Clause Subordination and Demography

Geoffrey Sampson

Sometime Fellow, The Queen's College, Oxford

To appear in Mouna Ayadi, ed., Approaches to Linguistic Complexity and Variation in Language Use*, 2027*

The complexity of conversation

My book *Complex Talk* (Sampson 2027) examines how the grammatical complexity of everyday conversational speech in England varies with demographic characteristics of speakers, such as their age, sex, level of education, and home region. It finds a number of statistically significant (sometimes highly significant) correlations which have not previously been suspected, to my knowledge, and some of which seem quite surprising.

The term "complexity" is used here in the schoolroom sense which distinguishes a *simple sentence*, consisting exclusively of a main clause, from a *complex sentence* which contains one or more subordinate clauses. Clause subordination is the central example of the phenomenon of grammatical recursion, which is widely seen as the crucial property distinguishing human language from communication systems used by other species. Thus, the very influential article by Hauser, Chomsky, and Fitch (2002) contains repeated remarks such as

> ... animal systems lack the rich expressive and open-ended power of human languages (based on humans' capacity for recursion)
>
> Why did humans, but no other animal, take the power of recursion to create an open-ended and limitless system of communication?

Consequently, complexity – the extent to which utterances make use of clause subordination – is an attractive precisely-quantifiable proxy for the vaguer idea of level of competence in a person's native language. Linguistic competence is commonly thought of as mastery of a set of linguistic rules, or parameter settings, but in practice we are never told just how many rules of grammar (or parameters) a given

language possesses in total (let alone what all of them are), so we cannot put figures on the proportion of language structures a given user has mastered, whether in comprehension or production. The average complexity of a speaker's utterances is of course not the *same thing* as his or her level of linguistic competence, but it is a phenomenon that lends itself to precise measurement and which common sense would suggest must be closely related to competence level. And it is a property which is as readily measurable in the context of casual, spontaneous conversation as in those of studied public speaking or carefully-edited written language; this matters, because it is the first of these kinds of language which seems likely to offer the most direct window onto a language user's competence.

There is a danger that I might be read as assuming that more complex utterances are "better" utterances – that grammatical complexity is desirable in its own right. Of course that is not so: whenever something can be said in simple grammar, simple is definitely best. But clause subordination allows many subtle and logically-complex things to be said that cannot be said, or cannot be efficiently said, using only simple grammar; and in a complex world we can expect that many such things will need to be expressed. So it is reasonable to expect that, in large samples, people capable of using complex grammar will have occasions to use it, making their average complexity indices higher than those of people less linguistically skilled.

Some generative linguists would reject any attempt to measure differences among adults' competence levels numerically, holding that individuals do not differ significantly in this respect. According to Chomsky,

> To a very good first approximation, individuals are indistinguishable (apart from gross deficits and abnormalities) in their ability to acquire grammar. ... Individuals of a given community each acquire a cognitive structure that is rich and comprehensive and essentially the same as the systems acquired by others. (1976: 144)

> ... every child ... acquires knowledge of his language, and the knowledge acquired is, to a very good approximation, identical to that acquired by others on the basis of their equally limited and somewhat different experience. (1975: 30)

But that idea has been exploded through experimental work by Ewa Dąbrowska (1997) and Ngoni Chipere (2001, 2003, 2009). Chipere (2009: 179) found

> that native speakers vary widely in their ability to comprehend complex

> grammatical structures ... [and that] grammatical competence, and not working memory capacity, was the source of the subjects' variations in ability to comprehend complex sentences.

The aspect of language usage for which these researchers have demonstrated large competence-level differences among speakers is the handling of clause embedding.

Measures of complexity

There are many ways in which one might calculate a numerical measure of the grammatical complexity of an utterance. Cheung and Kemper (1992) surveyed a range of metrics proposed by various scholars up to that date, and used experimental techniques to assess the adequacy of eleven such metrics. Rather reassuringly, they found that in practice conclusions about relative utterance complexity were not much affected by which particular measure was chosen.

