

Interview with
John Goldsmith

Khalil Iskarous (KI)

I'm really happy to be interviewing you, John. I hope that we'll be able to have a discussion of various aspects of your career. And the very first question is, how did you get into linguistics?

John Goldsmith (JG)

I was aware of and interested in linguistics, even when I was in high school. I remember reading an article about Noam Chomsky in *Time* magazine; it was the first time that I had come across his name. He was an up and coming academic that people had been talking about. I graduated from high school in 1968, and he was already publishing (on political subjects) in the *New York Review of Books* and the like, and making quite a name.

When I got to Swarthmore College, there wasn't any linguistics taught there. But I know I was interested in the area. I'd have to admit that my interests were very varied. Mathematics was my main interest, and I was also interested in philosophy and economics.

But in the spring of 1969, I was visiting a friend at Cornell University. There had recently been a sit-in at the Student Center—the Straight—by the African-American students. After that crisis was resolved, the SDS, Students for Democratic Society, had invited Noam Chomsky to come and give a talk.

I was there, more or less by accident. I just walked into the Student Center and saw him there, listened to him for an hour, bought some offprints of his papers from *Ramparts* magazine. I was just totally blown away by this guy. Quite, quite amazing.

Shortly after I got back to Swarthmore, there was an announcement about a new course to be offered in the fall, an introduction to linguistics. I said, wow, that's for me, and signed up for the course. It was taught by Lila Gleitman, who had just gotten her degree from Penn. And that was the beginning of my life in linguistics.¹

KI

Tell me more about your work with Lila Gleitman, inspiration that came at that stage before you went on to MIT?

JG

She was just such a warm and wonderful person, and intellectually inspiring. (Thinking back on that moment when she first stood in front of a classroom, I am sure that she knew at some level she was a warm and wonderful person, but she probably had no idea how intellectually inspiring she was!) So many people have had that experience with her. I'm really proud to have been her first student. When I met her, she had just recently gotten her degree. She must have been just about forty, though at my age I had no idea what that even meant. She'd been a grad student at

¹ I've added some references in footnotes, and all of my own papers can be found at my website, <https://people.cs.uchicago.edu/~jagoldsm/>.

Penn, where she'd first worked with Zellig Harris, for several years, but she was no longer his student at the point when she finished her degree. She got her degree with someone else; I think it was Henry Hoenigswald. She worked closely with her husband, Henry Gleitman, who was a professor of psychology at Penn—also a larger than life character, both inspiring and supportive.² After a few years at Swarthmore, Lila moved to Penn and eventually became a professor in the psychology department with Henry. But when she was at Swarthmore, she got a bunch of people like me interested in linguistics, several of whom went on in either linguistics or psychology.³ At the end of the first year that I studied with her (so it was the end of my sophomore year), she invited me to work on her grant over the summer. There were interesting things to do! I worked on interviewing a patient with aphasia, for example.

Ultimately she let me do anything I wanted to do. She asked me to write a paper, and I worked on one of Chomsky's papers at that point. He was criticizing generative semantics, so I was looking at criticisms, looking at George Lakoff's "Linguistics and natural logic." I just ate it all up. I had read *Aspects*, but I did not understand the issues that were dividing the field at that point. I went to the LSA meeting, the summer meeting in 1970, which was at Ohio State University, where I met Haj Ross, Arnold Zwicky, David Stampe, and Eric Hamp—and just had a fantastic time. I felt that linguistics was a community I wanted to be part of.

KI

Great, great. And since you're talking about this, so I think many people think of you as primarily a phonologist, and many people may not realize that you started in syntax. So how did you sort of what aspects of syntax really attracted you? And what was the transition point into becoming more of a phonology?

JG

Yeah, that's a great question. My answer isn't really very satisfactory in some ways.

I think I got attracted to syntax because that was clearly what Chomsky was most interested in, where he was making his major contributions. I remember just being totally blown away by *Aspects of the Theory of Syntax* (1965).

I read that in the spring of my first year, my sophomore year, the first year I was studying linguistics, and then I read Haj Ross's dissertation from 1967. I read the two at roughly the same time. Indiana University Linguistics Club had mimeographed Ross's dissertation, and it was in the library at Swarthmore College, and it too just blew me away. Haj would later be my teacher at MIT; we became close friends after that. He sadly passed away quite recently, the passing of one of the great syntacticians. So syntax was really where things were at in 1970, it seemed to me.

Oh, there was a psycholinguist at Penn named Harris Savin, who was a good friend of Lila's, who along with some grad students (including Lissa Newport, who works on American

² She went on to be a giant in the field of language acquisition. Her final and quite remarkable book was *Sentences First, Arguments Afterward*—what a brilliant title, from *Alice in Wonderland*—published by Oxford University Press in 2020. Susan Goldin-Meadow and I wrote an obituary for her in *Language* in 2022, <https://muse.jhu.edu/article/873753>. See also her wonderful reflections at <https://pubmed.ncbi.nlm.nih.gov/34623924/>.

³ I'm thinking of Muffy Siegel, Robert May, Emily Bushnell, and Margaret Allen.

Sign Language now) in psychology tried to read through *The Sound Pattern of English*, and I stuck it out with them for a couple chapters, and we struggled as best we could, but none of us really got much out of it, I suspect. Phonology was much harder to get into, and certainly generative phonology was hard to get into, from reading a book such as *Sound Pattern*.

Two years later, I applied to MIT for graduate school, and I was lucky enough to get in. So syntax was clearly where it was at as far as my head was concerned, and I loved taking courses with Haj Ross especially, and with Chomsky too, maybe to a lesser degree, although he was obviously very impressive. I learned a great deal with courses with Paul Kiparsky too, and from interacting with my fellow graduate students every single day. The first two courses in phonology were taught by Morris Halle, and they left an indelible mark. Morris was chairman of Linguistics, and he took a personal interest in every one of the students. I owe him a great deal for the chance to become the linguist I am today.

At MIT at the time, a graduate student had to submit two generals papers—essentially qualifying exams—which were research papers, and they were handed in in November of the second academic year.

[first work on autosegmental phonology]

Of the two papers, one had to be roughly in the phonology area, and the other had to be roughly in the syntax area, and so at the end of your first year, you had to come up with two topics. For syntax, no problem: I did a paper on infinitival relatives.

I was very inspired by Joe Emonds' work. Emonds had submitted his dissertation in 1969, and I remember reading it (everyone read everybody's dissertation).⁴ I thought it was fantastic work, and found it quite inspiring. So I worked on infinitival relatives following up on Joe's ideas. But for phonology? I had no idea what to do.

I had enjoyed Morris's two courses on phonology, but me and phonology didn't mix, or didn't combust. So sometime, probably at the very beginning of the summer, I went to Morris and said, can you give me a topic? He rummaged around on his desk and gave me a paper that had just arrived, a blue rexographed paper on Kalenjin vowel harmony by R.M.R. and Beatrice Hall.⁵ He said, Read this; you'll find something.

I read it a bunch of times, got nothing out of it, got no ideas, had no idea what was going on.

So at some point, I took my courage in my hand, I went back to him, probably towards the end of the summer, and said, I have gotten nowhere with this. Have you got another suggestion? So he said, I just got Will Leben's thesis in the mail. (I didn't know Will Leben.) And he says, read this. You'll find something in it. I read it—and yes, it was really interesting. I knew nothing about tone languages—African tone languages. And I didn't know much about phonology. But I found it fascinating. And the thing I tried to wrap my head around was a point that was for Leben, and Morris, central and terribly important.

⁴ A revised edition was published in 1976 as *Transformational Approach to English Syntax: Root, Structure Preserving and Local Transformations*, Academic Press.

⁵ It might have been "African vowel harmony systems from the vantage point of Kalenjin" by Beatrice L. Hall, R.M.R. Pam, Martin D. 1973. *Afrika und Übersee: Sprachen, Kulturen* 57:4 241-267.

It was the idea that a single vowel might have two tones. Morris had inherited an idea from his teacher, Roman Jakobson, which was that tones—well, you can have rising and falling tones, but when you do, they are really combinations of Highs and Lows. Rising tone is a sequence of a Low and High; Falling tone is a sequence of a High and a Low. And now, thinking of Japanese and classical Greek, Jakobson had stated that you can have rising and falling tones only on long vowels, because they are composed of two tones and a long vowel is made of two parts (moras, they're called), and a mora can have no more than one tone. You can have two such tones on long vowels because they have two moras.

