K + K = 120: Papers dedicated to L. Kálmán & A. Kornai on the occasion of their 60th birthdays, pp. 445–459

Intuition and decidability in grammar and number theory

GEOFFREY K. PULLUM
University of Edinburgh
Geoffrey.Pullum@ed.ac.uk

KEYWORDS

grammar
prepositions
decidability
algorithm
arithmetic

ABSTRACT

Could a grammatical English sentence contain three consecutive strictly transitive prepositions? One might easily think not: strictly transitive prepositions require NP complements. However, prepositions can be stranded, clausal constituents can begin with prepositions, and so on. Ideally one would like such questions to be algorithmically decidable. I examine the theoretical issue, note a parallel in number theory, reveal the solution to the empirical puzzle (but not the number-theoretic one), and conclude by noting that there is indeed an algorithm for deciding whether some sequence can appear as a proper subsequence of a grammatical string, provided English is context-free.

1. A puzzle in English grammar

Some English prepositions can be used either with a noun-phrase complement or without it if the context permits:

(1) a. Open the gate and walk through it.
b. Open the gate and walk through.

(2) a. After a while we went back inside the house.
b. After a while we went back inside.

Others, like *of* and *into*, strictly require a noun-phrase complement and cannot be used grammatically without one:

(3) a. She asked for dihydrocodeine, but I had never heard of it.
b. *She asked for dihydrocodeine, but I had never heard of.

(4) a. We went back to the car and got into it.
b. *We went back to the car and got into.

I'll refer to the latter kind of prepositions as *strictly transitive*. Consider now the question stated in (5).

(5) Is there a grammatical English sentence containing a sequence of three consecutive occurrences of a single strictly transitive preposition?

You might think that this question can be immediately answered in the negative on the grounds that the next word after a strictly transitive preposition would have to be the beginning of a noun phrase and noun phrases do not begin with prepositions. But not so fast: English syntax is much trickier than that. For one thing, the complement of a preposition does not have to immediately follow it, but can be displaced in at least three ways.

First, there are prepositional passives like (6b), where what is understood as the complement of a preposition (as in (6a)) is in grammatical terms the subject of the clause, and thus separated from the preposition:

(6) a. People seldom speak of this.
b. This is seldom spoken of.

Grammarians speak of such prepositions as *stranded*. Because of what are often called "subject raising" constructions, the subject can appear arbitrarily far away from the stranded preposition:

(7) This seems to have turned out under the circumstances to have been only very seldom spoken of.

Second, items like interrogative or relative words or phrases can be displaced an arbitrary distance to the beginning of a clause. This is the most common way in which a preposition can be stranded. Notice that these two sentences are both grammatical, and are synonymous (though while the first is normal in style, the second is rather formal and pompous):

(8) a. Which regulation do you think the committee imagines the provost's action might have been in violation of?
b. Of which regulation do you think the committee imagines the provost's action might have been in violation?

In both of these the preposition *of* is understood to have *which regulation* as its complement. Notice that in (8b) the construction involved allows the strictly transitive preposition *of* to be the first word in its clause.

Third, phrases can also be displaced toward the end of the clause, yielding a different way in which a strictly transitive preposition may fail

to be immediately followed by its complement, as when a parenthetical is inserted after a preposition:

(9) It was a painting of, or perhaps I should say a painting apparently intended to vaguely suggest, a cornfield in summer.

Thus in order to settle the question in (5) it will be necessary to ensure that no facts of this sort can interact to create a way in which three transitive prepositions could become adjacent. It is not just a matter of which words are allowed to be adjacent to which other words: the interactions of the many different syntactic constructions in English are not necessarily going to be easy to foresee.

2. Generative grammars and decidability

What (5) asks is whether some combination of grammatical configurations can permit a sentence to contain a sequence like *of of of* or *into into into*. It seems intuitively unlikely. But can we prove that it is impossible?

There are systematic computational ways of answering some kinds of questions about sentences in languages. The great majority of the relevant work has been based on systems of rules that Post (1943) originally called production systems, and computer scientists often call rewriting systems, and linguists call generative grammars. Basically they are sets of rules for nondeterministic random construction of abstract structures such as strings or trees.

