The counterrevolutionary

András Kornai
Dept. of Algebra, Budapest University of Technology and Economics
H-1111 Budapest Egry J. u. 1
kornai@math.bme.hu

***Abstract*—In this paper we offer a systematic overview of Chomsky's early work on formalizing linguistic theory. In 20/20 hindsight it is easy to see that this seminal work was not only one-sided, concentrating entirely on the discrete aspect of language and linguistic theory, but actively hostile to the continuous aspect, especially to the use of frequency data.**

Statistical thinking entered physics with the work of Maxwell, Boltzmann, and Gibbs on thermodynamics in the second half of the 19th century, and with the advent of quantum theory it began to pervade all aspects of physics. In the same period actuarial science, evolutionary theory, epidemiology (and medicine in general), econometrics, psychology, and much of social science were all put on a statistical foundation. Ideas from statistics and probability theory have penetrated our scientific thinking to such an extent that

> shortly after 1930 it became virtually certain that at bottom our world is run at best by laws of chance. In epistemology [...] it is now a commonplace that much of our learning from experience, and much our foundations for knowledge, are to be represented by probabilistic models. (Hacking, 1987)

Against this general background it is nothing short of remarkable that in the 1950s and 1960s a brilliant, and in many ways radical and revolutionary thinker, the young Noam Chomsky, tried to singlehandedly stem the tide and make linguistics the last stand of deterministic theorizing. In this, he had notable success, in that generations of linguists grew up accepting his arguments against the use of probabilistic models. Here we are less concerned with why he did this (other notable anti-probabilists include Einstein, who fought a losing battle against quantum mechanics until his death in 1955) and concentrate on the substance of his arguments, put forth in a remarkable set of papers and books in the late 1950s and early 1960s.

The beginnings are already present in his PhD thesis (1954-1955) published in shortened form as Chomsky (1975), and in some tech reports that appeared in the MIT Research Laboratory of Electronics *Quarterly Progress Report* (nos. 41 and 42, 1956), but we see no reason to discuss these fully here, since Chomsky's priority in these matters is undisputed, and it was only the subsequent, more widely accessible publications, which repeated and extended these arguments, that had broad impact. (According to Google Scholar, Chomsky's first journal publication on the subject (Chomsky, 1956) has 3212 citations, whereas the tech reports have a total of 12.)

As we shall see, most of his arguments rest on simple, but common misunderstandings regarding the nature and evidential value of putatively infinite sequences, while the rest are attacks on formal tools that were, in critical respects, immature at the time.

I. The fight against information theory

The probabilistic theory of communication was initiated by Shannon (1948). If anything, the famous entropy formula $H = -\sum_i p_i \log_2 p_i$ is even more a part of the formula-repository that undergirds 21st century intellectual developments than $E = mc^2$, for bits and bytes have become part of everyday life, on a par with volts and ampers, whereas rads and grays, thankfully, have not. Remarkably, the mathematical underpinnings of the theory are accessible to all highschoolers with a firm grasp of logarithms, and were commonly taught at MIT in the 1950s. As it happens, the leading local authority on the subject, Robert Fano, a coauthor of Shannon, whose textbook (Fano, 1961) was widely used until Cover and Thomas (1991) replaced it, actually worked in the same Building 20 as Chomsky for decades. Altogether, Chomsky clearly had the means and the opportunity to fully grasp the subject, there is simply no way to chalk up to blind ignorance what he did, which was to run a full-bore disinformation campaign against information theory.

We start with a typical example from his review of Greenberg (1957). In footnote 2 Chomsky (1959) writes:

> The assumption is that at each point in the sentence there are a certain number of remaining possibilities as to what the sentence

> will be, and that choice of an element out of a larger class reduces this number more severely than choice from a smaller set, thus giving more "information" to the listener as to which sentence is being selected. However, this interpretation is unacceptable as it stands. Since, by and large, there are, at a given point in a sentence, an infinite number of possible sequences which will complete this sentence, choice of a particular element does not reduce the *number* of remaining alternatives at all.

The plain fact of the matter is that Greenberg was right and Chomsky was wrong, since it is not the raw number, but the information that is at stake here. A sentence that begins *Hey have you heard that* and one that begins *There are many trivial ways* both generate a probability distribution on the set of words that can come next, and the entropy of the former, 4.6 bits, is significantly higher than the entropy of the latter distribution, 2.9 bits. Using a simple statistical language model trained on the Penn Treebank, the word predicted as most probable after the first five words is *the, p*=0.27 in the first case, *to, p*=0.35 in the second, while *to* has probability 0.006 in the first case and *the* 0.012 in the second. Six decades have passed, and the information-theoretic approach has proven its worth in many areas from speech recognition to machine translation, while the idea of basing a calculus on raw counts (without probabilistic weighting) remains elusive to this day.