When I first (Sampson 2001) investigated the issues now examined in *Complex Talk*, I used what seemed (and seems) to me the most obvious and straightforward metric, which I illustrate here via a simple example. Consider an utterance *no we don't like wasting food.* (Because the focus of the present chapter is spoken language, I avoid using features such as capitalization and punctuation marks which apply only to writing.) Any linguist, I believe, would recognize that this utterance has the following clause structure (I use square brackets to mark finite clauses and round brackets for non-finite clauses):

> no [we don't like (wasting food)]

The word *no* is outside any clause structure; the main clause is headed by the finite verb group *don't like*; the object of *don't like* is the present-participle clause *wasting food*. Each word of the utterance is assigned a "depth" figure, i.e. the number of clauses within which it is embedded, so we have:

no	we	don't	like	wasting	food
0	1	1	1	2	2

I define the *complexity index* of an utterance, a series of utterances, or a sample of the utterances of a speaker or a class of speakers, as ten times the mean of the depths of the words in the utterance or sample – the index of this example will thus be

10 × (7 / 6) = 11.67. (The reason for multiplying by ten is that, before multiplication, in practice complexity figures are commonly between 1.0 and 2.4 or so, so that the digit after the decimal point carries much more information than the initial digit; psychologically it is easier to think about differences between figures in the range 10 to 24 than to force oneself to attend to what follows the decimal point. This arithmetic operation is purely "cosmetic", and cannot affect any substantive findings about variations in complexity.)

As already said, in many cases choice of one complexity measure rather than another seems unlikely to make much practical difference. However, there is one way in which the metric defined above differs from the majority of those compared by Cheung and Kemper (and from those adopted by other psycholinguists since), and where, for the kind of data dealt with in *Complex Talk*, the difference matters.

The metric I use is word-based. Each word of an utterance is assigned a clause-depth figure; indices are assigned to utterances by averaging the depths of their individual words, and to speakers by averaging the depths of the words in a sample of their utterances. Many complexity measures in the literature, on the other hand, are sentence-based. Some suitable method is chosen for evaluating the complexity of a sentence, and overall measures of the complexity of multi-sentence passages are derived, by averaging or otherwise, from the figures for the sentences they comprise.

A sentence-based complexity metric may work well in connexion with written language, for instance in order to assess the reading difficulty of specimens of documentation. In written English the sentence is a unit clearly defined by capitalization and punctuation. However, as Miller and Weinert (2009) point out, in spontaneous spoken English "sentences" are not well-defined units. This is not mere pedantry: a sentence-based measure of complexity could give wildly different results depending on analytic questions that have no "correct" answers.

Consider again our very simple example *no we don't like wasting food*, and suppose it is uttered with a slight pause after *no*. Is it one sentence or two? – the question is meaningless. Now suppose we measure sentence complexity as the number of clauses contained in a sentence (there are many other ways it could be measured, but that is beside the point here). If we understand the example as *No. We don't like wasting food.* (i.e. as two sentences), then *No.* scores zero and the second sentence scores two (for the main clause and the *wasting* clause), and the average over the complete utterance is 1. As *No, we don't like wasting food.* (i.e. a single sentence), the score is 2. A metric which can yield a two-to-one difference (or more) depending on answers to meaningless questions is not acceptable. Even in speech, on the other hand, words are almost always well-defined. (There are occasional terms, such as *sea shell*, which might be considered two words or one hyphenated word, but this issue is

trivial relative to the debatability of sentence units.) Hence I was not tempted to abandon my word-based complexity metric.

Lifelong learning versus steady state

The linguistic issue which first led me to study numerical levels of complexity was the theoretical debate between "lifelong learning" and "steady-state" models of the mother-tongue acquisition process. Linguists of the nineteenth and early twentieth centuries assumed that mastering one's first language was a process that had no particular endpoint other than death. It might be that language learning proceeded faster in childhood than in adulthood, as seems to be true of many kinds of learning, but there was no point at which one ceased to be a learner. For instance, Leonard Bloomfield (1935: 46) wrote:

> there is no hour or day when we can say that a person has finished learning to speak, but, rather, to the end of his life, the speaker keeps on doing the very things which make up infantile language-learning

However, in 1967 the psychologist and neurobiologist Eric Lenneberg contradicted this, claiming that there is evidence for a "critical period" during childhood when language-acquisition has to take place if it is ever to take place successfully. Our faculty of language acquisition is biologically programmed to "switch off" at a certain age, perhaps around puberty, which among other things would explain why adult second-language acquisition is so relatively painful and imperfect. Noam Chomsky took up this idea of Lenneberg's and popularized it among the linguistics community. According to Chomsky, at the end of the "critical period" the language user enters a "steady state", in which his linguistic competence does not slow down but rather it altogether ceases to develop further, however long he lives:

