed as a formal language over P_L, will be context-sensitive (Type 1) in the sense of formal language theory. On the contrary, we have good reasons to believe that W is in fact regular (Type 3).

To go beyond segments, in Section 3.2 we introduce some subatomic components called *distinctive features* and the formal linguistic mechanisms required to handle them. To a limited extent, distinctive features pertaining to tone and stress are already useful in describing the *suprasegmental phonology* of languages. To get a full understanding of suprasegmentals in Section 3.3 we introduce *multitiered* data structures more complex than strings, composed of *autosegments.* Two generalizations of regular languages motivated by phonological considerations, regular transducers and regular k-languages, are introduced in Section 3.4. The notions of prosodic hierarchy and optimality, being equally relevant for phonology and morphology, are deferred to Chapter 4.

3.1 Phonemes

We are investigating the very complex *interpretation relation* that obtains between certain structured kinds of sounds and certain structured kinds of meanings; our eventual goal is to define it in a generative fashion. At the very least, we must have some notion of identity that tells us whether two signs sound the same and/or mean the same. The key idea is that we actually have access to more information, namely, whether two utterances are *partially similar* in form and/or meaning. To use Bloomfield's original examples:

> A needy stranger at the door says *I'm hungry*. A child who has eaten and merely wants to put off going to bed says *I'm hungry*. Linguistics considers only those vocal features which are alike in the two utterances . . . Similarly, *Put the book away* and *The book is interesting* are partly alike *(the book).*

That the same utterance can carry different meanings at different times is a fact we shall not explore until we introduce *disambiguation* in Chapter 6 – the only burden we now place on the theory of meanings is that it be capable of (i) distinguishing meaningful from meaningless and (ii) determining whether the meanings of two utterances share some aspect. Our expectations of the observational theory of sound are similarly modest: we assume we are capable of (i′) distinguishing pauses from speech and (ii′) determining whether the sounds of two utterances share some aspect.

We should emphasize at the outset that the theory developed on this basis does not rely on our ability to exercise these capabilities to the extreme. We have not formally defined what constitutes a pause or silence, though it is evident that observationally such phenomena correspond to very low acoustic energy when integrated over a period of noticeable duration, say 20 milliseconds. But it is not necessary to be able to decide whether a 19.2 millisecond stretch that contains exactly 1.001 times the physiological minimum of audible sound energy constitutes a pause or not. If this stretch is indeed a pause we can always produce another instance, one that will have a significantly larger duration, say 2000 milliseconds, and containing only one-tenth of the previous energy. This will show quite unambiguously that we had two utterances in the first place. If it was not a pause, but rather a functional part of sound formation such as a stop closure, the new ‘utterances’ with the artificially interposed pause will be deemed ill-formed by native speakers of the language. Similarly, we need not worry a great deal whether *Colorless green ideas sleep furiously* is meaningful, or what it exactly means. The techniques described here are robust enough to perform well on the basis of ordinary data without requiring us to make ad hoc decisions in the edge cases. The reason for this robustness comes from the fact that when viewed as a probabilistic ensemble, the edge cases have very little weight (see Chapter 8 for further discussion).

The domain of the interpretation relation I is the set of *forms* F, and the codomain is the set of *meanings* M, so we have $I \subset F \times M$. In addition, we have two *overlap* relations, $O_F \subset F \times F$ and $O_M \subset M \times M$, that determine partial similarity of form and meaning respectively. O_F is traditionally divided into *segmental* and *suprasegmental* overlaps. We will discuss mostly segmental overlap here and defer suprasegmentals such as tone and stress to Section 3.3 and Section 4.1, respectively. Since speech happens in time, we can define two forms α and β as *segmentally overlapping* if their temporal supports as intervals on the real line can be made to overlap, as in the *the book* example above. In the segmental domain at least, we therefore have a better notion than mere overlap: we have a partial ordering defined by the usual notion of interval containment. In addition to O_F, we will therefore use sub- and superset relations (denoted by $\subset_F, \supset_F$) as well as intersection, union, and complementation operations in the expected fashion, and we have

$$\alpha \cap_F \beta \neq \emptyset \Rightarrow \alpha O_F \beta \tag{3.1}$$

In the domain of I, we find obviously complex forms such as a full epic poem and some that are atomic in the sense that

$$\forall x \subset_F \alpha : x \notin dom(I) \tag{3.2}$$

These are called *minimum forms*. A form that can stand alone as an utterance is a *free form*; the rest (e.g. forms like *ity* or *al* as in *electricity*, *electrical*), which cannot normally appear between pauses, are called *bound forms*.