Jakobson was big into moras, you know—moras are those things that if a vowel has two of them, it is a long vowel, and if it has one, it is a short vowel (as Jim McCawley has cogently expressed it). And so long vowels could have a Low and a High tone, or a High and a Low tone, or who knows, maybe two Highs because they were really two vowels—two moras at least. That was Jakobson's belief. And Morris was very proud of Nancy Woo (who I also didn't know, but who had written a dissertation in 1969 with Morris, in which she tried to establish this Jakobsonian generalization: two tones are possible only on a long vowel).

Well, a major part of Will's dissertation was to show this was not true. You can have tone melodies of two and even three tones that are realized on a single short vowel.

So you can see that this was really important for Morris's way of thinking about things. He passed the dissertation, and the question, on to me.

So what the heck is going on? Well, it was clear you can't put two tones *into* a vowel. And that was what Leben and Edwin Williams (who had been Will's roommate in grad school) that's what they were trying to do.

But that makes no sense. A segment is not an interval of time; from the Jakobsonian and generativist point of view, a segment isn't an interval of time, it's an atom.

There *cannot* be temporal order within an atom. So the only way out? I remember walking around a park near where I live, just puzzling these things. How can this be? The answer was the tones can't become part of the vowel. If they are to maintain their temporal order, they must retain their existence outside of the vowel. There's no alternative.⁶

So I said, okay, I have an idea. This idea has got to be good enough for a second year graduate student paper. I wrote up a paper called "Tonemic Structure" in which I proposed that the tones didn't get merged with the vowel.

The committee that read and evaluated my papers was Morris Halle, Noam Chomsky, and David Perlmutter. Well, they passed my paper on infinitival relatives clauses, but they flunked my paper on tone. They said, you can do better. We're not throwing you out of the program, but this isn't good enough.

There has been a tradition there—a Hallean tradition, I'm sure: flunking students. I wasn't the first person to flunk a paper. Yuki Kuroda, for example, was flunked on his phonology paper

⁶ In the last few years, there's been discussion of what some see as an alternative, in some ways inspired by ideas of Donca Steriade, and developed as Q-theory, in "Subsegments and the emergence of segments," by Stephanie Shih and Sharon Inkelas, *Proceedings of the Linguistic Society of America* 4. 37:1-8, 2019, and others more recently. There's the seed for a long discussion on how phonology has shifted its understanding of what its central concepts and questions are. I do not see this as a move in a sensible direction.

and his makeup, his do-over, was his classic analysis of Yawelmani Yokuts. So I followed in a proud tradition.

They either gave me, or I took, six months to redo my paper (I don't remember exactly what they told me at the time). I said to myself, if I'm going to have a new framework, it needs a name. So I called it autosegmental phonology. The segments are entities in their own right: that's where the "auto" prefix came from. And I did some work on Igbo, and the committee said, yeah, this is good enough. This is fine.⁷

Dave Perlmutter, in the end, didn't seem to like the paper very much, but he was willing to go along with Morris and Noam, who were more than satisfied with the paper, it seemed to me at the time.

That, then, happened in the spring of 1974. That summer there was a linguistics institute at UMass Amherst. Frank Heny gave a course on tone languages, which if I had been a more conscientious student should have given me a broader grounding in that literature. But alas, I just got too involved in my own tonological world that summer. I also started thinking about English intonation from the point of view of an African tone language, and had other things to do and to think about. I worked on Igbo quite a bit, using a very good descriptive grammar by Margaret Green and G. E. Igwe (*A Descriptive Grammar of Igbo*, 1963). They were very much inspired by their Firthian training. It was an excellent grammar, very insightful, and it helped me enormously. By the time I got back to MIT in the fall, I was really gung-ho about autosegmental phonology. And I wrote a paper that I called "English as a tone language," which gave rise to a large amount of work on intonation in English and many other languages, as you know very well. It was the first paper that I wrote for other people to read.

I continued to work on phonology, and there were other people at MIT, including Mark Liberman, Ivan Sag, Nick Clements, and Morris Halle himself, and others who found this work interesting and were pursuing it.

[semantics]

So I continued to work on it, too though, truth to tell, my own interest was in syntax and in semantics. I was working with Eric Woisetschlaeger, who was a very good friend of mine, a classmate at MIT, on semantics.

He ended up doing a dissertation on the work we were doing on semantics—involving the concepts behind tense, aspect, and modal verbs. And when I went for job interviews, I didn't talk about phonology. I gave a talk at Stanford on semantics, on the English progressive. I gave a talk at Indiana, also on semantics—my job talk was on semantics.

[teaching at Indiana University]

And then I ended up teaching syntax. My first job was at Indiana, and I mostly taught syntax while I was there (though the year long courses I did on field methods were really the best: one year on Igbo, one on KiHunde, one on KiRundi, with Firmard Sabimana, a gifted linguist and native speaker of KiRundi). I sort of branched out a little bit into phonology, but I really didn't do much phonology then.

⁷ The first and second papers are available on line; see my homepage.

KI

I'm curious what aspects of semantics attracted you, and what kind of framework at that time were you using?

JG

[semantics: Whorf and Heidegger]

What interested me was Benjamin Lee Whorf's work. I didn't mention Whorf in answer to your first question about how I got interested in linguistics, but I realize that it was really Whorf who pulled me into linguistics. I'm certainly not alone in that. I read Whorf when I was in high school and then again in college, and I thought a lot of people misunderstood him. His view is often oversimplified.

I think what he said about English wasn't terribly interesting, but his major point was that if you study the grammar of a language, it can give you great insight into very mysterious concepts, very deep and primordial concepts that are reflected in the grammar and the morphosyntax of a language.

I think that's true: he was right. So Erich Woisetschlaeger and I started looking at very simple things, the English progressive, the tense system of English, and the English verbal auxiliaries were what we focused on. We spent a lot of time looking at the semantics of the English progressive.

We wrote a paper, which eventually was published in *Linguistic Inquiry* (which might be the *last* place it ought to have published! We tried to publish it in *Language*, and it wasn't accepted. And eventually it was accepted in *Linguistic Inquiry*—thank you, Jay Keyser). Oh, I should say also, it was very inspired by Heidegger. We never said that in the paper, but one of the years at MIT I took a seminar. There was a mathematician named Gian-Carlo Rota (who passed away a few years ago)—I don't know if you know him.

KI

A very famous mathematician. A combinatorialist.

JG

Yes, exactly.

Rota was interested in phenomenology. And he taught an unofficial course for students in general at MIT, which I took—I had known very little about phenomenology. I found Heidegger extremely hard to read, but with some of Rota's inspiration, I began to see something in it, an effort to understand the fundamental concepts of human experience. This seemed to give me permission to try to rethink the basic concepts that are accessible to a speaker of a language and to think about time that was different from what seemed to me a horrible way—!—of thinking about tense, that is to say, the (Hans) Reichenbachian way.

I didn't know much about Reichenbach at the time. And while I think Reichenbach was a very interesting philosopher, what he had to say about time—although, yes, it does shed light on *some* things about the tense system of languages, especially compound tenses—I think that tense logicians who are interested in natural language have gone much too far with it and have not chosen

to look at the profound concepts that must underlie notions of tense that are much earlier, much earlier than the notions of the timeline.

So that was the kind of semantics I was interested in. It didn't fit into any particular framework. Erich and I went and talked to Hans Kamp, who was a visitor at MIT. He was very interested in what we were doing. I saw him decades later, and he said, Yeah, so how is Woisetschlaeger doing? And what have you done with that?

KI

Let me use this as a segue into another question, which is personal in some respects, since I come from the world of articulatory phonology, in which there are multiple notions of time, but one of them is quite close to absolute time, another one more intrinsic time that flows with respect to regular time, etc.

One thing that has interested me is how your the notion of time built into logical representations through the graphical device of association lines really departed from what went before. You've given lots of credit to Hockett, Kenneth Pike, and Will Leben, but there is a theoretically deep notion of time that enters phonology through the autosegmental framework. Was that in any way sort of related to those other worries about semantics and so on?

JG

Good question. I fear that it would be easy to give a facile answer, like: time is at the center of any deep question that you might want to think about, whether it's human or scientific. I can't give a more interesting answer, one that isn't facile, at this point. Connecting the kinds of time we talk about in phonology and phonetics with the kind of time that I was referring, involved in basic concepts of tense, I would like to think I could give a good answer, but I'll have to think about it. Perhaps one simple answer is this: don't take time for granted. I'm stating that as an imperative, but it's an imperative addressed to myself: John, don't take time for granted. Don't assume that just because you know what a Cartesian axis is, and you can put time on a line that stretches from left to right—don't think that just because somebody showed you how to do that, that therefore that's what time *is*, and now you know how time works. No; there's so much more to it. We certainly haven't reached the end of our journey to finding the right ways to conceptualize time. That's something of an answer, a very inadequate answer, to your question.