> children proceed through a sequence of cognitive states ... [ending in] the "final state," a "steady state" attained fairly early in life and not changing in significant respects from that point on. When the child has attained this steady state, we say that he has learned the language. (Chomsky 1976: 119)

This steady-state idea appeared to me implausible and inadequately argued. In Sampson (2001), using data from the original British National Corpus of 1994 (see below), I found evidence which seemed to suggest (though at a low level of statistical

significance) that British speakers' complexity indices do continue to increase throughout adult life. But, as we shall see, the data of BNC1994 leave much to be desired; and with hindsight I believe my use of statistical methods in the 2001 article was questionable. So for the *Complex Talk* research I used newer data to reinvestigate the question, and found that for speakers over the age-range thirteen to eighty there is indeed a positive correlation between age and grammatical complexity index, with the probability of the null hypothesis as low as 0.017. That is, there is less than a one in fifty chance that the correlation is a merely chance property of the limited sample in my data, and does not reflect a property of the speech of the population as a whole. (After age eighty there is some suggestion of a drop-off in average complexity, though my data here are too sparse to be sure of this.)

However, the rate at which grammatical complexity increases after age thirteen is slow: in my data, 0.023 index points per year. So this finding implies that "lifelong learning" and "steady-state" models of first-language acquisition both contan a measure of truth. The data include too few under-13 child speakers to determine a precise trajectory for the growth of grammatical complexity through childhood, but we know that a newborn's complexity index is zero (babies do not speak, let alone speak in clauses), and the data show that the average index around age thirteen is in the region of 13 – so the *average* complexity growth rate up to that age must be about one index point per year. That rate is so much greater than 0.023 points/year that it seems there must be a rather abrupt inflexion point around age thirteen. The changing growth rates cannot be explained in terms of a continuous gradual decline in rate of increase from early childhood into middle age, as I previously imagined. Thus, generative linguists are correct to point to a sharp difference between a childhood "critical period" and adult life, but they are mistaken to see adulthood as a time when linguistic competence ceases to develop further.

Sources of data

In order to study relationships between grammatical complexity and age (or any other demographic characteristic), we need a representative body of accurately transcribed speech with information about speakers' ages and perhaps other demographic variables. Since transcribing recorded speech is very labour-intensive, we have few possible data sources.

The earliest English speech corpus I know of, the London–Lund Corpus compiled in the 1960s–70s (Svartvik and Quirk 1980), is in one respect unrivalled to this day: its representation of intonation patterns on utterances is superbly detailed.

Unfortunately it is useless for our purposes, since it made no attempt to achieve a representative set of contributors – the speakers appear mainly to be staff and students of University College London and their friends and families.

The first speech corpus to aim at a set of contributors representing all sections of British society, balanced by age, region, social class, and other demographic characteristics, was the "demographically sampled speech" section of the original British National Corpus published in 1994 (Burnard 1995). I used that resource for the research published as Sampson (2001); but I found its standards of accuracy to be quite problematic. A friend who witnessed the processing of sound recordings for BNC1994 has described a scenario in which low-level clerical workers transcribed recordings under time pressure, and in my experience the end-result is replete with the types of mistake which one would predict in such circumstances.

Not infrequently it is obvious that the words as transcribed cannot be what the speaker said. In a discussion between a mother and daughter about unsatisfactory child-minders, file KE6 makes one speaker say *... unless you've low and detest children or unless you're out purely and simply for the money ...* . The sequence *you've low and detest* is word salad, in reality the speaker must have used the standard piece of hyperbole *you loathe and detest*. And there is a suspiciously low incidence of speech edits in the BNC1994 transcriptions, suggesting that the transcribers often wrote what they heard the speaker as trying to say rather than what he actually said (the former being what a secretary taking dictation is expected to do).

Even with respect to attribution of utterances to speakers, and identification of speakers' demographic characteristics, there are absurd, obvious errors in BNC1994. File KDE contains a conversation between a 34-year-old engineer and his three-year-old son; one exchange between them runs *Did you want to have a shower with Daddy? — Umm yes*. The corpus assigns the question to the child and the response to the father. The occupation of a 44-year-old woman in file KDL is given as "doctor", and her social class in a four-level version of the standard occupation-based "ABC1" system is given as DE, semi-skilled or unskilled. Happily for British patients, semi-skilled and unskilled people do not practise as doctors. Any doctor is a member of one of the top top two social classes, A or B (Market Research Society 2010).