Typically, utterances are full phrases or sentences, but when circumstances are right, e.g. because a preceding question sets up the appropriate context, forms much smaller than sentences can stand alone as complete utterances. Bloomfield (1926) defines a *word* as a minimum free form. For example, *electrical* is a word because it is a free form (can appear e.g. as answer to the question *What kind of engine is in this car?*) and it cannot be decomposed further into free forms (*electric* would be free but *al* is bound). We will have reason to revise this definition in Chapter 4, but for now we can provisionally adopt it here because in defining phonemes it is sufficient to restrict ourselves to free forms.

For the rest of this section, we will only consider the set of words $W \subset F$, and we are in the happy position of being able to ignore the meanings of words entirely.

the existence of phonetic alphabets can be derived from postulates rooted in these limitations.

3.2 Natural classes and distinctive features

Isolating the atomic segmental units is a significant step toward characterizing the phonological system of a language. Using the phonemic alphabet P, we can write every word as a string $w \in P^*$, and by adding just one extra symbol # to denote the pause between words, we can write all utterances as strings over $P \cup \{\#\}$. Since in actual *connected* speech pauses between words need not be manifest, we need an interpretative convention that # can be *phonetically realized* either as silence or as the empty string (zero realization). Silence, of course, is distinctly audible and has positive duration (usually 20 milliseconds or longer), while λ cannot be heard and has zero duration.

In fact, similar interpretative conventions are required throughout the alphabet, e.g. to take care of the fact that in English word-initial t is *aspirated* (released with a puff of air similar in effect to h but much shorter), while in many other positions t is *unaspirated* (released without an audible puff of air): compare *ton* to *stun*. The task of relating the abstract units of the alphabet to their audible manifestations is a complex one, and we defer the details to Chapter 9. We note here that the interpretation process is by no means trivial, and there are many unassailable cases, such as aspirated vs. unaspirated t and silenceful vs. empty #, where we permit two or more alternative realizations for the same segment. (Here and in what follows we reserve the term **segment** for alphabetic units; i.e. strings of length one.)

Since λ can be one of the alternatives, an interesting technical possibility is to permit cases where it is the only choice: i.e. to declare elements of a phonemic alphabet that never get realized. The use of such *abstract* or **diacritic** elements *(anubandha)* is already pivotal in Pāṇini's system and remains characteristic of phonology to this day. This is our first example of the linguistic distinction between *underlying* (abstract) and *surface* (concrete) forms – we will see many others later.

Because in most cases alternative realizations of a symbol are governed by the symbols in its immediate neighborhood, the mathematical tool of choice for dealing with most of segmental phonology is string rewriting by means of context-sensitive rules. Here a word of caution is in order: from the fact that context-sensitive rules are used it does not follow that the generated stringset over P, or over a larger alphabet Q that includes abstract elements as well, will be context-sensitive. We defer this issue to Section 3.4, and for now emphasize only the convenience of context-sensitive rules, which offer an easy and well-understood mechanism to express the phonological regularities or *sound laws* that have been discovered over the centuries.

Example 3.2.1 Final devoicing in Russian. The nominative form of Russian nouns can be predicted from their dative forms by removing the dative suffix u and inspecting the final consonant: if it is b or p, the final consonant of the nominative form will be p. This could be expressed in a phonological rule of *final b devoicing*: $b \to p/\_\#$.

When it is evident that the change is caused by some piece of the environment where the rule applies, we speak of the piece *triggering* the change; here the trigger is the final #.

Remarkably, we find that a similar rule links d to t, g to k, and in fact any voiced obstruent to its voiceless counterpart. The phenomenon that the structural description and/or the structural change in rules extends to some disjunction of segments is extremely pervasive. Those sets of segments that frequently appear together in rules (either as triggers or as undergoers) are called *natural classes*; for example, the class $\{p, t, k\}$ of *unvoiced stops* and the class $\{b, d, g\}$ of *voiced stops* are both natural, while the class $\{p, t, d\}$ is not. Phonologists would be truly astonished to find a language where some rule or regularity affects *p, t*, and *d* but no other segment.

