Version de travail

Publiée dans :
Léon J. 2007. « Empiricism versus Rationalism revisited. Current Corpus Linguistics and Chomsky's arguments against corpus, statistics and probabilities in the 1950-1960s » In S. Matteos & P. Schmitter (eds.) *Linguistische und epistemologische Konzepte — diachron*. Münster: Nodus Publikationen:157-176.

Empiricism versus Rationalism revisited. Current Corpus Linguistics and Chomsky's arguments against corpus, statistics and probabilities in the 1950-1960s

What is currently called 'Corpus Linguistics' covers various heteregeonous fields ranging from lexicography, descriptive linguistics, applied linguistics - language teaching or Natural Language Processing - as well as domains where corpora are needed because introspection cannot be used, such as studies of language variation, dialect, register and style, or diachronic studies. The sole common point to these diverse fields is the use of large corpora of texts or spontaneous speech, available in machine-readable form.

Current Corpus Linguistics stems from British corpus-based research, which is the oldest and the best organized with international corpus projects, a journal and many collective books. In order to legitimate their claim to be an autonomous and unified linguistic field in spite of their noticeable heterogeneity, and to ensure their theoretical position, corpus linguists developed arguments against Chomskyan generative grammar. Actually two types of stances can be observed :

- the first one is anchored in the Firthian tradition of empiricism which has never ceased to discuss Chomsky's views since the 1960s.
- the second one, though also stemming from the London School, is a reconstruction : Chomsky's arguments against corpora and statistics dating back to the 1950-1960s have been used in the 1990s to legitimate corpus-based research as a new linguistics[1].

In this paper, I will first examine Chomsky's arguments against corpora, probabilities and statistics ; then I will consider how these arguments have been taken up by present corpus linguists to legitimate their claims.

1. Chomsky's position on corpora, statistics and probabilities

1.1. Corpora and inductive methods

First let us recall that corpora were at the core of the Neo-bloomfieldian approach. For Neo-bloomfieldians, linguistic theory should aim at a systematic taxonomy of linguistic elements (distributional classes), by using a set of methods belonging to empiricism, that is the use of inductive discovery procedures from a corpus of observed data.

Against this approach, Chomsky (b.1928) argued that these procedures only revealed surface phenomena since they yield no more than a static inventory of signs, devoided

[1] In fact, if we consider works outside the British tradition, there is a third stance, more recent, and which we will not address in this paper. This position advocates a reconciliation between algebraic grammars and probabilistic grammars or between formal linguistics and information theory.

of any significance and theoretical explanatory adequacy. For him, the description obtained by this method was limited to the data which had been collected, and did not lead to any insight into the nature of language (see in particular Chomsky 1964).
It should be said that Chomsky did not deny the use of corpora when the aim is to describe specific languages such as Amerindian languages. However corpora of observed data should play a specific role within the generative machinery, as can be seen in a debate with the Neo-Bloomfieldians, when he was invited by Archibald Hill (1902-1992) at the University of Arizona in 1958.
To study a language which he does not know, the linguist should start from a natural corpus of sentences provided by an informant. In a second step, a second corpus will be generated by the grammar. This new corpus, containing ill-formed as well as well-formed sentences, should be tested by the informant in order to validate the grammar, and therefore the linguistic theory :

(1)
Suppose I am working with an informant in a language which I do not know. I have gotten from the informant responses that tell me some formulations or guesses were good, some were not good. I also have in mind, from some source, a general theory of linguistic structure. This tells me what is the general form of grammars. I will revise my general theory whenever it turns out that there is a better formulation. As the result of a lot of operating with the data which I have now collected, I come out with a theory, a grammar of the proper form, which fits in with my general conception of grammatical forms. My grammar tells me that some things should be sentences, some should not. I go back to my informant, and try them out. If the informant agrees with my predictions, then I am content. How I got the theory in the first place is something I don't know. This is not properly a question belonging to the field of linguistics, it seems to me. (Chomsky, 1962 : 175).

As can be seen, the corpus appears twice in Chomsky's hypothetico-deductive machinery: first as the input which should be analyzed by the theory, and second as the output generated by the grammar. Only the first corpus is 'natural'. Besides, Chomsky's conception of corpora involves several options : language is non finite, unlike the Neo-Bloomfieldian conception of language as a finite set of utterances ; any corpus should be projected by the grammar ; lastly language is innovative, according the principle of linguistic creativity.
The projection of the corpus by the grammar appears as soon as 1956, in one of his first papers "Three models for the description of language" :

(2)
Similarly, a grammar is based on a finite number of observed sentences (the linguist's corpus) and it 'projects' this set to an infinite set of grammatical sentences by establishing general 'laws' (grammatical rules) framed in such hypothetical constructs as the particular phonemes, words, phrases, and so on, of the language under analysis. A properly formulated grammar should determine unambiguously the set of grammatical sentences. (Chomsky, 1956 : 113).