When the incidence of obvious errors is so high, one can feel fairly sure that there are also plenty of non-obvious errors. One feels that with a little quality assurance, many of these errors should have been avoidable.

A second attempt at a spoken British National Corpus was produced in 2014 (Love et al. 2017), and this is a great improvement over BNC1994. Spoken BNC2014 is the source used for the *Complex Talk* research.

Spoken BNC2014 is not a perfect research resource. Although its coverage of

the various regions of England is fairly comprehensive, the compilers admit (Love et al. 2017: 327) that the other UK nations (Scotland, Wales, Northern Ireland) are inadequately covered. But this is a deficiency only relative to the name "British". If one chooses to think of the population represented by the Corpus as that of England rather than Great Britain or the UK, the problem disappears (there are some speakers from the Celtic nations in the Corpus, but then for an inhabitant of England there is nothing unusual about hearing an occasional Scottish or Welsh voice). England is arguably more suitable than Britain as a unit for this kind of survey: a language is a social rather than a political institution, and Scotland in particular is far more separate from England as a society than as a polity (Ratti et al. 2010).

Another shortcoming in representation would be much more problematic for other kinds of linguistic research than it is for that of *Complex Talk*: few Spoken BNC2014 contributors are children. (There is internal evidence that the requirement of releases signed by responsible adults for recordings of children was sometimes interpreted as a *de facto* ban on such recording. The Corpus does contain quite a lot of child speech, but almost all of it is by the same two children.) Luckily, because the initial impetus for the *Complex Talk* work was to discover whether the generative steady-state model of language acquisition is correct, child speech was irrelevant: we know that children's competence advances rapidly, the interesting question is whether it ceases to advance in adulthood. Once the reality of a discontinuity between childhood and adult grammatical growth had been established, subsequent analyses focused on the usage of "linguistic adults" – speakers between thirteen and eighty.

One consideration motivating the creation of a new Spoken BNC2014 alongside its 1994 predecessor was to have a more up-to-date resource, but this consideration may have weighed more with funding agencies than with linguistics researchers. Languages change rapidly with respect to vocabulary and, less rapidly, grammar (consider e.g. the recent replacement, even in published writing, of the word *might* by *may* as its own past tense), but there seems little reason to expect the incidence of clause subordination to change much from one decade to the next – had it been more representative, I would have been happy to use London–Lund for the *Complex Talk* research. In any case, by the time of writing this, Spoken BNC2014 is itself more than a decade old – but then the research reported in *Complex Talk* has itself taken a large fraction of that time. To produce non-superficial research findings from thoroughly up-to-date language data would be a vain ambition.

The real reason for basing the present research on Spoken BNC2014 (and, I suspect, the main reason why its compilers felt the exercise worthwhile) is that the accuracy of its speech transcription is on an altogether higher plane than the 1994

corpus. It does not include intonation data, parallel to London–Lund, but for identifying clause structure intonation has little relevance. In other respects the care taken to produce precise records of the wording actually uttered by the various contributors, and to adopt disciplines ensuring consistency across the Corpus (for instance "filled pauses" – ums and ahs – are required to be spelled in one of a canonical set of representations, rather than at the whim of the transcriber), must be strikingly apparent to anyone who works intensively with it. As an accurate and representative sampling of conversational speech across England, Spoken BNC2014, referred to for short below as "the Corpus", is unrivalled and unlikely to be rivalled in the foreseeable future.

It must also be said that the BNC2014 speakers' demographic characteristics are not always logged to the same high standards of scholarly accountability as its speech transcriptions, even if much better than BNC1994; but that is a problem one can work round. By far the most important requirement on a speech corpus (and the most difficult to achieve) is that its speech should be accurately transcribed.

The Grammage database

For *Complex Talk* it was necessary to mark up clause boundaries in a sample of the Corpus. This is slow, painstaking work, so annotation of the entire Corpus was out of the question. I do not doubt that some IT mavens would suggest achieving this via parsing software, but I would regard such an exercise as worthless. Any parsing software will have its own hidden biases, and since the grammatical-complexity differences which emerge from carefully manually-annotated material are often quite delicate, they woud be very likely to be swamped by the biases introduced by software. Analyses of automatically-annotated speech would give us information about the software used, rather than information about human speech, and it goes without saying that the latter is what we ought to want.