The linguist has no control over the phonemic alphabet of a language: P is computed as the result of a specific (oracle-based, but otherwise deterministic) algorithm. Since the set $N \subset 2^P$ of natural classes is also externally given by the phonological patterning of the language, over the millennia a great deal of effort has been devoted to the problem of properly characterizing it, both in order to shed some light on the structure of P and to help simplify the statement of rules.

So far, we have treated P as an unordered set of alphabetic symbols. In the Ashṭādhyāyī, Pāṇini arranges elements of P in a linear sequence (the *śivasūtras*) with some abstract (phonetically unrealized) symbols *(anubandha)* interspersed. Simplifying his treatment somewhat (for a fuller discussion, see Staal 1962), natural classes *(pratyāhāra)* are defined in his 1.1.71 as those subintervals of the *śivasūtras* that end in some *anubandha*. If there are k symbols in P, in principle there could be as many as 2^k natural classes. However, the Pāṇinian method will generate at most $k(k+1)/2$ subintervals (or even fewer, if diacritics are used more sparingly), which is in accordance with the following postulate.

Postulate 3.2.1 In any language, the number of natural classes is small.

We do not exactly spell out what ‘small’ means here. Certainly it has to be polynomial, rather than exponential, in the size of P. The European tradition reserves names for many important natural classes such as the *apicals*, *aspirates*, *bilabials*, *consonants*, *continuants*, *dentals*, *fricatives*, *glides*, *labiodentals*, *linguals*, *liquids*, *nasals*, *obstruents*, *sibilants*, *stops*, *spirants*, *unaspirates*, *velars*, *vowels*, etc. – all told, there could be a few hundred, but certainly not a few thousand, such classes. As these names suggest, the reason why a certain class of sounds is natural can often be found in sharing some aspects of production (e.g. all sounds crucially involving a constriction at the lips are *labials*, and all sounds involving turbulent airflow are *fricatives*), but often the justification is far more complex and indirect. In some cases, the matter of whether a particular class is natural is heavily debated. For a particularly hard chestnut, the *ruki* class; see Section 9.2, Collinge’s (1985) discussion of Pedersen’s law I, and the references cited therein.

For the mathematician, the first question to ask about the set of natural classes N is neither its size nor its exact membership but rather its algebraic structure: under what operations is N closed? To the extent that Pāṇini is right, the structure is not fully Boolean: the complement of an interval typically will not be expressible as a single interval, but the intersection of two intervals *(pratyāhāra)* will again be an interval. We state this as the following postulate.

Postulate 3.2.2 In any language, the set of natural classes is closed under intersection.

This postulate makes N a meet semilattice, and it is clear that the structure is not closed under complementation since single segments are natural classes but their complements are not. The standard way of weakening the Boolean structure is to consider meet semilattices of linear subspaces. We embed P in a hypercube so that natural classes correspond to hyperplanes parallel to the axes. The basis vectors that give rise to the hypercube are called **distinctive features** and are generally assumed to be binary; a typical example is the *voiced/unvoiced* distinction that is defined by the presence/absence of periodic vocal fold movements. It is debatable whether the field underlying this vector space construct should be $\mathbb{R}$ or GF(2). We take the second option and use GF(2), but we will have reason to return to the notion of real-valued features in Chapters 8 and 9. Thus, we define a **feature assignment** as an injective mapping C from the set Q of segments into the linear space GF(2,n).

This is a special case of a general situation familiar from universal algebra: if A_i are algebras of the same signature and $A = \prod A_i$ is their direct product, we say that a subalgebra B of A is a **subdirect product** of the A_i if all its projections on the components A_i are surjective. A classic theorem of Birkhoff asserts that every algebra can be represented as a subdirect product of subdirectly irreducible algebras. Here the algebras are simply finite sets, and as the only subdirectly irreducible sets have one or two members (and one-member sets obviously cannot contribute to a product), we obtain distinctive feature representations (also called **feature decompositions**) for any set for free.