Post's systems were developed for the purpose of formalizing rules of inference in logic, and were very elaborate, allowing for conclusions to be derived from arbitrary-sized finite sets of premises of arbitrary complexity. As soon as the concept of recursively enumerable (r.e.) sets was clearly formulated in the 1930s, it was clear to Post that any r.e. set of strings could be generated by one of his "canonical production systems". In 1943 he proved that this remained true for dramatically restricted systems that he called "normal" systems, in which every rule had the form "$yX \Rightarrow Xz$" for specified strings y and z, where X is a free string-valued variable, and in 1947 he proved the same for another special case, where the rules all have the form "$X_1yX_2 \Rightarrow X_1zX_2$", for specified strings y and z and variables X_1 and X_2. This, of course, is exactly the form of the grammars that Chomsky (1959) later called "type 0".

For any form of grammar that has this kind of expressive power (i.e., that can generate any arbitrary r.e. set), questions of the form "Does

grammar G generate any string containing the substring w?" are always going to be undecidable. This follows from Rice's theorem (Hopcroft & Ullman 1979, 185–192) as applied to generative grammars rather than Turing machines. All non-trivial properties of r.e. sets (that is, properties that hold of some r.e. sets but not all) are undecidable.

I suspect it will also hold for the restricted class called *context-sensitive* languages (equivalent to the class of type 0 grammars in which the z is always at least as long as the y), though the conjecture needs a proof. The rationale for my conjecture is that nearly all decision problems asking for a property of an arbitrary context-sensitive stringset L are undecidable, the two exceptions being membership ("Is w in L?"), which is decidable, and complement type ("Is $\overline{L}$ context-sensitive?"), which was proved in 1987, surprisingly, to be trivial in the sense of having a positive answer for every context-sensitive L (this follows from the Immerman–Szelpcsényi theorem; see Immerman 1999, 149–151). Despite the decidability of membership, context-sensitive languages are extremely similar to arbitrary c.e. sets, and have essentially all of their complexity. Any type 0 grammar G over a symbol inventory Σ can be converted into a context-sensitive grammar G' over $\Sigma \cup b$ where b is a new dummy symbol used for padding the ends of rules in which the right hand side is shorter than the left hand side. $L(G)$ is then obtainable from $L(G')$ simply by ignoring b (where "ignoring b" means applying a homomorphism that erases b).

The most interesting family of stringsets for purposes of studying the properties of human languages is the much smaller subset known as the *context-free* stringsets (standardly called CFLs). This deserves closer attention. CFLs are generated by context-free grammars (CFGs). It is by no means implausible that the set of grammatical sentences of English could be exactly generated by a CFG: see Pullum & Gazdar (1982), Pullum (1985), and Pullum & Rawlins (2007) for discussion of some failed counterarguments.

CFGs are far more tractable than context-sensitive grammars in most respects. But even for a CFG, it is not always immediately obvious what it can do. As a very simple example, consider the CFG with terminals $\{a, b\}$, nonterminals $\{S, A, B\}$, start symbol S, and the rules shown in (10).

(10)
$S \to aB$	$A \to bAA$
$S \to bA$	$B \to b$
$A \to a$	$B \to bS$
$A \to aS$	$B \to aBB$

The strings that this grammar generates are jumbles of a's and b's in arbitrary orders, but they all meet a special condition: the number of a's is exactly the same as the number of b's. The grammar in (10) generates all and only the strings meeting that condition. This could hardly be said to be immediately evident from looking at the rules, but it can be proved inductively (see Hopcroft & Ullman 1979, 81–82).

It follows that the grammar definitely allows for the construction of a sentence with *aaa* in it, and that any such string will also contain at least three instances of b, and so on. Indeed, we know for any arbitrary string of a's and b's that the answer to whether (10) can generate it is yes. However, that is specific to (10), and depends on the proof concerning what stringset it generates. Can we decide such questions more generally?