Accordingly I constructed a database consisting of manually-annotated samples of the Corpus; since the initial purpose was to look at variation of grammatical complexity with age, I called this the "Grammage" database. (The eventual *Complex Talk* research went on to examine many demographic variables in addition to age.)

Most of the Grammage files consist of annotated stretches of a thousand-plus words extracted from the middle of Corpus files (avoiding beginnings and ends of those files, where conversations are often preoccupied with the mechanics of the recording process). At first I chose Corpus files at random, though ignoring any files

in which one or more speakers' ages were unknown, together with files involving more than three speakers (because the Corpus compilers warn that with those recordings their assignment of utterances to individual speakers was unreliable) and files with an unusually high incidence of wording too unclear to have been reliably transcribed. As the work proceeded I biased file selection in the attempt to compensate for imbalances in the Corpus – very many Corpus files include speakers in their twenties, so after a certain point I rejected any further such files; and I ensured that Grammage included every child speaker in the Corpus, since there are few. Towards the end, in order to improve the representation of under-represented classes of speaker, I annotated just the wording uttered by such speakers, excluding their conversational partners' utterances from the annotation and the subsequent analyses. I aimed for a database comprising at least 200,000 words.

The eventual Grammage database comprises 208,526 words from 240 identified speakers, amounting to 43 per cent of all identified Corpus speakers. ("Identified" speakers, because some files include a certain amount of wording from individuals who were not part of the data-gathering exercise, for instance one conversation is between two women having a "girls' lunch out" who are briefly interrupted by a waiter taking their order – the waiter was not invited to provide his demographic characteristics or sign a release form, and his words are ignored in the *Complex Talk* analyses.) The Grammage files are freely available for downloading from <www.grsampson.net/GrammageFiles.tgz>; as time permits, I shall add copies of software I wrote to analyse the material, together with documentation and intermediate results of analysis, enabling others to check the correctness of my results and to produce further results of their own.

Jane Edwards (1992: 139) made the wise remark:

> The single most important property of any data base for purposes of computer-aided research is that *similar instances be encoded in predictably similar ways*

(her italics). My SUSANNE scheme of grammatical annotation (Sampson 1995) has been recognized internationally (e.g. Lin 2003: 321) as particularly strong with respect to tight definition of analytic categories and their boundaries, and experimental research is available (Sampson and Babarczy 2008) on how closely its precision approaches the ideal maximum. The Grammage sample is annotated according to this scheme, but ignoring all structure other than clause boundaries, with clauses distinguished only as [finite] versus (non-finite), together with speech edits – false starts, either followed by replacement wording, or broken off without replacement –

which are enclosed in < angle brackets > and ignored in subsequent analysis. (I distinguished finite from non-finite clauses because, at the outset of research, I expected the former to need to "weigh" more than the latter in calculating a meaningful index of speech complexity. Early experimentation, however, suggested that this was not so, and the results reported in *Complex Talk*, and in some cases below, were calculated treating the two kinds of bracketing as equal.)

The complicated XML markup in the Corpus files is replaced in Grammage by a user-friendly notation, in which, for instance, truncated words are marked by a final hyphen, and many irrelevant phenomena covered by the XML notation are ignored.

Spontaneous does not imply simple

Written language, where the writer can unhurriedly consider his wording and if necessary edit it into better shape, can sometimes involve numerous layers of subordinate clauses embedded within subordinate clauses. But one might imagine that spontaneous conversation would always be grammatically fairly simple. Dąbrowska (1997) is an outstandingly insightful and logically watertight empirical study of competence in handling certain types of grammatical complexity, yet even Ewa Dąbrowska treats as truisms the statements "Unplanned informal discourse is very simple syntactically", and "normal everyday language ... tends to be very simple syntactically". The special kinds of syntactic complexity which Dąbrowska studied surely are rare in informal talk, but it is just not true that everyday speech is always devoid of syntactic complexity in the sense of deeply-recursive clause subordination.