KI

Even if we ignore that semantic side, and the like, would you like to talk more about your theoretical conceptualization inspired by some aspects of earlier work, but this kind of very novel idea with an analogy of other segments linked to each other with association lines?

JG

Let me back up a step and say something about autosegmental phonology, which departs from the most familiar view of phonology, certainly strongly embedded in the view described by Saussure and passed down to us.

In his tradition, the language consists of a string of atomic units, and if you asked how long each atomic unit was, the answer would either be that they don't have a length because from the

physical point of view, they can overlap, so they don't have a sharp beginning or end—that would be one answer. Another answer would be, some segments are longer than others, vowels are much longer than most consonants and so forth. So the departure of autosegmental phonology is to say that there are two (or potentially more than two) parallel sequences of these units that we call segments.

And what's always been unsatisfactory about that, as you know every bit as much as I do, is that there's so much that involves the quantitative notion of time that the segmental perspective leaves out completely. The funny thing about scientific models (this is certainly true about linguistics in spades), is that when a model is powerful enough to give you a handle on difficult problems, its success always comes in part from getting you to not think about things that are both obvious and obviously important. Like the fact that some vowels are very long from a temporal point of view and some segments are very short, yet as a phonologist, you don't pay any attention to that at all.

probably went years as a budding phonologist without ever thinking about this—and that's the power of a model, the model of “phonological segments”. It allows you to not think about timing, probably because we can write our segments down on a piece of paper.

But there's so much about the physical models, as you and I have talked about for many years. So that simple fact drew me towards thinking about rhythm and then, oh I should say too, this is important, and we're sort of skipping around in my career a little bit. But in the late 1980s and early 1990s, the late 1980s was a period in which work on neural networks and connectionism, in particular inspired by the Parallel Distributed Processing (PDP) group centered at UC San Diego, made an impression on a lot of people, including me.⁸

I get very interested in that in the late 1980s, as you know, and into the early 1990s. People were doing really interesting work on the brain at that time, as they've done before and after.

But that was a time during which there's a lot of interest in oscillators, essentially, seeing what very basic principles of circuitry can give rise to oscillators, what it means for oscillating systems to be in sync and the like (what's called *entrainment*). And I found that extremely interesting because it seemed to be calling out to phonologists. Oscillators were a very natural way to think about time from a quantitative point of view, and also to the possibility of having oscillators of different frequencies and seeing how they interact.

So my answer to your question is that that interest in another aspect or another way to look at time that went beyond the usual phonological way, that came as I got interested in neural nets and in what people were saying about the brain in the late 1980s, early 1990s.

KI

There aren't that many scientists in our field who do a major amount of work in one field, subfield, go on to another, go on to another. If there's a student who is actually intrigued by many different aspects of language—which feels like what being a good linguist should be like, not obliged to stay within two blocks of one's neighborhood. What is that like, and what has that been like within your

⁸ Take a look at *Parallel Distributed Processing*—volume 1 can be found online at <https://gwern.net/doc/ai/nn/1986-rumelhart-pdp-v1.pdf>.

career? Positive aspects, negative aspects? As potential advice for others who may find themselves in a similar situation as you were in different parts of your career.

JG

The way that question projects onto my own career lies in the simple fact that I've wandered through many different subdisciplines either in or very close to linguistics (linguistics-adjacent?). I could formulate a local version of your question: how did this shifting of research focus happen? And what has the impact been on me as a person, and as an academic? Is it a good thing, is it a bad thing? If somebody else wants to be that way, what do they have to do? Lots of questions, none of them very simple.

The first thought that comes to mind in response to your question is one that has a slight tinge of bitterness associated with it. But since it does come to mind first, I'm going to go with it, recognizing that it's only a really partial answer. But a couple of times I've simply left a subdiscipline.

"Left"? That may sound like a funny word to use there, but it's the right one. I left, because I was unhappy with the state of the field that I was leaving. There were two times this happened. One involved syntax, and the other phonology.

I was really unhappy about the state of those fields, even while I felt a pretty big commitment to them at the time. And I was really grateful that linguistics was big-hearted enough and big-roomed enough that it wasn't a difficult thing to say goodbye.

As I was said earlier, I started off as a syntactician, and my first job was in 1976. I was at Indiana University from 1976 to 1984, and although, towards the end of that time, I was doing a lot of work on Bantu tone, I think I was still primarily thinking about syntax at that point. Certainly the dissertations that I was advising were primarily about syntax. And in fact, when I went to Chicago in 1984, the job there was defined as a syntax job, and I started off the first quarter I was there, we read through Chomsky's *Lectures on Government and Binding*.

That was a period in which I was quite disappointed with the direction Chomskian syntax was taking. Actually, already by 1982, when *Concepts and Consequences* came out. I just thought, Really? What is this? What are people *doing*? I had already been saying to people I knew who I took to be my friends and colleagues, why are you doing this kind of work? They would say, well, I cite Chomsky's papers and ideas because I want to be read. If you don't work within the framework that Chomsky's setting out, people won't read you.

Now, of course, they meant the people that they read and cited wouldn't read their work anymore, but regardless—I thought this was a terrible answer. In fact, what I thought was: if you're going to be an academic, you're not allowed to talk that way, or even think that way. You have to do what you think is right from a scientific point of view. If you don't, you're in the wrong profession. Eventually, around 1985 or so, after I got to Chicago, I switched gears and stopped paying attention to the kind of work that I had been following in syntax. (I wrote a paper that expressed my sentiment about what Chomsky was doing at that point that is quite different from any other paper I've written; it was entitled "Lectures on bubblemint and grinding."⁹.)

⁹ <http://people.cs.uchicago.edu/jagoldsm/papers/2006-bubblemint.pdf>

Since I had been moving over to phonology, and some people thought of me as a legitimate phonologist, moving into phonology wasn't a hard thing to do, and around 1985, I started shifting my focus to phonology.

I had spent a year at Harvard in 1980, working there with Nick Clements, and the work we did there got me quite excited about Bantu tone systems. So there I was: doing phonology. I profited immensely from the active Africanist community in the Indiana University department of linguistics. And as I say, around 1985, I started thinking much more seriously about phonology. In 1991, as you know, optimality theory came on the scene, due to work by Alan Prince and Paul Smolensky at the 1991 Linguistics Institute. And in short order, in two years or so, the field of theoretical phonology got pulled over to optimality theory and OT points of view, and for reasons that we will no doubt talk about, I was underwhelmed by that work. I didn't feel like a lot of critical thinking went into the enthusiasm that many people that I respected a good deal showed for OT. At that point I decided I really didn't want to have my work evaluated by people who were so enthusiastic about OT. So there, too, I felt like I needed to leave to find something else to do. Time number two, in about ten years.

Now, curiously, at that point, I was chairman of the linguistics department at Chicago, and I was not in a position where I wanted to pick something else to do in linguistics. (By the way, it is not at all uncommon for researchers at that age to check out of doing research entirely. I did wonder if that was happening to me at that point.)

And actually, just as a footnote, I chose to work in a kind of an area that was related to the social and political world in which we live, for a year or two at that point, while I was being chair. I worked on the John F. Kennedy assassination, and I taught a course on that subject twice. It was a wonderful teaching experience. The point of studying the Kennedy assassination is not to figure out who shot Kennedy—we've known that for a long time. The work that was done already in the 1960s and 1970s, work that laid out a great deal of information that clarifies what really happened, if one is willing to put in the time and read the material with an open mind. But the point of studying it is to understand something larger about politics and how media is controlled, and how the intelligence community cooperated with organized crime in the early 1960s. So it was a very interesting way to spend some time engaged in undergraduate teaching; and I was busy with administrative work, so I did that for two years, and then went off on sabbatical. A time to rethink what I was going to focus on.

[machine learning in the 1990s]

That sabbatical time allowed me to work with computer scientists and computational linguists at Microsoft Research, and I'm very grateful that they allowed me to spend time there; it was a great opportunity for me intellectually. My wife at that point was a researcher at Microsoft, which was an essential element of what made that possible.

That was the point at which machine learning was taking off, and I found that very exciting. From where I sat, the invention of machine learning was, more than anything, the realization by people in statistics and computer science, beginning in the mid 1980s, that what was interesting about neural networks was not the hardware but rather the mathematics that they implemented—a most reasonable conclusion. Now, I had always had a strong connection to computer programming, ever since high school, and my first love was mathematics. So it was natural for me to move in that direction, and people at the University of Chicago—both colleagues and

administrators, if I can draw the distinction (which is not at all obvious, at the University of Chicago, unlike many other places!)—were extremely supportive.