After all, although membership ("Can the string w be generated by the CFG G?") is decidable, and so is emptiness ("Are any strings at all generated by the CFG G?"), many other semantic questions about CFGs (i.e., questions not about their form but about their meaning, in the sense of what strings they can generate under their usual interpretation) are formally undecidable. These include *intersection emptiness* ("Does CFL L_1 have a non-empty intersection with CFL L_2?"), *stringset inclusion* ("Are all the sentences of CFL L_1 included among the sentences of CFL L_2?"), *regularity* ("Is CFL L accepted by a finite-state automaton?"), (see Hopcroft & Ullman 1979, 281), and many others.

The set of all strings of English words (whether grammatical or not) in which the substring *of of of* appears is clearly regular (finite-state), assuming only that English has a finite vocabulary V of words.[1] The finite automaton accepting the set remains always in its start state q_0, checking only that each word is in V, and always rejects, except that if it encounters an *of* it switches to q_1, and if another one immediately follows that it goes into q_2, and if another immediately follows that it goes into q_3. Once in q_3 it always accepts provided only that all subsequent words are in V. We seek a general algorithm for finding out whether some specific CFL has a non-null intersection with that regular set. But the general question of whether two stringsets have a non-null intersection is undecidable.

That is not in contradiction with what was said above about the rules in (10) and the set of strings containing *aaa*. There we had a specific CFG

[1] Kornai (2002) gives an interesting argument against this assumption, based on empirical facts about statistical properties of text: English text exhibits properties that are best modelled in terms of an infinite word stock. But for the sake of the present argument we continue under the usual formal language theory assumption of a finite terminal vocabulary.

for which it happened to be possible to construct a proof that all the generated strings had equinumerous *a*s and *b*s, and that some generated strings contained *aaa*. This shows that in certain special cases we may find out the answer. That does not give us a general algorithm for all cases.

I will return later to the question of whether, given a complete generative grammar for English, there would be a systematic general way of using it to guarantee answers to questions like (5). But first I want to note an interesting similarity to a question in mathematics.

3. A parallel in number theory

Question (5) has a particular logical property in common with the question in (11), which derives from a famous conjecture in number theory.

(11) Is there an even number that is not equal to the sum of two primes?

This can be easily stated using first-order logic interpreted in the usual number-theory model where the domain is the non-negative integers with the operations "+" (addition) and "·" (multiplication). The predicate "even" can be defined as in (12a); "prime" can be defined as in (12b); and then (11) is the question of whether (12c) is true in the specified model.

(12) a. $\mathbf{even}(x) =_{\text{df}} (\exists y[y \geq 1 \wedge y \cdot 2 = x])$

b. $\mathbf{prime}(x) =_{\text{df}} (\neg(\exists y \exists z[y \geq 2 \wedge z \geq 2 \wedge y \cdot z = x]))$

c. $\exists x[\mathbf{even}(x) \wedge \neg(\exists y \exists z[\mathbf{prime}(y) \wedge \mathbf{prime}(z) \wedge y + z = x])]$

What (11) is in effect asking for is a counterexample to the strong Goldbach conjecture, henceforth GC, which claims that every even number greater than 2 is the sum of a pair of primes. Most number theorists are inclined to think this conjecture is true. One reason is that as we consider larger and larger even integers n, the number of different pairs of primes that sum to n increases, so that for any large n it is overwhelmingly likely that there is at least one pair that sums to n. But GC is a non-probabilistic claim, and as is well known, no proof of it has been found, so currently it cannot be guaranteed that the answer to (11) is negative.

The logically interesting property that (5) and (11) share is that for each of them, showing that the question is algorithmically undecidable would ipso facto (though indirectly) tell us the answer. This sounds paradoxical, but it is not. It is a fairly simple point, fairly well known among number theorists and mathematical logicians.

Consider (11) first. To say that the answer to (11) cannot be discovered by an algorithm would mean that GC is unprovable within our system for proving things in arithmetic. And we know from Gödel (1931), of course, that some truths of arithmetic are unprovable in any system capable of expressing all arithmetical truths.