The speaker with the highest index (23.4) in the Grammage sample is a 25-year-old female graduate, living in Oxford and working as a publisher's marketing executive. She has been off sick through work stress, and her conversational partner, a woman friend, has just said *you're getting a bit better and you were worked too hard but if you were crying in a car because you didn't want to go back to work you should not be spending your twenties crying in a car hating your job* – to which the marketing executive responds:

> yeah exactly (doing that) no of course not [but it's also that kind of thing [that it's more [that < I've got to > I feel [that I've got (to prove [that I'm well enough (to do the work [*that I've got*] and feel comfortable within it myself)] but also from that lesson be like no [I'm not doing that])]]]]

The emotional charge here might be expected to make it particularly difficult for a speaker to develop intricate grammatical structure successfully, but in fact the words

I have italicized, *that I've got*, are eight clauses deep; and, once we exclude the angle-bracketed false start earlier in the utterance (where the speaker replaces *I've got to* with *I feel that I've got to*), what remains is well-formed structurally and makes good sense.

An embedding depth of eight is extreme, but it demonstrates that there is plenty of scope for variation in mean conversation complexity among different individuals and different demographically-defined classes of individuals.

The findings are reliable ...

It should be emphasized that the correlations between grammatical complexity and demographic variables revealed by the research described in *Complex Talk* are not mere vague tendencies: they are solidly statistically-significant findings. In some cases the findings achieve levels of significance which are not routinely encountered outside the physical sciences.

For instance, one variable considered is the length of individuals' formal education. For each Corpus participant we are told whether he or she left school at the minimum legal leaving age or stayed on into the sixth form, and whether he or she gained a bachelor's degree and, if so, also a postgraduate qualification. In Britain the legal age for starting school is five; the minimum leaving age was raised in 1972 from fifteen to sixteen, so, depending on their date of birth, Corpus participants who left school at the minimum age will have received either eleven or twelve years of formal education, while an education ending after the sixth-form stage will have lasted fourteen years. Most first-degree courses in Britain last three years, so a graduate can be estimated to have had seventeen years' education. Postgraduate qualifications can take different lengths of time, from one-year master's courses to at least three years for a PhD; the Corpus does not specify which particular type of postgraduate qualification a given contributor had, but it is reasonable to estimate two years as a compromise between the various types, in which case we would attribute nineteen years of education to holders of postgraduate qualifications.

Looking only at Grammage speakers at least 25 years old (by which age their formal education is likely to be complete), and adjusting their individual complexity indices by subtracting the complexity increase expected as a consequence of age alone, average indices for the five lengths of formal education are plotted in Figure 1. As can be seen, most of the data points fall very close to the overall line of best fit or "regression line" – Pearson's coefficient of correlation, r, has the high value of 0.969. (The symbol r always has a value between one and minus-one: $r = 1$ means that all

data points fall exactly on the regression line and that the correlation is positive; *r* = minus-1 means the same but the correlation is negative, so that increase in one of the correlated variables goes with decrease in the other; and *r* = 0 means that the variables represented by the two axes are completely unrelated.) The value *p*, the probability of observing a fit as close as we have here on the null hypothesis that, in the population at large, complexity indices are unrelated to length of education, is 0.007 – less than one chance in a hundred that the correlation is spurious.

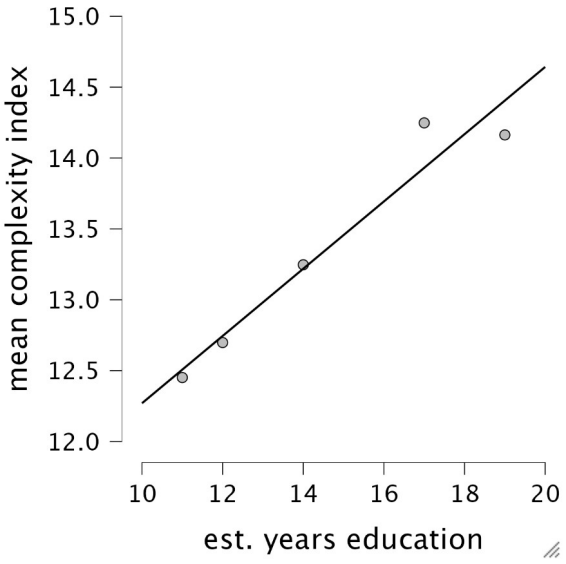

Figure 1

The points for holders of first degrees and of postgraduate qualifications in Figure 1 are less close to the line of best fit. But it is not too surprising to see that adding a postgraduate qualification to a first degree seems to have no effect on grammatical complexity: the activities leading to postgraduate qualifications have traditionally, in England, been rather different in kind from education up to the BA level, which is often described as less about teaching students "*what* to think" than teaching them "*how* to think". Over recent years the English higher education system has radically reshaped itself, under American influence and under pressure to develop new income streams for universities, but for much of my lifetime (and many Corpus contributors were older than me) postgraduate qualifications have tended either to be about "training" rather than "education" – teaching students the facts and skills needed to function in particular professional careers, rather than "how to think" – or else to be essentially honorific in nature. The MA degree was traditionally required for entry to the university-teaching profession, but it was not awarded on the basis of taught courses and exams: it was merely a certificate that the holder, in addition to