At this point, it is hard to say what is driving Chomsky, wilful ignorance (hard to countenance) or conscious choice, but as time goes by, the arguments get more explicit, so we will assume the latter. The highly cited (Chomsky, 1956) takes direct issue with the simple *n-gram* model (section 2.4):

> [W]e might inquire into the possibility of constructing a sequence of [finite state grammars] that, in some nontrivial way, come closer and closer to matching the output of a satisfactory English grammar. Suppose, for example, that for fixed n we construct a finite state grammar in the following manner: one state of the grammar is associated with each sequence of English words of length n and the probability that the word X will be produced when the system is in the state S_i is equal to the conditional probability of X, given the sequence of n words which defines S_i. The output of such grammar is customarily called an n+1st order approximation to English. Evidently, as n increases, the output of such grammars will come to look more and more like English, since longer and longer sequences have a high probability of being taken directly from the sample of English in which the probabilities were determined. This fact has occasionally led to the suggestion that a theory of linguistic structure might be fashioned on such a model. Whatever the other interest of statistical approximation in this sense may be, it is clear that it can shed no light on the problems of grammar. There is no general relation between the frequency of a string (or its component parts) and its grammaticalness. We can see this moat clearly by considering such strings as *colorless green ideas sleep furiously* which is a grammatical sentence, even though it is fair to assume that no pair of its words may ever have occurred together in the past ...

At the time, few linguists had the statistical background required for spotting the fallacy, namely, that the probability of a string is not equal to its frequency in a corpus, however large, since the corpus is just a finite sample from an infinite distribution. While a detailed critique specifically aimed at the *colorless green ideas* argument had to wait until Pereira (2000), the speech (Jelinek and Mercer, 1980) and character recognition (Kornai, 1994) communities kept up the "other interest" in obtaining estimates of the probability of the next word, a task commonly referred to as statistical *language modeling*. In these communities, and in the world of information retrieval (Ponte and Croft, 1998) and machine translation (Brown et al., 1990) it was well understood from the beginning that zero estimates are not acceptable even if zero frequencies are observed, the distribution needs to be *smoothed*. As a matter of historical interest, we mention here that the first widespread smoothing method was in great part due to Fred Jelinek, whose doctoral advisor was Robert Fano.

Let us now consider Chomsky and Miller (1958), one of the very appealing papers in the flagship first volume of *Information and Control*, which clearly helped establish *I&C* as the preeminent journal of formal language theory. Among other results, the paper contains the first systematic investigation of finite state languages (FSLs), proving several theorems that are now standard parts of the formal language curriculum, e.g. that the family of FSLs is closed under Boolean operations. Finite state automata (FSA) have received a great deal of attention at the time, with the definitive reformulation of the McCulloch and Pitts (1943) brain model as FSA (Kleene, 1956) occupying place of honor in the influen-

tial Shannon and McCarthy (1956) collection of papers, the same collection where the equivalent Moore machines (Moore, 1956) were introduced. Mealy machines, another equivalent formulation, were introduced the year before, in Mealy (1955). These were all deterministic models: the key observation, that nondeterminism (free will) makes no difference in the class of languages accepted/generated was eventually proven in Rabin and Scott (1959). How Chomsky and Miller (1958) fits in the tenor of Chomsky's disinformation campaign is clear:

> In recent years the representation of a message source by a stochastic process, usually a finite Markov process, has become a familiar procedure (Shannon, 1948). A finite Markov process is essentially a finite state generator that is supplemented by probability distributions for the choices available in each state. There are two important questions we could ask about such generators: (1) What properties characterize languages produced by such generators, and (2) Do natural languages have these properties? It seems clear that a finite state grammar is not adequate for most natural languages (Chomsky, 1956), so the answer to the second question is negative. Nevertheless, the mathematical properties of such processes are well suited to the needs of communication engineers and it is likely that they will continue to be of interest in many applications of communication theory.

Since the leading brain model at the time was equivalent to FSA, the logical conclusion was that natural languages are finite state. Chomsky (1957) devoted special effort to demonstrating, with the aid of center-embedding constructions, that this was not so, and subsequently added further epicycles to the theory, most notably, extending the scope of the otherwise clearly necessary competence/performance distinction (Chomsky, 1965) to center-embedding just to defend the validity of this demonstration. This extension of scope was rather controversial from the beginning (and it is hard to find anybody outside the ever-narrowing group of Chomsky devotees who still accepts it), but we shall leave this well-documented controversy to the side here, concentrating entirely on the probabilistic aspects.