For concreteness, assume the standard system defined by the Peano axioms, known as PA, and a monadic second-order logic interpreted on $\langle \mathbb{N}, +, \cdot \rangle$ (the natural numbers with addition and multiplication). To say that GC is incapable of proof within PA is in effect to say that GC is *independent* of PA, in the sense that we could add GC to the set of PA's consequences, or add its negation, without losing consistency either way. You could believe all the truths of PA plus GC, or believe all the truths of PA plus the negation of GC, and no one would be able to use PA to prove you wrong either way.

Yet if GC were shown to be independent of PA, we would immediately know whether it was true or not: it would have to be true, so the answer to (11) would be negative. Here is the reasoning.

If GC were false, there would be a counterexample, an even number that cannot be expressed as the sum of any two primes. Let g be that number. We could demonstrate GC's falsity in an elementary way by simply exhibiting the list of all triples $\langle p_1, p_2, k \rangle$ such that p_1 and p_2 are primes and $k \leq g$ and $p_1 + p_2 = k$. The list might be very long, if g were very large, but it would be finite, and could be constructed by a very straightforward computer program. The absence from the list of any case where $k = g$ would falsify GC, and thus answer (11) in the affirmative.

If (11) cannot be answered in the affirmative by a proof, the answer to it must be negative, i.e., GC must be true. The answer to (11) cannot be positive yet unprovably so.

An analogous result holds for (5). If we found some way, using facts about a generative grammar for English, to show that (5) cannot be answered within some system of mathematical reasoning, then we would immediately (but indirectly) know that the answer to (5) is negative, because otherwise there would exist a sentence containing three consecutive occurrences of a single transitive preposition, and simply exhibiting the derivation of that sentence would settle the question, offering a proof of the positive answer. The answer to (5) cannot be positive yet unprovably so.

4. The answer to the grammatical puzzle

A key difference between question (5) and question (11) is that (5) is in a sense empirical. It is an empirical fact that those who describe themselves as speakers of English invariably regard *All cows eat grass* as grammatical but **Grass eat cows all* as ungrammatical; they regard *Never have I heard such nonsense* as grammatical but **Never I have heard such nonsense* as ungrammatical; and so on. If there is a sentence containing three instances of some transitive preposition in a row that English speakers treat as grammatical when it is presented to them, then that is an empirical fact (subject to all the usual epistemological caveats, to be discussed very shortly).

My guess, on the basis of 40 years' experience of working on English syntax and techniques for formalizing syntactic theories, and six years working with Rodney Huddleston on the largest and most complete reference grammar currently available for English (Huddleston & Pullum 2002), would have been that the answer to (5) was negative: I would have thought that the rules of English grammar could not allow three consecutive occurrences of a strictly transitive preposition, on the grounds that there wouldn't appear to be any context in which all three of them could have the obligatory noun-phrase complements they require.

But it is a very important fact about argument and evidence in syntax that the intuition of a grammarian regarding generalizations of this sort cannot be trusted.

It is true that the intuition of a native speaker (whether a grammarian or not) can generally be trusted on individual sentences. This is why determining grammatical well-formedness for a sentence of reasonable length normally involves little more than having a native speaker look at it or listen to it, provided some minimal conditions of attentiveness are respected. But caveats are necessary even to that claim, because aspects of meaning, style, phonology, or processing may interfere with intuitive judgments about sentencehood. For example, (13a) will generally be judged grammatical, but the synonymous (13b) will not.

(13) a. Everybody whom everybody left departed.

b. Everybody everybody left left.

This apparently because center-embedding a phrase or clause inside another, even once increases the processing load substantially. (Notice that the relative clause everybody left is embedded with parts of the main clause either side of it, which means processing of the main clause must be

interrupted by the processing of another clause and then resumed where it left off; this is discussed in Miller & Chomsky (1963) and much subsequent psycholinguistic literature.) The combination of that with two pairs of adjacent duplicate words is confusing enough to completely wreck the chances of recognizing the grammatical structure.

Likewise, it is well known that there are sentences that confuse us into thinking they are ungrammatical by (as it were) tempting us to process them incorrectly. They are known as *garden-path sentences* (Bever 1970). One celebrated example, well known from the psycholinguistic literature, is (14):

(14) The horse raced past the barn fell.