Chomsky also presented his views on projection in *Syntactic Structures* ; it is grammar that projects the finite corpus of observed utterances to a set of infinite grammatical utterances :

(3)

First, it is obvious that the set of grammatical sentences cannot be identified with any particular corpus of utterances obtained by the linguist in his field work. Any grammar of a language will *project* the finite and somewhat accidental corpus of observed utterances to a set (presumably infinite) of grammatical utterances. In this respect, a grammar mirrors the behavior of the speaker who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences. (Chomsky, 1957 : 15)

Chomsky borrowed the notion of projection from Nelson Goodman (1906-1998)[2]. In his book *Fact, Fiction and Forecast*, first published in 1955, Nelson Goodman proposed a theory of projection according to which properties can be projected by induction from a sample to the general population. Projection belongs to predictive methods, and Chomsky suggested that the set of infinite grammatical sentences be projected by the grammar from the input (i.e. the corpus of observed data).

Bourdeau (1979) shows how taxonomists, in particular Charles F. Hockett (1916-2000), had already identified the issue of projection. See Hockett (1948) and (1954):

(4)

The task of the structural linguist, as a scientist, is as Preston implies, essentially one of classification. The purpose, however, is not simply to account for all the utterances which comprise his corpus at a given time ; a simple alphabetical list would do that. Rather, the analysis of the linguistic <u>scientist</u> is to be of such a nature that the linguist can account also for utterances which are <u>not</u> in his corpus at a given time. That is, as a result of his examination he must be able to predict what <u>other</u> utterances the speakers of the language might produce, and, ideally the circumstances under which those other utterances might be produced. (Hockett, [1948], 1957 : 279). [underlied by Hockett himself]

Besides, Hockett adressed the issue of projection in association with the necessity of tests of acceptability performed by native speakers:

(5)

… one must be able to generate any number of utterances in the language, above and beyond those observed in advance by the analyst - new utterances most, if not all, of which will pass the test of casual acceptance by a native speaker. (Hockett, 1954 : 232)

Therefore the notions of corpus projection and of testing by native speakers appeared first in Hockett's work. Still the fact remains Chomsky was the one who linked these notions with the idea of infiniteness and innatess of language, with some consequences on language creativity.

Moreover, the Neo-bloomfieldians disagreed with Chomsky on the adequacy of natural corpora. This was discussed too when Chomsky was invited at the University of Arizona. He argued that any natural corpus is skewed and cannot be generated, since it may produce non-sentences (ill-formed sentences) or be incomplete :

(6)

[2] On the notion of projection borrowed by Chomsky from Goodman, see Bourdeau (1979).

HILL : It seems to me that if I were working with transformations, I would first select a representative sample of English sentences for my corpus. I would then try to see if by selection of kernel sentences within the corpus I could then generate the whole of the corpus. This is all that I would do.
CHOMSKY : It is almost impossible to generate a corpus without going beyond it. Any natural corpus will be skewed. Some sentences won't occur because they are obvious, others because they are false, still others because they are impolite. The corpus, if natural, will be so wildly skewed that the description would be no more than a mere list.
HATCHER : I have a corpus of about one hundred and twenty-five thousand sentences, and I do not find that it is skewed.
CHOMSKY : But you do not have a machine which generates all of your sentences. I don't believe you could get a machine which would generate just these sentences. If you want to generate just the corpus and nothing beyond it, it would be a miracle if you could give any description shorter than the corpus itself.
(Chomsky, 1962 : 159f.).

In Chomsky's second response, it is worth noting his argument against inductive discovery procedures which, as we know, were dear to Neo-Bloomfieldians.
Further arguments against this kind of procedures, and more generally against empiricist methods, were developed later in *Aspects* (1965), by the time Chomsky introduced the distinction between competence and performance. Corpora are useless to study competence. Neither observed data nor inductive procedures from observed data will provide reliable information on the linguistic intuition of the speaker :

(7)
There is first of all, the question of how one is to obtain information about the speaker-hearer's competence, about his knowledge of the language. Like most facts of interest and importance, this is neither presented for direct observation nor extractable from data by inductive procedures of any known sort. … There are, in other words, very few reliable experimental or data-processing procedures for obtaining significant information concerning the linguistic intuition of the native speaker. (Chomsky, 1965 :18f.).

1.2. Linguistic creativity, memory and innateness

Linguistic creativity appears in 1956 and is defined as the ability of a native speaker to produce or understand new sentences and to reject ungrammatical sentences :

(8)
In other words, linguistic theory attempts to explain the ability of a speaker to produce and understand new sentences, and to reject as ungrammatical other new sequences, on the basis of his limited linguistic experience. (1956 : 113).