So, drawing these stories together: for me, the two changes of direction that we've talked about (from syntax to phonology, from phonology to computational linguistics) were very easy to do. In my head, easy, and it was easy to find other people to work with, and institutionally it was easy. That's always been one of the great things about linguistics: it has so many subfields, and since the entire discipline itself is not that large, it's easy to know (or get to know) people who are in these different areas. This has been a long answer to your question! And we haven't even touched on my move into studying the history of linguistics and its neighboring fields. That's another story.

KI

I'd like to go back to the beginning of your work on autosegmental phonology. You were interested in very local issues in tonal systems, both English and other languages, such as Igbo. You proposed theoretical constructs to be able to talk about how tonal systems and vocalic systems can sort of be related to each other. What I find interesting is even though you proposed the theory for such a local context, within the 1970s, these association lines started to take on more and more meaning in phonology and morphology, and then through other lexical syntax, potentially other areas binding. At what point did you realize that kind of the how those those very, very specific the very specific context in which it arose, that this, this way of looking at the structure of linguistic representations, it's something well, well beyond the kind of the problems of addressing tone, and how many tones can fit on a vowel?

JG

Another great question, Khalil. Let me start by saying I'm not sure that people who read my dissertation—which is, I think, how most people found out about autosegmental phonology—understood the central proposals, which fell into two quite different clusters.

What were they? The first has to do with representations in geometry, and the second has to do with alternatives to rules (and therefore to derivations).

The first point is that one of the primary jobs of the phonologist should be to study the geometrical representations that we use in our models. Now, that was already obvious in the domain of syntax. Although Chomsky wasn't really big into geometrical representations in his original manuscript, *The Logical Structure of Linguistic Theory* (1955)—although he wasn't that much into syntactic trees there, he very quickly adopted them into his work, and the world that was banging on his door made it clear that understanding syntax through syntactic trees was huge; it was, indeed, the key to thinking about syntax. (Yes, some people had been working on the problem before Chomsky, but nobody really hit the nail on the head till he got there with his trees.)

But people had, for the most part, not tried to do anything parallel in phonology. So the first thing I was trying to say was, no, you have to study geometry, you have to propose new and interesting geometries, and the picture in my head that was chemistry and quantum mechanics. I had studied quantum mechanics in college (or rather, the mathematics of quantum mechanics, not entirely the same thing), and doing that encourages a person to use both data and a sense of mathematical aesthetics to follow their nose in finding new models. So, physics and chemistry was

the model that was leading me, and I was asking, how do we understand the relationship between segments?—trying to ask how this is like the relationship between atoms, that is, trying to think about what corresponds to electron shells and the like. That was the inspiration for association lines. The question was always not just describing phenomena in languages, but asking, what’s the underlying object? What is its geometry? Geometry in a rather abstract sense, to be sure.

And the second big idea was what I called *the Well-formedness Condition*, which was the idea that the theory as a whole would add and delete association lines in a minimal fashion in order to maximally satisfy the requirements that autosegments be associated (to the extent that it’s possible in a given representation). Much like shells wanting to be filled by electrons in a fashion that is independent of how many protons happen to be in a nucleus.¹⁰

So you had this central theoretical constraint, the Well-formedness Condition, which was violable, and which, in a certain sense, caused, logically caused association lines to appear or vanish, which are the essential things in this game. Association lines were like, as I say, electrons appearing in certain shells. They were the conceptual equivalent. Association lines would be inserted or deleted automatically without rules doing that work.

So the idea was that most of the interesting work, hopefully, would be done by these principles that were not rules as such. So did this involve derivations? That depends to some degree on exactly what we mean when we use the term “derivation.” A well-formedness condition that adds a set of association lines to a representation: does the new representation that contains the added association lines form a new step in the derivation. Well, yes, possibly, if we’re looking at changes in representations and decide that that’s what it means to form a link in the chain of a derivation.

But if derivations mean are that are justified by the presence of rules—then no, the effects of the Well-formedness Condition stand outside of the rule-based derivation. That was certainly my view.

So it’s a dynamic theory—not static, nor limited to the sharing of features, like British anti-derivational theories would have it; but it is or isn’t derivational, depending on how you define derivation.

So those are the two basic ideas, and that first idea (i.e., we have to focus our attention on the geometry of phonological representations) certainly grabbed the attention of many phonologists.

And it’s disappeared now, in large part though certainly not completely. Since the advent of optimality theory (OT), that’s to say now, 30-plus years, the field has moved away from that kind of thinking.

Obviously there are areas where autosegmental theory is central. Anybody who works on tone uses some version of autosegmental phonology. But then the notion of finding forces that are larger than individual generative rules to accomplish changes in the phonological representations across a language, or even across languages, that is something that people don’t talk about in those

¹⁰ Let me interrupt here to spell this out a bit more. The original Well-formedness Condition had three parts, and an enforcer. The three parts were: a. All vowels are associated with at least one tone; b. All tones are associated with at least one vowel; c. Association lines do not cross. The enforcer was: Add or delete association lines minimally in order to maximally satisfy the three conditions.

terms. The Well-formedness condition was an attempt to say what structures are *good*, are sought by a phonology, are tendencies towards which phonological structures are pushed.

Putting it that way makes it sound like something that Trubetzkoy and Jakobson in the 1930s would have gone for: they hated nothing so much as the “positivism” of the Neogrammarians, who thought that phonology moved randomly in any generation; they thought that to the contrary, language changed by responding to the forces that their structuralism was attempting to uncover, notably in the theory of oppositions that uncovers the structure of phoneme inventories. It is remarkable that Chomsky and Halle never understood that this is what Trubetzkoy and Jakobson had in mind. They made no secret of it in their writings.

Back to the Well-formedness Condition, which was a violable constraint, connected to a mechanism for repair of its violation in a minimal way. This was a central idea, but I did not pursue it enough, and no one else picked up on it. I emphasized this point in the final chapter of my 1990 book (*Autosegmental and Metrical Phonology*). I’ll come back to it again in the context of the linear dynamic models that I developed around that time with Gary Larson.

Looking back, I can see that an important factor in this was the fact that some people—I think first of Doug Pulleyblank—pointed out quite rightly that the formulation in my dissertation of the WFC was wrong; while it made many right predictions, it made consistently wrong predictions (over-predicting associations in cases that could be clearly identified). That was fine, great, and I’m happy to be shown wrong, but then the next question is, what do we do now? The answer surely can’t be that we replace it with rules for each language. No, the question is, how do you do the job better. I did not work on that (as I say, that was when I had slid out of phonology, in the late 1970s!), nor did anyone else.

KI

I appreciate the way that you’ve developed the answer, because it’s what I think of as a trade-off between representations and the computational component. That was a foreign idea in 1974-1975, when you introduced it. Then it takes on the role for a good 15 years as the central concept in the field. And then surprise! In the period 1991-1993, we see the exact opposite—trying to define purely the computational component and just feeling representations, just bring your own theory of representation. But there are many people who do assume autosegmental association lines. So the so the trade-off is not really well managed, because they do assume very complex representations of the other mental types, so it’s yeah, but it’s it’s an interesting development in the field.

JG

Maybe this would be a great time to talk a little bit about something which you and I have talked about at length. And that’s the relationship between some of the ideas that went into OT, brought in by Paul Smolensky, and the development of this second interpretation of the Well-formedness Condition.

In the late 1980s, second half of the 1980s, after I moved to the University of Chicago, I was working a lot on Bantu tone languages. The most interesting theoretical idea I was pursuing was an attempt to develop the Well-formedness Condition to deal with the interaction between tone association, on the one hand, and accentual structure on the other.

I taught a course on this with Nick Clements at the 1987 Linguistics Institute (I should say, that was what I was talking about in my half of the course), which Ivan Sag organized at Stanford University. There I met Paul Smolensky, and I also learned about connectionism. David Rumelhart was teaching a course on parallel distributed processing (PDP). Paul Smolensky had just come off of developing harmonic theory in the context of the PDP project.¹¹ George Lakoff suggested to me that I should talk to him, because he thought that what I was doing in phonology would resonate with what Paul was proposing, and vice versa. So we talked and got to know each other, and I agreed with George that there was indeed a resonance between what I was trying to do and what Paul was trying to do.