having a BA, was now old enough to be taken seriously as a teacher (a system dating from centuries when undergraduates were often much younger than today). I converted my 1965 Cambridge BA to an MA simply by waiting two years and paying a £2 administration fee, and, later, to a PhD by showing the University a sample of my publications. No teaching or examining was involved; my only contact with Cambridge University at the time was to apply for the further degrees.

Where postgraduate qualifications did, or do, involve studying and passing exams, they may be very valuable to their holders and to society. But, at least until the recent reorganization of higher education, there was good reason to see them as something different in kind from education up to the bachelor's level – in which case it could easily be that one of these categories has a relationship with grammatical complexity of speech while the other does not. So it would be interesting to see how Figure 1 changes, if the "length of education" variable is redefined to count only years of education up to bachelor's level, so that the Grammage talkers we have been crediting with nineteen years' education are instead counted as having received seventeen.

The result is Figure 2. I find the closeness of fit in Figure 2 astonishing. The coefficient *r* has risen to 0.999; $p = 0.001$, just one chance in a thousand that the correlation is spurious. Figure 2 is telling us that each additional year of education up to first-degree level is equivalent, in terms of its effect on the grammatical complexity of conversational speech, to an additional thirteen years of growing older. (Here I am assuming that the arrow of causality runs from education to speech complexity, rather than *vice versa*, or than from some third factor to each of these separately.)

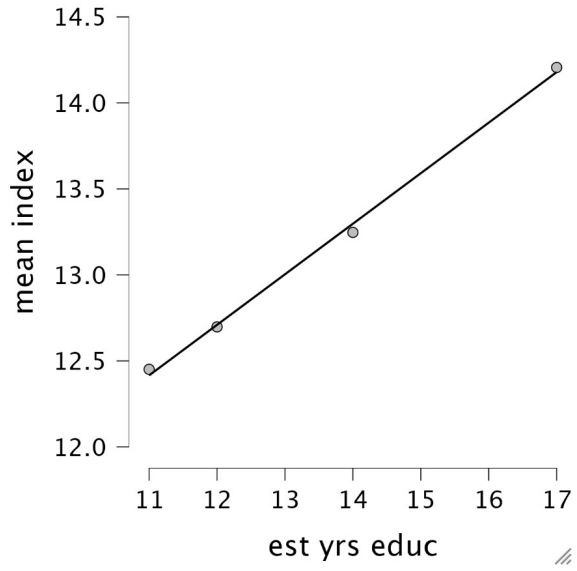

Figure 2

A correlation coefficient of 0.999 is so high that I wondered whether there might be some flaw in my calculations, so that the mathematical operations I had applied to my data were somehow bound to deliver a near-perfect fit of data points to regression line, whatever the data. But I could find no flaw. I believe Figures 1 and 2 represent a genuine discovery about the effects of education on conversational speech.

... but they are sometimes baffling

While various observed correlations between demographics and grammatical complexity are well founded statistically, some of them are extremely difficult to explain.

For instance, the data show a tendency for people living further south in England to have higher complexity indices. England is a highly centralized society in which it often feels as though everything that really matters happens in London, which could be a reason for London speech to be on average more logically complex than speech in the "provinces"; and since London lies towards the south coast of England, my first thought was that the tendency for southern speech to be more complex than northern somehow had to do with proximity to the capital (even though it was quite hard to imagine how the influence of London on speech could be stronger at, say, 250 miles than at 300 miles – television and other communications media are not weakened by distance, as gravitational forces between heavenly bodies are). But plenty of England, including the entire south-western peninsula, lies further south than London. Detailed comparison of the figures shows that the correlation is with latitude rather than with London distance.