Our tendency to process this with *raced* as the preterite-tense verb of the main clause, and an unneeded extra verb *fell* on the end, is almost irresistible, and blinds us to the fact that *raced* can also be a past participle, so *raced past the barn* could be a nonfinite passive clause modifying *horse*. In other words, the sentence can be read with the same structure as (15):

(15) The car driven past the barn crashed.

Many other similar examples could be given of the ways in which poor acceptability may wrongly make a properly-formed sentence seem ungrammatical.

However, even if native speakers can in typical cases intuitively perceive the well-formed structure of an individual sentence, even skilled syntacticians cannot reliably intuit the truth of generalizations about wide ranges of sentences or phrases.

The young Noam Chomsky ventured in a conference discussion the assertion that "The verb *perform* cannot be used with mass-word objects: one can perform *a task*, but one cannot perform *labor*" (Hill 1962, 29). Challenged by another participant (Anna Granville Hatcher) to say how he knew this, he answered: "Because I am a native speaker of the English Language". Later in the discussion Hatcher asked him what he would say if the non-count noun were *magic*, and Chomsky was immediately forced to confess: "I think I would have to say that my generalization was wrong" (*ibid.*, 31).

On the specific point of whether a short sentence like *They can perform magic* is grammatical, or whether a short string of words like **They can perform justice* is ungrammatical, he could supply reliable intuition

reports, like most native speakers; but confirming a broader generalization about English sentence structure is a very different matter.

And to return to the case at hand, judging whether three consecutive transitive prepositions is possible in English is a judgment concerning an indefinitely large range of sentences. I would have hazarded the guess that the answer was negative, but I would have been wrong. The answer to question (5) is now known, thanks to Wells Hansen (personal communication), and it is positive. Hansen showed this by constructing and exhibiting, rather surprisingly, a grammatical sentence with an *at at at* sequence. A similar one is given in (16).

(16) Donald Trump was laughed at at at least three dinner parties in Manhattan this year.

It is fully grammatical (as well as probably also true), and surprisingly simple to understand. Of course, it might be impugned for style: a writer who notices that some word has been used three times within a short space, or that a jingle effect has been created by two or three words with a similar sound will generally reword. But that is about stylistic acceptability, not grammaticality.

Retrospectively, we can see why and how the Hansen sentence has to be regarded as grammatical. *At* occurs as a grammaticized preposition syntactically required in the construction *laugh at*:

(17) a. They laughed at him.
b. *They laughed to him.
c. *They laughed by him
d. *They laughed on him.

And the choice of preposition is determined by the choice of verb; other verbs require different prepositions:

(18) a. *We spoke at him. (*speak* does not take *at*)
b. We spoke to him. (*speak* does take *to*)
c. *We spoke on him. (*speak* does not take *on*)

(19) a. *They rely at him. (*rely* does not take *at*)
b. *They rely to him. (*rely* does not take *to*)
c. They rely on him. (*rely* does take *on*)

Verb-preposition combinations of this sort readily yield prepositional passives (*was laughed at*, *was spoken to*, *was relied on*, etc.). Hansen's sentence has the form of a prepositional passive clause, with the first *at* of the sequence as its stranded preposition. The subject of the clause (*Donald Trump*) is understood as the complement of the first *at*.

But *at* is somewhat more syntactically versatile than *of* in one respect: it can also serve as the head of a locative modifier like *at three parties*, which can occur following a prepositional passive; and it occurs in the idiomatic Preposition + Adjective combination *at least*, which can serve as an adjunct modifying a determinative like *three*, hence, crucially, can stand at the beginning of a noun phrase, as in *at least three parties*, and thus can begin a noun phrase serving as the complement of the preposition *at*.

Thus when the first *at* is stranded in a prepositional passive construction it is possible for a second *at*-phrase heading a locative adjunct to follow, and for a third *at*-phrase to begin the noun phrase within that locative adjunct. All those facts are relevant to why it is that *at at at* can be a possible subsequence in a grammatical sentence.