Since any set, not just phonological segments, could be defined as vectors (also called *bundles*) of features, to give feature decomposition some content that is specific to phonology we must go a step further and link natural classes to this decomposition. This is achieved by defining as **natural classes** those sets of segments that can be expressed by fewer features than their individual members (see Halle 1964:328). To further simplify the use of natural classes, we assume a theory of *markedness* (Chomsky and Halle 1968 Ch. IX) that supplies those features that are predictable from the values already given (see Section 7.3). For example, high vowels will be written as $\begin{bmatrix} +\text{syll} \\ +\text{high} \end{bmatrix}$, requiring only two features, because the other features that define this class, such as [−low] or [+voice], are predictable values already given.

In addition to using *pratyāhāra*, Pāṇini employs a variety of other devices, most notably the concept of ‘homogeneity’ *(sāvarṇya)*, as a means of cross-classification (see Cardona 1965). This device enables him to treat quality distinctions in vowels separately from length, nasality, and tone distinctions, as well as to treat place of articulation distinctions in consonants separately from nasality, voicing, and aspiration contrasts. Another subsidiary concept, that of *antara* ‘nearness’, is required to handle the details of mappings between natural classes. Since Pāṇinian rules always map classes onto classes, the image of a segment under a rule is decided by P1.1.50

sthāne 'ntaratamaḥ ‘in replacement, the nearest’. The modern equivalent of P1.1.50 is the convention that features unchanged by a rule need not be explicitly mentioned, so that the Russian final devoicing rule that we began with may simply be stated as [+obstruent] → [−voice] / _#.

For very much the same empirical reasons that forced Pāṇini to introduce additional devices like *sāvarṇya*, the contemporary theory of features also relaxes the requirement of full orthogonality. One place where the standard (Chomsky and Halle 1968) theory of distinctive features shows some signs of strain is the treatment of vowel height. Phonologists and phoneticians are in broad agreement that vowels come in three varieties, *high*, *mid*, and *low*, which form an interval structure: we often have reason to group high and mid vowels together or to group mid and low vowels together, but we never see a reason to group high and low vowels together to the exclusion of mid vowels. The solution adopted in the standard theory is to use two binary features, [± high] and [± low], and to declare the conjunction [+high, +low] ill-formed.

Similar issues arise in many other corners of the system; e.g. in the treatment of *place of articulation* features. Depending on where the major constriction that determines the type of a consonant occurs, we distinguish several places of articulation, such as *bilabial*, *labiodental*, *dental*, *alveolar*, *postalveolar*, *retroflex*, *palatar*, *velar*, *pharyngeal*, *epiglottal*, and *glottal*, moving back from the lips to the glottis inside the vocal tract. No single language has phonemes at every point of articulation, but many show five-, or six-way contrasts. For example, Korean distinguishes bilabial, dental, alveolar, velar, and glottal, and the difference is noted in the basic letter shape (□, ∨, ↼, ⇁, and ◯, respectively). Generally, there is more than one consonant per point of articulation; for example, English has alveolars $n,\ t,\ d,\ s,\ z,\ l$. Consonants sharing the same place of articulation are said to be *homorganic* and they form a natural class (as can be seen e.g. from rules of nasal assimilation that replace e.g. *input* by *imput*).

Since the major classes (labial, coronal, dorsal, radical, laryngeal) show a five-way contrast, the natural way to deal with the situation would be the use of one GF(5)-valued feature rather than three (or more) underutilized GF(2) values, but for reasons to be discussed presently this is not a very attractive solution. What the system really needs to express is the fact that some features tend to occur together in rules to the exclusion of others, a situation somewhat akin to that observed among the segments. The first idea that leaps to mind would be to utilize the same solution, using features of features *(metafeatures)* to express natural classes of features. The Cartesian product operation that is used in the feature decomposition (subdirect product form) of P is associative, and therefore it makes no difference whether we perform the feature decomposition twice in a metafeature setup, or just once at the segment level. Also, the inherent ordering of places of articulation (for consonants) or height (for vowels) is very hard to cenvey by features, be they 2-valued or n-valued, without recourse to arithmetic notions, something we would very much like to avoid as it would make the system overly expressive.