Paul was trying to use connectionist networks, fully distributed connectionist networks, and using ideas that came from thermodynamics. Most neural networks, both before the PDP group's work and continuing to this day, are organized in layers, and most of the astonishing work takes place in the computation between successive layers. Physicist John Hopfield proposed a different kind of fully connected network in 1982, one which was fully connected inside a layer, so to speak, and this got a lot of people excited, because its dynamics were very similar to the models that were familiar to physicists describing thermodynamics. Paul Smolensky's work in the PDP project developed that intuition in very interesting ways, and I very much resonated with the idea that he explored (that was also explored by others, typically former physicists, such as John Hopfield and notably Daniel Amit, whose work I found inspiring¹²).

Paul and I talked and communicated after that institute, at that point, by U.S. mail. Paul got interested in phonology, and he read and carefully commented on my 1990 book, *Autosegmental and Metrical Phonology* (which he curiously referred to later as a textbook, though it was nothing of the sort).

Paul and Alan Prince then developed optimality theory, and applied it to phonology. To me the disappointment there was that he didn't use the power of the neural nets, the recurrent neural nets that he had explored in the PDP chapter of his career. (When I say recurrent neural nets, I should add as a footnote, I'm using the notion of recurrent that was used in the 1980s, the Hopfield-style network—which is not the way the term “recurrent network” is used now; today's recurrent network was called an Elman network then.)

Anyway, I was disappointed (and remained disappointed) that the ideas that Paul had worked on in the 1980s did not get integrated into the OT. Paul tried to do so in the years that followed in his book *The Harmonic Mind*, but in my view it wasn't a success. The relevant concepts were retrofitted onto OT; OT was not a theory based on those ideas, it seemed to me.

KI

In your early work on autosegmental phonology, you proposed very specific things for very specific phenomena having to do with questions like how many tones can go on a vowel. Then Nick Clements, John McCarthy, Donca Steriade, Morris Halle — a lot of people adopted this way of

¹¹ See my remarks above on the PDP project.

¹² See his *Modeling Brain Function: The world of attractor neural networks*, for example, at <https://archive.org/details/modelingbrainfun0000amit>

thinking about phonological representations, going way beyond working on tone. I'm curious about how you realized that what you had proposed was so consequential. Especially in light of the fact that you said you don't really think of yourself as exclusively a phonologist, or maybe even at some stage, mainly a phonologist. Yet there was this idea that in some ways took over the field, and moved beyond syntax, through Jerry Sadock's work on autolexical syntax. Even though that didn't have a huge readership, it still was trying to link morphology, syntax, and semantics, and phonology using this tier/association line idea. My question is a historical question about the evolution of the idea and the contribution of so many people to the nucleus of an idea.

JG

I started to give an answer yesterday, and it focused on two subparts, and these are the two central ideas of my thesis. The geometry of representations on the one hand, and an alternative to derivations on the other, which meant in more concrete terms, the role of the Well-formedness Condition.

I started my answer to your question in that way first, because that's how I think of my dissertation, but also because it's a particular way to answer your (to me quite interesting) question. Generativists are always looking for alternatives to derivations! Now, that sounds a little bit ironic, or is it paradoxical? Probably both.

But the fact is that both of Morris Halle and Noam Chomsky were deeply committed to a derivational point of view. (And I never have been.) Chomsky created a number of theories and approaches leading up to minimalism, the final and greatest of his theories, some might say. And they've all been so deeply derivational in terms of a linear sequence of representations, and these representations are related by these explicit rules. That last part has changed over time, but still. And generativists, coming from many, many different places, conceptually speaking, have occasionally been in sync and resonated with these feelings of Morris and Noam—but mostly not! They've *accepted* derivationalism as part of belonging to Team Generativist, but viewed certain aspects of the generativist picture—most notably extrinsic ordering of rules—as not appealing at all.

From that follows something of an explanation for why autosegmental phonology attracted people, because they saw it as an alternative. You mentioned Jerry Sadock, for example, and his autolexical grammar (making clear its relationship to autosegmental representation); Ray Jackendoff has developed similar ideas, and Ray has always been very explicit about the degree to which autosegmental representations inspired his work in that direction.

Autosegmental representations offer a sense of how to think about the relationship between the representations for each component (morphology, syntax, etc.). Here's an example: there's this big tree for a relative clause in Igbo in chapter 2 of my dissertation. And it has a syntactic tree assigned to it. But it's a double tree, one with a root at the top of the page, going downward to the terminal elements, the words and morphemes, and another rooted at the bottom of the page, going upwards, towards the words and morphemes. One tree for the syllables, so to speak, the segmental morphemes, and another tree for the tones. For the most part, the trees are the same, except for this one high tone, which is a relative clause morpheme. And it is very cool to look at these two syntactic trees, and how they fit together, they're almost identical, except in one spot. It gives an

interesting way to think about how the syntax and the autosegmental phonology affect each other mutually.

In 1977, Dan Dinnsen, who was a colleague of mine at Indiana, organized a big conference on theories and phonology, and he included autosegmental phonology as one of those theories. There was a lot of talk at the time about constraining theories.

That was because Chomsky had spent a good deal of time talking about the importance of constraining theories in syntax in order to claim the methodological high road in relation to the benighted generative semanticists, like Paul Postal and Jim McCawley. That was really a bunch of nonsense in the arguments in syntax.

[against constraining theories]

One of the points I tried to make in my presentation at Dan Dinnsen's phonological conference was this—and it was a response to what people were saying at the conference. Presenters were talking about constraining theory this way, and that way, and another way. My response to all this was, no, come on, just get off this “constraining the theory” stuff. The crucial point is to have *insights* into what's going on in language, and to use those insights to *enrich* the theory. When you enrich the theory and come up with new and deeper insights about what's going on in language—then you will see that there are all sorts of things that your theory allowed you to do that you no longer need to do, things that you don't use, and you can cut them out of the theory. The *goal* is not “to constrain the theory”. The *goal* is to come up with new and more insightful analyses of the data, and to integrate them into your theory. Constraints will fall out of the process of integrating new insights. (That's what people have always been looking for: new ways of thinking about things, which they can then use to understand the phenomena that they're studying.) Getting back to your question: autosegmental phonology set set people's minds going, set the juices flowing, so to speak, so they could think about things in new ways. That's my answer to your question; it may sound vague, but to make it concrete is simply to go back to what I was saying last time, to the two central points of my dissertation, of which the first was: come up with a better understanding of the geometry of the representations you're studying, and that's exactly what Jerry Sadock did with autolexical grammar. Constructing trees that he could put on the blackboard in his office. And then also, looking for alternatives to rules that “simply” tell you to change one representation into another. Instead of that, and moving towards a theory in which certain representations are *better* formed than others, which aligns with a tendency in that direction. I mentioned earlier Trubetzkoy and Jakobson's despair in the 1930s with what they considered the positivism of the Neogrammarians, who thought that phonetic or phonological changes could go in any direction; for Trubetzkoy and Jakobson, the crucial thing was that according to *their* understanding of structuralism, changes are essentially directional, and not just because of the arrow of time; the direction of change is due to forces that are inherent to the phonemic inventory. Anyway—there's a connection there to autosegmental phonology. The well-formedness condition is an example of a direction in which association changes addition and deletion, should go in order to maximize the well-formedness of a representation. Those are the reasons that the work on autosegmental phonology had the impact, because basically it allowed other people to think about their problems in new ways, and that's always a good thing, and it's hard to come up with new ideas.

KI

There are a couple of things arising from previous discussions that I've always wondered about. I'd like to know more about the ideas inherent to generative semantics and interpretive semantics — about the syntax/semantics relation. There is one particular thing (since you mentioned Jackendoff) that I find extremely peculiar. Within a three to four year period, there was your work on English as a tone language. And there was the work by Jackendoff and Chomsky that was part of the argumentation against generative semantics based on focus and tonal accent.

JG

I fear I'm not too knowledgeable about that. Those are good questions, but my experience with that is a bit limited. The work that you're talking about, which Chomsky was doing, and Jackendoff's work that he cites, was before my time at MIT.

It's certainly interesting that Chomsky had a handle on these observations about intonation and focus. I've been told by someone whose judgment and experience I trust that Chomsky's familiarity with that material on focus and intonation came from his interactions with Albert Kraak, whose work Chomsky actually cites.

There are some other interesting things to say about the generative semantics, interpretive semantics debate or dispute. You want to talk about that a little bit?

KI

Yes, let's.

JG

[the generative semantics period, and sociological failures]

I wrote a book with Geoffrey Huck about the generative semantics/interpretive semantics rift.¹³ When I think back on that period, I tend to think about the major lessons that we as a field should learn are primarily sociological.