5. Proper substring possibility for CFLs

We should never forget that English syntax constitutes a vast domain of exploration, within which are many known unknowns, and an unknown number of unknown unknowns. This domain cannot be explored via the simplistic appeals to "logic" that purists and usage advisers so often advocate. Which sentences are grammatical is not determined by any kind of common-sense or formal logic. The grammatical sentences are simply the ones that happen to be permissible under the set of rules or constraints that defines the language – the large set of exception-ridden and often rather quirky rules that define English as it happens to be today. Discovering how we are to precisely formulate the content of those rules is a major scientific enterprise. Even an informal survey of the ground that must be covered takes up more than 1,700 pages of text (Huddleston & Pullum 2002, henceforth *CGEL*).

And we cannot blithely assume, even when we have produced such a grammar, that there will exist an algorithm for finding out whether it is possible for a grammatical sentence to meet some given condition, such as having three consecutive transitive prepositions, or containing the sequence *and the of*, or any other syntactically definable property of symbol strings. Indefinitely many precisely framed questions about human languages (considered as stringsets over a word vocabulary) are undecidable,

even given a full, exact, and correct grammar for the language (which even for English, of course, we do not have as yet).

While in general native speakers (whether grammarians or not) have intuitive reactions concerning the grammaticality of specific strings of words presented to them, they do not have intuitional access to the truth values of generalizations about the entire range of sentences that are grammatical in their language, any more than mathematicians have intuitional access to the truth values of generalizations about the integers. The key difference is that we take the truths of number theory to be a priori and necessary, substantiable through rigorous proof as in the other formal sciences, while the true statements about English grammar are at root empirical.

We have qualified intuitional access to the status of specific sentences because we subconsciously respond to them as if we were encountering them in actual use, and to some extent we can report on our responses (see Devitt 2006, chapter 7, for a discussion of this topic that I find very perceptive). But we have no veridical intuitional access to broader generalizations about the grammar of our language, and can be surprised by discovering them.

Questions about whether some word sequence like *at at at* or *of of of* can form a subsequence of a grammatical sentence in some human language, if we assume for concreteness that human languages can be generated by CFGs, have the general form seen in (20):

(20) Proper substring possibility

Given a CFG G with terminal vocabulary V and an arbitrary string w in V^*, does G generate any string that has w as a proper substring?

One might ask whether there could ever be practical reasons for needing answers to such questions. Practical importance is of course something that in general we discover only retrospectively, but it is not impossible to imagine a context in which information about occurrence of substrings might be of practical use to an engineer. A robot equipped to parse and understand spoken English might be designed with a special simple word sequence that would immobilize it to permit servicing (or to block a *Westworld*-style disaster in which robots become malign). To make sure the robot could not be immobilized unintentionally through ordinary conversation, one might want the word sequence to be one that definitely could not form part of any sentence of the language. We know, thanks to Wells Hansen's discovery, that *at at at* could not serve that purpose. Maybe *of of of* could.

So is (20) formally decidable? The answer turns out to be yes. A problem closely related to it was studied by Lang (1988) in the context of devising a context-free parsing algorithm that will yield useful output even when faced with sentences containing unknown parts of unknown length, by producing a finite representation of the set of all possible parses (perhaps infinitely many) that could allow for the missing parts. Subsequently Osorio and Navarro (2001) tackled more directly the problem of solving proper substring possibility as stated in (20), using the CKY algorithm as the basis for their proof and showing that the problem can be decided in cubic time.

Osorio and Navarro point out that the problem actually has many more areas of potential application than you might think, since CFG parsing is so closely related to other computational tasks like matrix multiplication and is so widely applicable: it could be relevant to DNA analysis in bioinformatics, syntax-driven development of language tools, and shape analysis in pattern recognition.

Given a CFG for English, therefore, we could use a fully general algorithm to find out (in cubic time) whether, for example, there is a grammatical string with *of of of* as a substring. (I think there probably is, but I leave the exercise of constructing one for the reader to pursue in idle moments.) Of course, the algorithm presupposes the completion of a CFG that fully and accurately generates all and only the sentences of English. The informal account in *CGEL*, mentioned above, is not expressed in anything like the form of a CFG.