The solution now widely accepted in phonology (Clements 1985, McCarthy 1988) is to arrange the features in a tree structure, using intermediate **class nodes**

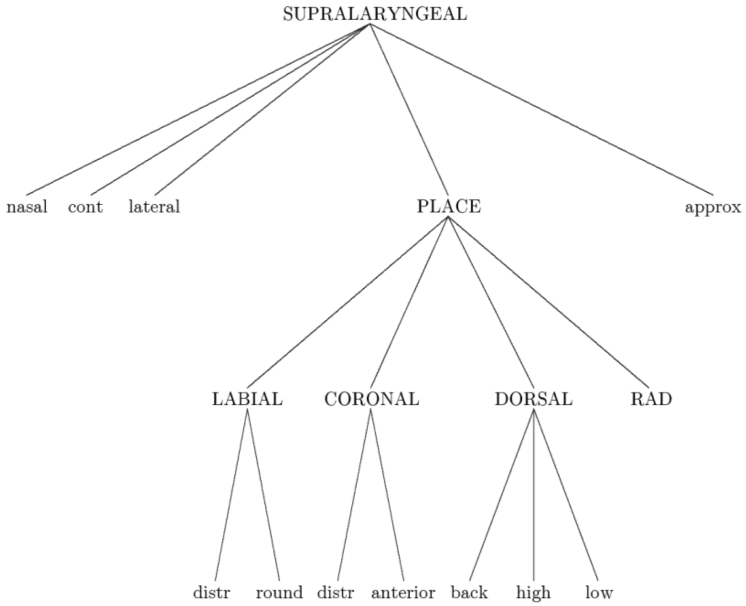

Fig. 3.1. Feature geometry tree. Rules that required the special principle of *sāvarṇya* can be stated using the *supralaryngeal class node*

to express the grouping together of some features to the exclusion of others (see Fig. 3.1). This solution, now permanently (mis)named **feature geometry**, is in fact a generalization of both the pratyāhāra and the standard feature decomposition methods. The linear intervals of the Pāṇinian model are replaced by generalized (lattice-theoretic) intervals in the subsumption lattice of the tree, and the Cartesian product appearing in the feature decomposition corresponds to the special case where the feature geometry tree is a star (one distinguished root node, all other nodes being leaves).

Discussion The segmental inventories P developed in Section 3.1 are clearly different from language to language. As far as natural classes and feature decomposition are concerned, many phonologists look for a single universal inventory of features arranged in a universally fixed geometry such as the one depicted in Fig. 3.1. Since the cross-linguistic identity of features such as [nasal] is anchored in their phonetic (acoustic and articulatory) properties rather than in some combinatorial subtleties of their intralanguage phonological patterning, this search can lead to a single object, unique up to isomorphism, that will, much like Mendeleyev's periodic table, encode a large number of regularities in a compact format.

Among other useful distinctions, Chomsky and Halle (1968) introduce the notion of *formal* vs. *substantive universals*. Using this terminology, meet semilattices are a formal, and a unique feature geometry tree such as the one in Fig. 3.1 would be a substantive, universal. To the extent that phonological research succeeds in identifying a

Each tier N has its own **tier alphabet** T_N, and we can assume without loss of generality that the alphabets of different tiers are disjoint except for a distinguished **blank** symbol G (purposely kept distinct from the pause symbol #) that is adjoined to every tier alphabet. Two tiers bearing identical names can only be distinguished by inspecting their contents. We define a tier containing a string $t_0t_1 \ldots t_n$ starting at position k by a mapping that maps k on t_0, $k+1$ on $t_1, \ldots, k+n$ on t_n, and everything else on G. Abstracting away from the starting position, we have the following definition.

Definition 3.3.2 A tier N **containing** a string $t_0t_1 \ldots t_n$ over the alphabet $T_N \cup^* G$ is defined as the class of mappings F_k that take $k+i$ into t_i for $0 \leq i \leq n$ and to G if i is outside this range. Unless noted otherwise, this class will be represented by the mapping F_0. Strings containing any number of successive G symbols are treated as equivalent to those strings that contain only a single G at the same position. G-free strings on a given tier are called **melodies**.