When I say they are sociological, I don't mean they are not scientific. I actually think the lessons lie at the very heart of what it is to be science. Let me emphasize that: if you don't study and think about science from a sociological point of view, you're not thinking about science in a broad enough perspective.

Let me drop a footnote here—I just read a statement that Fritz Newmeyer makes in his recent book, *American Linguistics in Transition*—an interesting book. I disagree with a lot that he says, but he certainly represents some widespread points of view. At one point, he criticizes something he wrote earlier, several decades ago, about whether the Chomskian turn in linguistics was a Kuhnian revolution. He writes that he disagrees with what he wrote earlier because it was too sociological in character. I couldn't agree less on that! To understand science is to understand its sociological character. Western science as we understand it began with the creation of the Royal Society in 1660, because it that represented the transition from an individual enterprise to a social

¹³ *Ideology and Linguistic Theory: Noam Chomsky and the Deep Structure Debates*, Geoffrey Huck and John Goldsmith, Routledge 1996.

enterprise. Scientific knowledge is not individual knowledge; it is knowledge that has passed through a complex social structure that has been constructed (consciously and with intent, for the most part) by human beings for particular ends. Those ends are scientific knowledge.

Let's look at the work that was being done at MIT, in the period we're talking about, the mid-1960s. From the point of view of the way science ought to work, all was not well. The most obvious way in which it wasn't being done well is that people weren't going through the review and publication process. There are good things about that, but a lot of bad things. The good things about it involved the ease with which ideas could move from one person to another, which is great (that's also something to study sociologically, how ideas move around¹⁴).

But the way in which people needed to convince one certain other person! —that was the terrifying aspect of this process, whereby Haj Ross, for example, felt like he absolutely needed to convince Chomsky of his position, and kept on failing to do so. Science cannot be based on convincing one great figure in a field.

Haj and Chomsky shared letters going back and forth—odd, since they were colleagues in the department, but oral communication had quite literally broken down between the two of them. Chomsky displayed tremendous intellectual immodesty.

At a purely personal level, Chomsky is a tremendously modest person, and I admire him greatly in many respects. From an intellectual point of view, he's quite the opposite, and I can't say that I fault him for it. It's part of his great intellectual strength. But at the same time, his intellectual immodesty and his insistence that everybody has to look at things his way doesn't fit with the nature of a healthy scientific enterprise. In science, everybody is playing on a level playing field. Science works well when there are referees (in the sports sense, not the publishing sense) who are not fans of one team or another.

Chomsky doesn't work that way. That's fine when you're having a conversation, or giving a lecture, or exchanging letters, as happened then. But science...science is different. When science is done right, you write articles and have them reviewed anonymously by journals and publishing them, and so there's a final form that can be shared by the discipline.

This process should have filtered a lot of the noise—the crap—out, but that process didn't happen, and of course the noise didn't get filtered out; it enraged everyone, until it didn't, and everyone stopped arguing with each other. What is very striking, when you read the documents that illustrate what was really going on, was that there was far too much effort expended trying to convince Chomsky. What should have happened was that each person and each perspective should have continued to refine their perspectives as best they could, recognizing that there would ultimately remain multiple perspectives in the field: there would not be a single winner surrounded by corpses in the field of battle.

When I look back on that period, I think we could have done better as a discipline. We let one person—Chomsky, in the event—influence the interpretation of the questions, and a large part of that out-sized influence was due to the failure to treat publication seriously.¹⁵

¹⁴ Which is the central question of *Battle in the Mind Fields*, University of Chicago Press 2019, John Goldsmith and Bernard Laks

¹⁵ I do not take myself to be a paragon in this respect in any way, I should add. My publication practices leave a great deal to be desired.

KI

Let's go back to something you said yesterday, where you started to introduce the ideas of the dynamic linear model.¹⁶ I see a trade-off between computation and representation in phonological theory, and that lay at the heart of what you were doing.

JG

I mentioned that in 1987, I taught a course at the Linguistics Institute, where I was thinking a lot about quantifying the notion of well-formedness. In the particular context, the problems I was looking at were relations between accentual patterns and tone assignment in Bantu languages. I use the metrical grid as the formalism for the grid that Mark Liberman proposed, and Alan Prince developed.¹⁷

I was looking at that, only making what seemed like relatively small steps. But then about two years later I started looking at some problems in syllabification and its relation to sonority. I was very much struck by—and inspired by—a paper by François Dell and Mohammed Elmedlaoui that got a lot of people excited at the time.¹⁸

In their paper, although they did not phrase it this way, you could see the role that competition played in the proper syllabification of words and phrases. Basically, different segments would compete for being the nucleus of a syllable. The greater the sonority of a segment, the greater was its ability to claim the role as nucleus—though there were forces that could come in and counter that. And bear in mind: the notion of competition plays no role, and can play no role, in generative grammar.¹⁹

It occurred to me that I could write a couple of equations that would provide a solution to the problem without a set of ordered rules and derivations, as Dell and Elmedlaoui had done (they were inspired in their work by ideas of Lisa Selkirk at the time).

I worked on that for a while, working closely with Gary Larson, a grad student at Chicago at the time. We did some really interesting work. The connection that I want to make, though, is that those computations were, as I see it, my attempt to make good on promises that the Well-formedness Condition of autosegmental phonology made but didn't deliver on.

The well-formedness condition said this: you need to add and delete association lines in a minimal fashion to maximally satisfy three very simple and natural structural conditions relating

¹⁶ See, for example, "A Dynamic Computational Theory of Accent Systems." In *Perspectives in Phonology*, edited by Jennifer Cole and Charles Kisseberth, pp. 1-28. Stanford: Center for the Study of Language and Information. <https://people.cs.uchicago.edu/~jagoldsm/papers/1991-UrbanaDynamicComputational.pdf>.

¹⁷ For example, "Relating to the grid." Alan Prince, *Linguistic Inquiry* 1983 14:19-100, a paper that impressed me greatly.

¹⁸ "Syllabic Consonants and Syllabification in Imdlawn Tashlhiyt Berber," Francois Dell and Mohammed Elmedlaoui, *Journal of African Languages and Linguistics* 7:105-130.

¹⁹ That could be the subject of an interesting conversation: how alternative mechanisms had been explored in generative grammar, most notably rule ordering, but also the difference between underspecification and full specification, in the attempt to deal with phenomena that reflect competition between alternative analyses.

tones and vowels. Well, that's fine, but when you look at concrete examples, rare are the examples where that prose is enough to really solve the problem. You need something that is explicit, quantitative, something you can sink your teeth into.²⁰

And I didn't do that further work with the Well-formedness Condition in autosegmental phonology (nor did anyone else, alas), but this work on these dynamic linear models was work on the same *kind* of question. In this new model, the theory consisted of a very small number of simple equations. You plug in certain numbers—the parameters—and out come the results that you need, and they talk specifically about the relationship between adjacent symbols on a (linear) tier—i.e., a sequence of symbols. The symbols could be segments; they could also be syllables. In the case of a model of sonority, the symbols are segments. The idea, essentially, is that segments have an inherent sonority, and a derived sonority, i.e., a sonority in context. I leaned heavily on an idea developed by Lisa Selkirk and Nick Clements, that sonority could be viewed as a variable that takes on numerical values. But, in addition, as I say, a segment has not only an inherent sonority, but a derived sonority. And its derived sonority is affected by that of the segments to their left, to their right, and to their position inside a word.

I'm not going to go into any details here, but the interested reader can get a sense of what is going on with an equation showing part of the interaction. We'll call the activation of the i^{th} unit a_i (which is the derived sonority in this case). It has an inherent sonority, $Inh(a_i)$, and that is its activation level on the first calculation; we update that calculation, noting the time as superscript t , with the equation $\alpha_i^{t+1} = Inh(a_i) + \alpha a_{i+1}^t + \beta a_{i-1}^t$. (In the fullness of time you showed us how the quantity could be found without an iterative calculation, Khalil.²¹)

What does this have to do with the Well-formedness Condition? Both this equation and the WFC describe local effects that have global consequences which “intercede,” so to speak, to keep the representation maximally well-formed at all times. (“times”? Yes, times.) Only this time, we see it in its full glory, and can calculate it.

[Information theory]

KI

There was a stage in your career where information theory played a major role. Could we start this by you telling me how you think about information theory, and how it structured some of your research into both phonology and unsupervised learning of morphology, and other things that you feel were very much touched by information theory?