What we see here is a systematic effort to present Markovian systems as FSA "supplemented" by probability distributions, and then study FSA without the probabilistic aspect Chomsky deems supplementary. The last sections of Chomsky and Miller (1958), where the number of sentences of length λ is counted, already display this strategy of making frequencies disappear by a clever shell game. First arbitrary sentences are replaced by sentences with explicit end-markers. This is without loss of generality according to Theorem 1, which deals only with stringsets shorn of frequencies. Next finite state grammars are replaced by unambiguous ones, again without loss of generality according to Theorem 3 as long as boundary markers are present. Once the switcharoo is complete, the methods of (Shannon, 1948) are used to count the strings of length λ. That the whole shell game is useless, because there are probabilistic languages with finite state support that cannot be generated by FSGs "supplemented" by probabilities was proven only a decade later, by Ellis (1969). A remarkable piece of rearguard action was performed by Suppes (1970) and canonicalized in Levelt (1974) who simply presented Suppes' false conjecture as a theorem, see Kornai (2011) for details.

II. What happened?

What we have documented here is not an error. There is simply no way for anyone familiar with the body of Chomsky's work on formal grammars (a field that, for all intents and purposes, he originated) to bring themselves to believe that he somehow committed a 'thinko', a mental misstep akin to a typo. In hindsight it is evident that Chomsky was within striking distance of some of the greatest discoveries of the 20th century, nondeterminism (Rabin and Scott, 1959), and Hidden Markov Models (HMMs) (Stratonovich, 1960; Baum and Petrie, 1966). He was occupied with the same issues, and the extensive (at 70 pages, near monograph-length) treatment of probabilistic models presented in Miller and Chomsky (1963) is ample testament to his mathematical sophistication. But in his eagerness to present his notion of transformational grammar as the only adequate model of natural language (an attempt destroyed by Peters and Ritchie (1973) long before the rebirth of probabilistic models) he missed out both on nondeterminism, which is highly relevant to Artificial Intelligence in general and to programming language semantics in particular (Floyd, 1967), and as we will try to show here, on HMMs, which have far-reaching uses not just for the study of natural language, but also for the language of nature, the genetic code (Durbin et al., 1998).

It is hard to speculate about the mind-state of others, but there is no deep reason to assume that the disinformation campaign, such as it was, was run cynically. More likely is that Chomsky thought hard about these matters, and first convinced himself that probabilities are irrelevant to grammar. One glimpse into his thinking process is provided in Section 2.4 of (Chomsky, 1956)

that we already quoted: 'colorless green ideas sleep furiously' is (...) grammatical (...), even though it is fair to assume that *no pair of its words may ever have occurred together in the past* (emphasis added). Obviously, the bigrams 'colorless green', 'green ideas', 'ideas sleep', and 'sleep furiously' are all extremely rare.

Now that we have access to really large corpora (subsequent search results are from Google Books) we do find an instance of 'colorless green' in Volume 8 of the *School of Mines Quarterly* (1887) in an article devoted to qualitative blowpipe analysis (p. 363), and 'Groovy Green Ideas: Environmental Education for the Kids' effortlessly transcends what at first blush looks a very strong selectional restriction ruling out the use of color terms with abstract objects. While 20th century material is dominated by Chomsky's example, in an 1868 book we find 'lain hushed in that mysterious dormitory, where ideas sleep, all ready to awake again into life' and in a 1932 volume we find 'After a day in the ditches we would sleep furiously'. This is a situation every generative grammarian who relies on intuition is familiar with: we may swear something is ungrammatical, until we are confronted with real life examples of it. Trusting our intuition is good, but verifying it is better.

The examples drive home not just the trivial point that 'unattested' is not a reasonable proxy for 'ungrammatical', but also show how Chomsky got within an inch of the Hidden Markov breakthrough. He knew that the class bigrams, Adjective Adjective, Adjective Noun, Noun Verb, and Verb Adverb are perfectly common, and had the critical insight that the sentence is actually grammatical. All that was missing was to bring his very own notion of distinguishing between deep and surface structure to the case of preterminals (part of speech categories) and lexical entries!

On the occasion of his 90th birthday a celebration of Chomsky's lasting accomplishments would perhaps be more fitting than the impartial analysis, sine ira et studio, of a failure that lesser intellects could have committed even more easily. But the laudatory literature is vast, and almost always commits the very same errors as the master, which brings up the issue of collective responsibility. Chomsky, by missing out on great discoveries within his reach, has already paid dearly for his counterrevolutionary stance, the rest of the blame must be laid at the feet of his blind followers. Instead of becoming a vital part of 21st century research, by the 1990s "mainstream" generative syntax increasingly resembled a rust-belt town, with the local industries dead or dying, old transportation arteries falling into disuse, and no hope for urban renewal. The revolution could not be stopped.