The climate in southerly parts of Britain is more clement than further north, so (grasping at straws) I wondered whether there might be a tendency for people to prefer living towards the south, with those having greater linguistic competence tending to find more opportunities to move. But the Corpus records birthplaces as well as places of residence at the time of data collection, and any tendency in these data for current residences to lie further south than birthplaces is far too slight to account for the correlation mentioned.

So far as the evidence goes, it appears that simply living further south, whether since birth or more recently, raises a person's grammatical complexity index: each degree of latitude further south is associated with an extra 0.56 on one's index, $p = 0.002$. (The latitude difference between the Lizard in the extreme south and the

Scottish border at Berwick is not quite six degrees.) Putting it another way, one degree of latitude is equivalent, with respect to grammatical complexity, to 25 years of growing older. But I am entirely at a loss to understand how or why such an effect could arise.

Conclusion

The *Complex Talk* research revealed further relationships between grammatical complexity and various demographic characteristics of speakers. I do not list these here, but some of them seem as remarkable in their various ways as those detailed above. The relationships are not always straightforwardly linear correlations, like those I have discussed here.

This research area has turned out to be richer in interesting linguistic phenomena than I initially supposed possible.

References

Bloomfield, L., 1935. *Language*. Allen & Unwin.

Burnard, L., 1995. *Users Reference Guide for the British National Corpus*, version 1.0. Oxford University Computing Services. The current version is available from <http://www.natcorp.ox.ac.uk/docs/userManual/>.

Cheung, H. and Susan Kemper, 1992. Competing complexity metrics and adults' production of complex sentences. *Applied Psycholinguistics* 13.53–76.

Chipere, Ng., 2001. Native speaker variations in syntactic competence: implications for first language teaching. *Language Awareness* 10.107–24.

Chipere, Ng., 2003. *Understanding Complex Sentences: native speaker variation in syntactic competence*. Palgrave Macmillan (Basingstoke).

Chipere, Ng., 2009. Individual differences in processing complex grammatical structures. In G.R. Sampson, D Gil, and P. Trudgill, eds, *Language Complexity as an Evolving Variable*, 178–91. Oxford University Press.

Chomsky, A.N., 1975. *The Logical Structure of Linguistic Theory*. Plenum (New York).

Chomsky, A.N., 1976. *Reflections on Language*. Temple Smith.

Dąbrowska, Ewa, 1997. The LAD goes to school: a cautionary tale for nativists. *Linguistics* 35.735–66.

Edwards, Jane, 1992. Design principles in the transcription of spoken discourse. In J. Svartvik, ed., *Directions in Corpus Linguistics: proceedings of Nobel Symposium 82*,

129–44. Mouton de Gruyter (Berlin).

Hauser, M.D., A.N. Chomsky, and W.T. Fitch, 2002. The faculty of language: what is it, who has it, and how did it evolve? *Science* 298.1569–79.

Lenneberg, E.H., 1967. *Biological Foundations of Language*. Wiley (New York).

Lin, D., 2003. Dependency-based evaluation of Minipar. In Anne Abeillé, ed., *Treebanks*, 317–29. Kluwer (Dordrecht).

Love, R., et al., 2017. The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22.319–44.

Market Research Society, 2010. *Occupation Groupings: a job dictionary*, 7th edn. Market Research Society.

Miller, J. and Regina Weinert, 2009. *Spontaneous Spoken Language: syntax and discourse*. Oxford University Press.

Ratti, C., et al., 2010. Redrawing the map of Great Britain from a network of human interactions. *PLoS One* Dec. 2010, vol. 5, issue 12, e14248.

Sampson, G.R., 1995. *English for the Computer: the SUSANNE corpus and analytic scheme*. Clarendon Press (Oxford).

Sampson, G.R., 2001. Demographic correlates of complexity in British speech. In Sampson, *Empirical Linguistics*, 57–73. Continuum.

Sampson, G.R., 2027. *Complex Talk: quantifying grammatical recursion in English conversation*. Bloomsbury.

Sampson, G.R. and Anna Babarczy, 2008. Definitional and human constraints on structural annotation of English. *Journal of Natural Language Engineering* 14.471–94. A version is reprinted as "The bounds of grammatical refinement" in Sampson and Babarczy, *Grammar Without Grammaticality*, De Gruyter (Berlin), 2014.

Svartvik, J. and R. Quirk, 1980. *A Corpus of English Conversation*. C.W.K. Gleerup (Lund).