Between strings on the same tier and within the individual strings, temporal ordering is encoded by their usual left-to-right ordering. The temporal ordering of strings on different tiers is encoded by association relations.

Definition 3.3.3 An **association relation** between two tiers N and M containing the strings n $= n_0n_1 \ldots n_k$ and m $= m_0m_1 \ldots m_l$ is a subset of $\{0, 1, \ldots, k\} \times \{0, 1, \ldots, l\}$. An element that is not in the domain or range of the association relation is called **floating**.

Note that the association relation, being an abstract pattern of synchrony between the tiers, is one step removed from the content of the tiers: association is defined on the *domain* of the representative mappings, while content also involves their *range*. By Definition 3.3.3, there are 2^{kl} association relations possible between two strings of length k and l. Of these relations, the *no crossing constraint* (NCC; see Goldsmith 1976) rules out as ill-formed all relations that contain pairs (i, v) and (j, u) such that $0 \leq i < j \leq k$ and $0 \leq u < v \leq l$ are both true. We define the **span** of an element x with respect to some association relation A as those elements y for which (x, y) is in A. Rolling the definitions above into one, we have the following definition.

Definition 3.3.4 A **bistring** is an ordered triple (f, g, A), where f and g are strings not containing G, and A is a well-formed association relation over two tiers containing f and g.

In the general case, we have several tiers arranged in a tree structure called the geometry of the representation (see Section 3.2). Association relations are permitted only among those tiers that are connected by an edge of this tree, so if there are k tiers there will be $k-1$ relations. Thus, in the general case, we define a k-**string** as a $(2k-1)$-tuple $(s_1, \ldots, s_k, A_1, \ldots, A_{k-1})$, where the s_i are strings and the A_i are association relations.

Theorem 3.3.1 The number of well-formed association relations over two tiers, each containing a string of length n, is asymptotically $(6 + 4\sqrt{2})^n$.

Proof Let us denote the number of well-formed association relations with n symbols on the top tier and k symbols on the bottom tier by $f(n, k)$. By symmetry, $f(n, k) = f(k, n)$, and obviously $f(n, 1) = f(1, n) = 2^n$. By enumerating relations according to the pair (i, j) such that no $i' < i$ is in the span of any j' and no $j'' > j$ is in the span of i, we get

$$f(n+1, k+1) = \sum_{i=1}^{k+1} f(n,i)2^{k+1-i} + f(n, k+1) \tag{3.3}$$

From (3.3) we can derive the following recursion:

$$f(n+1, k+1) = 2f(n+1, k) + 2f(n, k+1) - 2f(n, k) \tag{3.4}$$

For the first few values of $a_n = f(n, n)$, we can use (3.4) to calculate forward: $a_1 = 2$, $a_2 = 12$, $a_3 = 104$, $a_4 = 1008$, $a_5 = 10272$, $a_6 = 107712$, and so on. Using (3.4) we can also calculate backward and define $f(0, n) = f(n, 0)$ to be 1 so as to preserve the recursion. The generating function

$$F(z, w) = \sum_{i,j=0}^{\infty} f(i, j) z^i w^j \tag{3.5}$$

will therefore satisfy the equation

$$F(z, w) = \frac{1 - \frac{z}{1-z} - \frac{w}{1-w}}{1 - 2z - 2w + 2zw} \tag{3.6}$$

If we substitute $w = t/z$ and consider the integral

$$\frac{1}{2\pi i} \int_C \frac{F(z, t/z)}{z} dz \tag{3.7}$$

this will yield the constant term $\sum_{n=0}^{\infty} f(n, n) t^n$ by Cauchy's formula. Therefore, in order to get the generating function

$$d(t) = \sum_{i=0}^{\infty} a_n t^n \tag{3.8}$$

we have to evaluate

$$\frac{1}{2\pi i} \int_C \frac{1 - \frac{z}{1-z} - \frac{t/z}{1-t/z}}{z(1 - 2z - 2t/z + 2t)} dz \tag{3.9}$$

which yields

$$d(t) = 1 + \frac{2t}{\sqrt{1 - 12t + 4t^2}} \tag{3.10}$$

$d(t)$ will thus have its first singularity when $\sqrt{1 - 12t + 4t^2}$ vanishes at $t_0 = (3 - \sqrt{8})/2$, yielding