Acknowledgments

I am grateful to Paul Kiparsky for incisive comments on an earlier version of the paper.

References

Baum, Leonard E. and Ted Petrie (1966). "Statistical inference for probabilistic functions of finite state Markov chains". In: *Annals of Mathematical Statistics* 37, pp. 1554–1563.

Brown, Peter et al. (1990). "A statistical approach to machine translation". In: *Computational Linguistics* 16, pp. 79–85.

Chomsky, Noam (1956). "Three models for the description of language". In: *IRE Transactions on Information Theory* 2, pp. 113–124.

– (1957). *Syntactic Structures*. The Hague: Mouton.

– (1959). "On certain formal properties of grammars". In: *Information and Control* 2, pp. 137–167.

– (1965). *Aspects of the Theory of Syntax*. MIT Press.

– (1975). *The logical structure of linguistic theory*. Springer Verlag.

Chomsky, Noam and George Miller (1958). "Finite-state languages". In: *Information and Control* 1, pp. 91–112.

Cover, Thomas M. and Joy A. Thomas (1991). *Elements of Information Theory*. Wiley.

Durbin, R. et al. (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press.

Ellis, Clarence A. (1969). *Probabilistic Languages and Automata*. University of Illinois, Urbana: PhD thesis.

Fano, Robert (1961). *Transmission of Information: A Statistical Theory of Communications*. MIT Press.

Floyd, Robert W (1967). "Nondeterministic algorithms". In: *Journal of the ACM (JACM)* 14.4, pp. 636–644.

Greenberg, Joseph H., ed. (1957). *Essays in linguistics*. University of Chicago Press.

Hacking, Ian (1987). "Was There a Probabilistic Revolution 1800-1930?" In: *The Probabilistic Revolution*. Ed. by L. Krüger et al. Cambridge MA: MIT Press.

Jelinek, Frederick and Robert Mercer (1980). "Interpolated estimation of Markov source parameters from sparse data". In: *Proceedings of the Workshop on Pattern Recognition in Practice*. Ed. by E. S. Geltsema and L. N. Kanal. Amsterdam: North-Holland.

Kleene, Stephen C. (1956). "Representation of events in nerve nets and finite automata". In: *Automata Studies*. Ed. by C. Shannon and J. McCarthy. Princeton University Press, pp. 3–41.

Kornai, András (1994). "Language models: where are the bottlenecks?" In: *AISB Quarterly* 88, pp. 36–40.

– (2011). “Probabilistic grammars and languages”. In: *Journal of Logic, Language, and Information* 20, pp. 317–328.

Levelt, Willem J.M. (1974). *Formal Grammars in Linguistics and Psycholinguistics*. Vol. 1–3. The Hague: Mouton.

McCulloch, W.S. and W. Pitts (1943). “A logical calculus of the ideas immanent in nervous activity”. In: *Bulletin of mathematical biophysics* 5, pp. 115–133.

Mealy, George H. (1955). “A method for synthesizing sequential circuits”. In: *Bell System Technical Journal* 34, pp. 1045–1079.

Miller, George A. and Noam Chomsky (1963). “Finitary models of language users”. In: *Handbook of Mathematical Psychology*. Ed. by R.D. Luce, R.R. Bush, and E. Galanter. Wiley, pp. 419–491.

Moore, E.F. (1956). “Gedanken-experiments on sequential machines”. In: *Automata studies*. Ed. by Shannon and McCarthy. Princeton University Press, pp. 129–153.

Pereira, Fernando (2000). “Formal grammar and information theory: Together again?” In: *Philosophical Transactions of the Royal Society* A 358, pp. 1239–1253.

Peters, Stanley and Robert W. Ritchie (1973). “On the generative power of transformational grammars”. In: *Information Sciences* 6, pp. 49–83.

Ponte, Jay M. and W. Bruce Croft (1998). “A language modeling approach to information retrieval”. In: *Proc SIGIR*. ACM Press, pp. 275–281.

Rabin, M.O. and D. Scott (1959). “Finite automata and their decision problems”. In: *IBM journal of research and development* 3.2, pp. 114–125. ISSN: 0018-8646.

Shannon, Claude E. (1948). “A Mathematical Theory of Communication”. In: *Bell System Technical Journal* 27, pp. 379–423, 623–656.

Shannon, Claude and John McCarthy, eds. (1956). *Automata studies*. Princeton University Press.

Stratonovich, R. L. (1960). “Conditional Markov Processes”. In: *Theory of Probability and Its Applications* 5.2, pp. 156–178.

Suppes, Patrick (1970). “Probabilistic grammars for natural languages”. In: *Synthese* 22, pp. 95–116.