JG

This involves a series of moments that I haven't talked about much—yet most of those moments are very, very clear in my own personal trajectory. There was a period around 1993 when I felt like distancing myself from phonology—we talked about this briefly, in connection with

²⁰ A not particularly successful effort along those lines can be found in a paper I wrote in 1984, “Meeussen's Rule.” In Mark Aronoff & Richard T. Oehrle (eds.) *Language Sound Structure*. MIT Press, 245-259.

²¹ “The Dynamics of Prominence Profiles: From Local Computation to Global Patterns,” Khalil Iskarous and Louis Goldstein. In *Shaping Phonology*, edited by Diane Brentari and Jackson L. Lee, University of Chicago Press 2018.)

optimality theory. In 1995, my wife took a job at Microsoft Research, working in natural language, and I spent a year there as a visitor working on intonation. Rick Rashid was head of Microsoft Research at the time, and I reported to Dan Ling, part of Rashid's group. Dan was a terrific person—I really enjoyed having him as my boss during my time there. I was working in X. D. Huang's group, which was focused on text to speech development at that point. A historical aside: this was the moment that cell phones and email were both coming in, and it was widely thought that the next killer app would be the one that let you get your email over your phone by reading it to you.

So I worked for a year on intonation, using autosegmental phonology. Over the next couple of years I had some further opportunities to stay there, but I felt my work on intonation had come to a close.

Some of the people in the group of linguists that my wife worked with were working on the CJK languages (Chinese/Japanese/Korean), and the problem there is a problem of long standing: when you build a syntactic parser, you'd like to have your input parsed into words, but that is not normally provided to you from the standard orthography. So there was a lot of discussion about how you infer what the words are in the CJK languages when all you have is input from the keyboard, without spaces or word-separators.

Somebody on the internet (which was very young at that point for things like that) pointed me in the direction of a dissertation by Carl de Marcken which had just come out. It was about how to take a large textual corpus in which there are no indications of where words begin and end (spaces have been eliminated), and to reverse engineer the words: putting the spaces back in, even when you don't know the language that the text came from in the first place.

I downloaded the thesis, and I started to read it. De Marcken begins by saying that he's going to use Minimum Description Length analysis to figure out what the words are in a corpus from which all the spaces have been removed. I had never heard of Minimum Description Length analysis (which henceforth I'll call *MDL*). I thought, "Oh my God, here's another dissertation coming from the AI lab at MIT, across the street from Building 20 where the linguists are, once my home. One more CS dissertation from people who know nothing about linguistics. Do I really want to read this?"

Well, I started reading it, and within a very short period of time—maybe by the end of the foreword or the acknowledgments—I realized that this guy knows a lot about linguistics. I started reading it and I could see that I didn't understand what was going on. I didn't understand how information theory, which seemed so disconnected from the empirical world (I thought that then, but I realize now how wrong that is), could possibly connect with data in a meaningful way and find linguistic structure: that was beyond my imagination. How could information theory possibly make any sense out of inferring what words are?

I got to the end of the dissertation, and knew that I didn't understand it, so I read it again. The second time, I understood a fair amount more. I kept rereading it until finally I said, wow. This was certainly one of the biggest *wow* experiences for me in my career as a linguist, comparable to encountering Chomsky's idea of an evaluation metric (which, I eventually realized, was a very closely related idea).

I need to explain what the central idea is behind MDL. Minimum Description Length is a framework for data analysis based on information theory and developed by the Finnish-American

statistician Jorma Rissanen, who passed away a few years ago (I was lucky enough to spend a few very interesting days talking with him at a workshop organized by the Brazilian mathematician Antonio Galves).

Let's say you have a set of data, and you want to find the best analysis of it. What a question! But that is the situation we are very often in. And the right way to think about that problem from an MDL point of view is this. You commit to a probabilistic analysis of the data, which means generating models that assign a probability to the data that you have.

All other things being equal, you prefer a model that assigns a higher probability to your data, because by virtue of assigning a higher probability, it is making an claim (of varying degrees of explicitness) that there is structure in the data. That is not an easy notion to grasp a first: that there is a very direct connection between developing a probabilistic model which assigns a high(er) probability to a set of data and having a model which recognizes a certain kind of structure in data when it finds it.

This leads directly to the work of linguists. Linguists often understand their job to be to discover, or to write, grammars, with the understanding that the structure of the grammar is the structure of the language. If all this is leading in the right direction, then a grammar should be very close to a probabilistic model that assigns a high probability to utterances consistent with the grammar.

If one is familiar with information theory, then one understands that the natural way to think of probability is not directly, but rather through the lens of inverse log probability (the logarithm of the reciprocal of the probability). This sounds bizarre to the uninitiated and yet is a totally humdrum workaday fact in the context of information theory! One of the outstanding consequences of this is that this quantity has the bit as its natural unit of measurement. When we compute the log probability of a data set with a particular model, that quantity can naturally be called the *information content* of the data, under the model.

Now we can get to the heart of MDL. MDL asks us to consider any and every (probabilistic) way that we can imagine to analyze our data set, and for each model to compute the information content of that data. To the information content we add the length of the model (written as the most economical computer program that we can find), which is also expressed in bits. These two terms are the data term and the model term, we might say, and their sum is the *description length* of the pair consisting of the data and the model.

With all that, MDL can express its idea simply: the best analysis of the data is the one for which the description length is the shortest.

This is a *wow* moment—it was a wow moment for me when I understood that. And I had understood all along (as I was studying this) that the measure of the length of the model was very, very closely linked to the notion of the evaluation metric in generative grammar. Finally a deep place could be found for why an evaluation metric based on the complexity of grammar was of importance for linguistics.

So this was what de Marcken's thesis introduced me to—Rissanen's MDL perspective (what a stroke of good luck for me that de Marcken chose to make his dissertation a model of expository elegance).²²

²² <https://arxiv.org/abs/cmp-lg/9611002>.

Carl de Marcken was in (my colleague) Partha Niyogi's generation at MIT, as was Michael Brent, who also worked on an MDL approach to the word discovery problem, and they were influenced by Bob Berwick, who was Carl's advisor.

When I understood de Marcken's MDL-based claim, I sat down to write my own program based on his algorithmic insights. The bottom line for me was that although his work was amazing, it was not going to provide real linguistic insight until and unless it began with a deeper understanding of the fact that there is such a thing as word-internal structure—that is, morphology—and word-external structure—that is, syntax. So I set myself the task of adding a minimal knowledge of the structure of morphology, with the intent of improving the performance of a de Marcken-style learning. I've been working on that project, on and off, ever since—and learned a whole lot about language and computation (and corpora) along the way. And, I should say, about empiricism.²³

I'm trying to tell a story today. Let's go all the way back to Zellig Harris, because the story certainly goes that far back. In Harris's (1951) *Methods in Structural Linguistics*, he tried to create foundations for the Sapir-Bloomfield conception of descriptive linguistics. To any careful reader it is abundantly clear that to Harris the central goal of the linguist is to create a compact grammar. There's no goal that he sets in that book that's more focused in his sights. There's no goal that is sought more fervently than that of a compact description.

But he never explains why. The people who reviewed his book threw up their hands and said, what is this? Why are you focusing on compact descriptions when we have absolutely no reason to think that languages focus on a compact description? Harris wasn't prepared to answer their questions. Yet in a sense the answer was perfectly clear: it was because he was trying to explain to people what the best scientific analysis is of a set of data, and he knew perfectly well that a compact description was what scientists are looking for in their scientific analysis of data. He couldn't carry that conviction an inch further beyond saying it as it did, though.

Well, Chomsky understood all this, at some level—and I would say, at a deeper level than Harris, in part because he was studying with Nelson Goodman. Goodman was a philosopher at Penn who Zellig Harris had approached, asking him if he, Goodman, would help mentor Chomsky as an undergraduate (and then later a master's degree student) at Penn. Goodman's whole professional career at that point had been based on pushing Rudolf Carnap's ideas about the foundations of scientific theorizing, and theory-construction.

Goodman's central idea involved formalizing and making explicit the notion of simplicity. I can't say that he made any progress that spoke to me—alas, I've tried to understand it. Instead, MDL is the inheritor of that perspective, in my opinion, which is *not* to say that there's a direct intellectual descendant from Carnap to MDL—not at all—but Carnap and Rissanen were in some fashion, or fashions, attacking the same problem.

Chomsky is carefully following both Harris, and his notion of compactness, and Goodman and his notion of simplicity. Chomsky both embraces what they are saying. He embraces it in the sense that he says, yes, we want our grammars to be as small as possible. But, he says, following Goodman, even the notion of shortest analysis is malleable: it is subject to the creativity of the

²³ I'm alluding here to Chapter 1 of *Empiricism and Language Learnability*, by Nick Chater, Alexander Clark, John Goldsmith and A. Perfors, Oxford University Press 2015.

scientist or philosopher. That is, a human being is involved in the decision as to what is the shortest analysis. You can change the notational system in order to radically change what is the shortest analysis.