$$a_n \approx (6 + 4\sqrt{2})^n \tag{3.11}$$

the desired asymptotics. ■

The base 2 logarithm of this number, $n \cdot 3.543$, measures how many bits we need to encode a bistring of length n. Note that this number grows linearly in the length of the bistring, while the number of (possibly ill-formed) association relations was 2^{n^2}, with the base 2 log growing quadratically. Association relations in general are depicted as bipartite graphs (pairs in the relation are called **association lines**) and encoded as two-dimensional arrays (the incidence matrix of the graph). However, the linear growth of information content suggests that well-formed association relations should be encoded as one-dimensional arrays or strings. Before turning to this matter in Section 3.4, let us first consider two particularly well-behaved classes of bistrings. A bistring is **fully associated** if there are no floating elements and **proper** if the span of any element on one tier will form a single substring on the other tier (Levin 1985). Proper relations are well-formed but not necessarily fully associated.

Let us define $g(i, j)$ as the number of association relations containing no unassociated (floating) elements and define b_n as $g(n, n)$. By counting arguments similar to those used above, we get the recursion

$$g(n+1, k+1) = g(n+1, k) + g(n, k+1) + g(n, k) \tag{3.12}$$

Using this recursion, the first few values of b_n can be computed as 1, 3, 13, 63, 321, 1683, 8989, and so on. Using (3.12) we can calculate backward and define $g(0,0)$ to be 1 and $g(i,0) = g(0,i)$ to be 0 (for $i > 0$) so as to preserve the recursion. The generating function

$$G(z, w) = \sum_{i,j=0}^{\infty} g(i, j) z^i w^j \tag{3.13}$$

will therefore satisfy the equation

$$G(z, w) = \frac{1 - z - w}{1 - z - w - zw} = 1 + \frac{zw}{1 - z - w - zw} \tag{3.14}$$

Again we substitute $w = t/z$ and consider the integral

$$\frac{1}{2\pi i} \int_C \frac{G(z, t/z)}{z} dz \tag{3.15}$$

which will yield the constant term $\sum_{n=0}^{\infty} g(n, n) t^n$ by Cauchy's formula. Therefore, in order to get the generating function

$$e(t) = \sum_{i=0}^{\infty} b_n t^n \tag{3.16}$$

we have to evaluate

$$\frac{1}{2\pi i} \int_C \frac{1}{z} + \frac{t}{z(1 - z - t/z - t)} dz = 1 - \frac{t}{2\pi i} \int_C \frac{dz}{(z - p)(z - q)} \tag{3.17}$$

which yields

$$e(t) = 1 + \frac{t}{\sqrt{1 - 6t + t^2}} \tag{3.18}$$

Notice that

$$e(2t) = 1 + \frac{2t}{\sqrt{1 - 6 \cdot 2t + (2t)^2}} = d(t) \tag{3.19}$$

and thus

$$\sum_{i=0}^{\infty} b_n (2t)^n = \sum_{i=0}^{\infty} a_n t^n \tag{3.20}$$

Since the functions $d(t)$ and $e(t)$ are analytic in a disk of radius 1/10, the coefficients of their Taylor series are uniquely determined, and we can conclude that

$$b_n 2^n = a_n \tag{3.21}$$

meaning that fully associated bistrings over n points are only an exponentially vanishing fraction of all well-formed bistrings. In terms of information content, the result means that fully associated bistrings of length n can be encoded using *exactly* one bit less per unit length than arbitrary well-formed bistrings.

Exercise 3.3* Find a 'bijective' proof establishing (3.21) by direct combinatorial methods.

Now, for proper representations, denoting their number by $h(n, k)$, the generating function $H = H(z, w)$ will satisfy a functional equation

$$H - zH - wH - 2zwH + zw^2H + z^2wH - z^2w^2H = r(z, w) \tag{3.22}$$

where $r(z, w)$ is rational. Using the same diagonalizing substitution $w = t/z$, we have to evaluate

$$\frac{1}{2\pi i} \int_C \frac{s(z, t)}{z(1 - z - t/z - 2t + t^2/z + tz - t^2)} dz \tag{3.23}$$

Again, the denominator is quadratic in z, and the radius of convergence is determined by the roots of the discriminant

$$(t^2 + 2t - 1)^2 - 4(t - 1)(t^2 - t) = t^4 + 10t^2 - 8t + 1 \tag{3.24}$$

The reciprocal of the smallest root of this equation, approximately 6.445, gives the base for the asymptotics for c_n, the number of proper bistrings over n points. By taking the base 2 logarithm, we have the following theorem.