For what it is worth, Pullum and Rogers (2008) provide, in a rather unexpected way, good reason to believe that there is nothing in *CGEL* that is beyond the power of CFGs. They note that although the objects that *CGEL* uses as structural representations are not trees, they are very close to being trees, and the very limited departure from treehood that is employed (downward convergence of branches in certain particular kinds of noun phrase) can be described by a transduction to covering trees expressible in weak monadic second-order logic (wMSO), and wMSO-describable sets of trees always have CFLs as their frontier sets (by the theorem of Doner 1970). Hence *CGEL* appears to covertly entail that English (considered as a stringset over a vocabulary of dictionary words) is a CFL.

In principle, then, there almost certainly exists a CFG for English on which we could run an algorithm of the sort sketched by Osorio and Navarro to find out whether *of of of* (or any other word sequence) can be a substring of a sentence.

Acknowledgements

My thanks to Wells Hansen, whose serendipitous observation of an *at at at* sentence inspired this paper; to Colin Stirling, who discovered that the substring admissibility problem had been shown to be decidable by Osorio and Navarro (2001); to Marcus Kracht (originally an anonymous referee), who suspected that this was true and drafted a different proof from Osorio and Navarro's, later generalizing it to cover all multiple context-free grammars (MCFGs); and to my old friend and research collaborator András Kornai, who has taught me so much (though it will never be enough) about mathematically informed work on human languages.

References

Bever, T. G. 1970. The cognitive basis for linguistic structures. In J. R. Hayes (ed.) Cognition and the development of language. New York: Wiley. 279–362.

Chomsky, N. 1959. On certain formal properties of grammars. Information and Control 2. 137–167.

Devitt, M. 2006. Ignorance of language. Oxford: Clarendon.

Doner, J. 1970. Tree acceptors and some of their applications. Journal of Computer and System Sciences 4. 406–451.

Gödel, K. 1931. Über formal unentscheidbare Sätze der *Principia mathematica* und verwandter Systeme I. Monatshefte für Mathematik und Physik 38. 173–198. (English translation in *Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*, ed. by Jean van Heijenoort, Harvard University Press, Cambridge MA, 1967, 596–616.)

Hill, A. A. 1962. Third Texas Conference on Problems of Linguistic Analysis in English. Austin, TX: University of Texas.

Hopcroft, J. E. and J. D. Ullman. 1979. Introduction to automata theory, languages and computation. Reading, MA: Addison-Wesley.

Huddleston, R. and G. K. Pullum. 2002. The Cambridge grammar of the English language. Cambridge: Cambridge University Press.

Immerman, N. 1999. Descriptive complexity. New York: Springer.

Kornai, A. 2002. How many words are there? Glottometrics 4. 61–86.

Lang, B. 1988. Parsing incomplete sentences. In Proceedings of the 12th Conference on Computational Linguistics, Volume 1, COLING '88. Stroudsburg, PA: Association for Computational Linguistics, 365–371.

Miller, G. A. and N. Chomsky. 1963. Finitary models of language users. In R. D. Luce, R. R. Bush and E. Galanter (eds.) Handbook of mathematical psychology, vol. 2. New York: Wiley. 419–491.

Osorio, M. and J. A. Navarro. 2001. Decision problem of substrings in context free languages. In CIC-X: Memorias del X Congreso Internacional de Computación. 239–249.

Post, E. L. 1943. Formal reductions of the general combinatory decision problem. American Journal of Mathematics 65. 197–215.

Post, E. L. 1947. Recursive unsolvability of a problem of Thue. Journal of Symbolic Logic 12. 1–11.

Pullum, G. K. 1985. On two recent attempts to show that English is not a CFL. Computational Linguistics 10. 182–186.

Pullum, G. K. and G. Gazdar. 1982. Natural languages and context-free languages. Linguistics and Philosophy 4. 471–504.

Pullum, G. K. and K. Rawlins. 2007. Argument or no argument? Linguistics and Philosophy 30. 277–287.

Pullum, G. K. and J. Rogers. 2008. Expressive power of the syntactic theory implicit in *The Cambridge grammar of the English language*. Paper presented at the 49th Annual Meeting of the Linguistics Association of Great Britain, University of Essex.