This, by the way, is very different from the MDL notion of shortest analysis.²⁴ There are two things involved with figuring out what the shortest which is based on the choice of a Turing machine for determining what's shortest, which is open to the scientists' decisions to a small degree, but not to a large degree. On the one hand, Chomsky agrees with the notion that compactness is crucial, and he embeds it in his early notion of generative grammar. But, on the other hand, he adopts a notion of simplicity, which ultimately, he takes to be arbitrary. It can be informed by the linguist as researcher. For him, that's important, because he wants to insist that rule ordering be related to this notion of simplicity.

KI

One exception would be in his formal language theory, where the notion of power comes in, and this desire within this framework to stay as low as possible on the hierarchy. Because if you move towards the power of a Turing machine, then grammar is unlimited. So that's one exception to the idea that it's up to the philosopher or linguist.

JG

Yes, you're right. I wonder how much Chomsky really understood that at the time. That's in the late 1950s, when he was doing that work. Once he became an assistant professor at MIT, and he was working with the people at MIT.

I've been trying to reconstruct when I first read *Logical Structure of Linguistic Theory*. The book didn't come out until the fall of my last year at MIT.

KI

Do you mean it was a manuscript from 1950, and never published?

JG

No, I mean the book, because I never saw the manuscript.

KI

So it really was a manuscript for 20 years?

²⁴ Recall that with MDL, one has to provide a program that implements the model that is used for a particular "description length (given a model)". There are two ways in which finding the shortest program is not obvious. The first is that no matter how short the program is that we come up with at a given time, we can't be sure beyond a theoretical doubt that there isn't a shorter program, and the second is that the choice of the programming language we use to write the program is essentially bound to the choice of the Turing machine we use for the entire enterprise. While the choice of Turing machine can have an impact on the ultimate decision, the size of the impact is bounded, as I try to show or explain in "Towards a new empiricism for linguistics," in *Empiricism and Language Learnability* (OUP 2015) or at <https://people.cs.uchicago.edu/~jagoldsm/papers/empiricism.pdf>

JG

Right. I never saw it as a manuscript. I think it came out in the fall. The summer of 1975 I spent in Montreal—it was a 4-month summer, and I spent a lot of time thinking about the evaluation metric. I wrote some pages which were supposed to go into my dissertation, but never did, in which I presented an argument for autosegmental phonology based on an application of the evaluation metric to language learning in a completely theoretical way. In retrospect, Jim McCawley had already developed a similar idea, but not in the detail that I did it, because I actually came up with some numbers for what that's worth.

I sent this to Chomsky, mailing him stuff to his Cape Cod address over the summer, since he was on my dissertation committee. I recall he was really impressed; it was only years later that I understood why. And that is because Chomsky had basically given up on the evaluation metric, and he was going to abandon it four years later.

In *Aspects*, Chomsky has worked this out in outline what the evaluation metric was to do, though it's really not more than a few pages, and I think only a few people really picked up on it. The idea was in the *Aspects* model of generative grammar: the idea that the evaluation metric is the crucial tool used by the language learner to choose a grammar. The thing that most people never understood was that the theory of grammar could allow an infinite number of grammars compatible with the data, as long as the evaluation metric as part of the learner's attention could clearly focus on the simplest grammar inside the set of grammars that were compatible with the data.

That was the crucial element, having a device that could *find* the simplest grammars compatible with the data. It didn't matter what the mass of grammars were that were not compatible with the data, but were compatible with UG, with universal grammar. Didn't matter.

I remember just being blown away by that. (That must have been my first huge wow moment with generative grammar, and that was when I was a sophomore in college.) I understood that notion, in *Aspects*, and I spent a lot of time thinking about that in the summer of 1975, when I was writing my dissertation.

I read LSLT in the fall of 1975, and that seemed to me to be the beautiful face of generative grammar. That fall Chomsky started using the term *universal grammar*, which was another term he used was the initial state of the learner—this made a lot of sense to me.

Then when he gave it up with the “principles and parameters” approach, about two years later, it seemed to me that that was the beginning of the end of generative grammar as an interesting scientific approach to language. Yeah, once you start thinking about language acquisition as setting a small set of parameters to a small set of values, we've really given up on a serious theory of language, it seemed to me.

KI

There is a very unsatisfying aspect to all of this, and I wonder what your reflection on this is. If we give a starting date of the mid-1940s with Zellig Harris, we're in an 80-year history where true battle lines were drawn.

At a certain point there's a symbolic grammar approach in which discovery procedures were kind of impossible, and it's heretical to suggest that they could even exist, and that their existence just shows complete lack of understanding of what language is about. On the other side,

approaches that take any kind of deep delving into actual languages with actual phenomena to be a sort of non-technical dabbling into something that has a technical solution.

Here we are, 80 years later, at a point in this long history where research on deep learning and large language models have seemed to side with the notion of discovery procedure at a very, very deep level. Using a very tiny, agnostic little computation, repeated endlessly, a remarkable language model emerges. Oh, people can still thump their fists on the table and say, oh, AI doesn't know anything, and I just showed in an experiment that it doesn't know this, but the results are simply astonishing.

Of course there are challenges for the deep learning models. These modern systems are probably far larger than they ought to be. If only the battle lines weren't so stark and so inhospitable to efforts to communicate across them. And if the personalities involved weren't incredibly strong, and if each side didn't believe that the other is just a complete waste of time.

That's also a thread through all of your work. You're a major contributor to generative linguistics; since autosegmental phonology is theoretical linguistics. But you're also a major contributor to a neural network, dynamical system, and probabilistic approaches. I would say very few people have that kind of double life.

JG

Yeah, and I don't understand why. I don't understand why there aren't more people who look to machine learning to find ways to rethink how we write grammars. Not to limit what goes into grammar, but rather to rethink the ways grammars are written so that they are learnable.

What do I mean by that? I mean basic things like being able to think of grammars as existing in points in a high-dimensional space, with distances between grammars. Is that so hard? It's very much out of keeping with the old view of grammar, but it's very natural from the point of view of machine learning.

My view of all this was radically altered by the years I spent working on *Linguistica*, a software project that takes a raw corpus from a language as its input and infers the morphological structures—or some aspects of the morphological structure—of the language.²⁵

I've been working on this project on and off for more than twenty-five years, though I mostly worked on it from 1998 to around 2012. Most recently I've been using it to analyze the language of the Voynich manuscript, along with a colleague from Yale University, Claire Bowerman. (It continues to surprise me that my colleagues in linguistics do not know this work in computational linguistics, despite the fact that the citations of this publication are more than double that of the citations of any of my papers on phonology.)

This work has taught me what it means to think about learning a grammar from data. It does not mean what they used to call "expert systems" back in the 1970s and 1980s, when programmers would ask human experts what questions they ask when looking at data.²⁶ It means

²⁵ Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics* 27:2 pp. 153-198. Also https://github.com/JohnAGoldsmith/lxa5_0.

²⁶ A good example of that approach is B. Elan Dresher, and Jonathan D. Kaye, A computational learning model for metrical phonology, *Cognition* 34:2, 1990, 137-195.

finding fundamentally new ways to think about language, using insights that machine learning can lend us.

As I alluded to earlier, the central idea that I find myself being drawn to again and again in this work is the notion of finding the simplest grammar that describes the data, within the limits set to the grammar as a whole. It's the intuition that Zellig Harris had; it's the picture that Chomsky had up through *Aspects*. Harris tried to make it work with computers; Chomsky did not.

And now we're in a brave new world. If I were twenty years younger, there's no question in my mind that I would be focused on developing ways of understanding grammar as reflecting the geometry of deep learning systems. At this point, I leave it to others to do that.

Over the last ten years, I've spent a good deal of time writing about the history of linguistics and its relation to its neighboring disciplines—something that I think is very important for any working linguist, and something that becomes increasingly important to one as one spends more time watching a field evolve.

It astonishes me to realize that the work I did on autosegmental phonology began more than fifty years ago now. When I think back on the mid 1970s and realize what had happened fifty years before then, it gives me pause. I'm pleased that others are still thinking about it today.

In SKASE Journal of Theoretical Linguistics [online]. 2025, vol. 22, no. 3 [cit. 2025-12-12]. Available on web page <http://www.skase.sk/Volumes/JTL60/07.pdf>. ISSN 1336-782X