Theorem 3.3.2 The information content of a fully associated (proper) well-formed bistring is 2.543 (2.688) bits per unit length.

Exercise 3.4 Count the number of well-formed (fully associated, proper) k-strings of length n assuming each tier alphabet has only one element besides G.

Sets of well-formed (fully associated, proper) bistrings will be called well-formed (fully associated, proper) **bilanguages**. These can undergo the usual set-theoretic operations of **intersection**, **union**, and **complementation** (relative to the 'universal set' of well-formed, fully associated, resp. proper bistrings). **Reversal** (mirror image) is defined by reversing the constituent strings together with the association relation. The concatenation of bistrings is defined by concatenating both the strings and the relations:

Definition 3.3.5 Given two bistrings (f, h, A) and (k, l, B) on tiers N and M, their **concatenation** (fk, hl, AB) is constructed via the tier-alphabet functions $F_0, H_0, K_{|f|}$, and $L_{|g|}$ as follows. $FK_0(i) = F(i)$ for $0 \leq i < |f|$, $K_{|f|}(i)$ for $|f| \leq i < |f| + |k|$, G otherwise. $HL_0(j) = H(j)$ for $0 \leq j < |k|$, $L_{|k|}(j)$ for $|k| \leq j < |f|+|k|$, G otherwise. Finally, $AB = A \cup \{(i+|f|, j+|k|)|(i, j) \in B\}$.

Notice that the concatenation of two connected bistrings will not be connected (as a bipartite graph). This is remedied by the following definition.

Definition 3.3.6 Given two bistrings as in 3.3.5, their **t-catenation (b-catenation)** is defined as (fk, hl, AtB) (fk, hl, AbB), where $AtB = AB \cup \{(|f| - 1, |k|)\}$ $(AbB = AB \cup \{(|f|, |k| - 1)\})$.

Using phonological terminology, in t-catenation the last element of the *t*op tier of the first bistring is *spread* on the first element of the bottom tier of the second bistring, and in b-catenation the last element of the *b*ottom tier of the first string is spread on the first element of the top tier of the second bistring.

The only autosegmental operation that is not the straightforward generalization of some well-known string operation is that of **alignment**. Given two bistrings $x = (f, g, A)$ and $y = (g, h, B)$, their alignment $z = x \parallel y$ is defined to be (f, h, C), where C is the relation composition of A and B. In other words, the pair (i, k) will be in C iff there is some j such that (i, j) is in A and (j, k) is in B. Now we are in a position to define projections. These involve some subset S of the tier alphabet T. A **projector** $P_S(h)$ of a string $g = h_0h_1 \ldots h_m$ with respect to a set S is the bistring (h, h, Id_S), where (i, j) is in Id_S iff $i = j$ and h_i is in S. The **normal bistring** $I(h)$ corresponding to a string h is simply its projector with respect to the full alphabet: $I(h) = P_T(h)$. A **projection** of a string with respect to some subalphabet S can now be defined as the alignment of the corresponding normal bistring with the projector.

The alignment of well-formed bistrings is not necessarily well-formed, as the following example shows. Let f $= ab$, g $= c$, h $= de$, and suppose that the following associations hold: $(0, 0)$ and $(1, 0)$ in x; $(0, 0)$ and $(0, 1)$ in y. By definition, C should contain $(0, 0), (0, 1), (1, 0)$, and $(1, 1)$ and will thus violate the No Crossing Constraint. Note also that a projector, as defined here, will not necessarily be proper. In order to capture the phonologically relevant sense of properness, it is useful to relativize the definition above to 'P-bearing units' (Clements and Ford 1979). We will say that a bistring (f, h, A) is **proper with respect to a subset** S of the tier alphabet T underlying the string h, iff $(f, h, A) \parallel P_S(h)$ is proper.