In excerpt (3), it could be seen that, in *Syntactic Structures* corpus projection is strongly connected with linguistic creativity, and that this ability is infinite.

(9)
In this respect, a grammar mirrors the behavior of the speaker who, on the basis of a finite and accidental experience with language, can produce or understand an indefinite number of new sentences. (Chomsky, 1957 : 15)

Later in the text, Chomsky claims that frequency of use does not take any part to the recognition of grammatical sentences. Hence linguistic creativity is independent of frequency :

(10)
In the context 'I saw a fragile-' the words 'whale' and 'of' may have equal (i.e., zero) frequency in the past linguistic experience of a speaker who will immediately recognize that one of these substitutions, but not the other, gives a grammatical sentence. (Chomsky, 1957 : 16).

Linguistic creativity also appears a little later in Chomsky's review of B. F. Skinner's *Verbal Behavior* (1959). He specifies that the faculty of recognize grammatical sentences is not formal, nor semantic, nor statistical, but belongs to infinite linguistic creativity where remembrance is of no use :

(11)
We constantly read and hear new sequences of words, recognize them as sentences, and understand them. It is easy to show that the new events that we accept and understand as sentences are not related to those with which we are familiar by any simple notion of formal (or semantic or statistical) similarity or identity of grammatical frame. (Chomsky, 1959 : 56).

However, it is only in 1962 at the 9th International Congress of Linguists, that his views on linguistic creativity became central to his linguistic theory. As can be seen in excerpt (12), some fundamental options of Chomsky's linguistic theory are linked to linguistic creativity, such as the infiniteness and the innateness of the faculty of language. Besides, Chomsky insists on the ability of hearers not only to identify deviant sentences and but to give them an interpretation[3]. Finally, in the name of linguistic creativity and quoting Hermann Paul, Chomsky only allotts very small place to memory (rote recall) and learning by heart in the use of language :

(12)
The central fact to which any significant linguistic theory must address itself is this : a mature speaker can produce a new sentence of his language on the appropriate occasion, and other speakers can understand it immediately, though it is equally new to them. … Normal mastery of a language involves not only the ability to understand immediately an indefinite number of entirely new sentences, but also the ability to identify deviant sentences and, on occasion, to impose an interpretation on them. It is evident that rote recall is a factor of minute importance in ordinary use of language, that ' a minimum of the sentences which we utter is learnt by heart as such – that most of them, on the contrary, are composed on the spur of the moment ', and that ' one of the fundamental errors of the old science of language was to deal with all human utterances, as long as they remain constant to the common usage, as with something

[3] See Joseph (2003) for his analysis of the asymetry of Chomsky's linguistic creativity focused on the speaker's production rather than the hearer's understanding. Hearers can only register passively what speakers have created. Furthermore two mechanisms of interpretation are at work in the hearer's understanding: for well-formed sentences, the interpretation is automatic and straightforward. For ill-formed sentences, a mechanism of imposing interpretation is often at play.

merely reproduced by memory ' (Paul, 1886, 97-8). A theory of language that neglects this ' creative ' aspect of language is of only marginal interest. (Chomsky, [1962] 1964 : 914-915).

Further in the same text, he specifies that it is ' 'rule-governed creativity' by means of an explicit generative grammar and not 'rule-changing creativity' which is involved in the ordinary everyday use of language '. (Chomsky, [1962] 1964 : 921). For Chomsky ' the 'creative' aspect of language ' is associated with ' the system of generative rules that assign structural descriptions to arbitrary utterances and thus embody the speaker's competence in and knowledge of his language. ' (Chomsky, [1962] 1964 : 922). This argument is repeated in *Aspects* where the role of remembrance is denied in the use of language :

(13)

the fundamental fact about the normal use of language, namely the speaker's ability to produce and understand instantly new sentences that are not similar to those heard in any physically defined sense, or in terms of any notion of frames or classes of elements, nor associated with those previously heard by conditioning, nor obtainable from them by any sort of 'generalization' known to psychology or philosophy. Chomsky (1965 : 57)

1.3. Chomsky, statistics and probabilities

Like many linguists, logicians and philosophers of sciences at this time, Chomsky paid much attention to Shannon and Weaver's book *Theory of Mathematical Communication* published in 1948, as well as to Zipf's and Mandelbrot's works on statistical models of vocabulary. As early as 1956, having probably read Zipf (1902-1950), he rejected any statistical definition of grammaticality[4]:

(14)

There is no significant correlation between order of approximation and grammaticalness. If we order the strings of a given length in terms of order of approximation to English, we shall find both grammatical and ungrammatical strings scattered throughout the list… (Chomsky 1956 : 116).

The same argument is used in *Syntactic Structures* :

(15)

If we rank the sequences of a given length in order of statistical approximation to English, we will find both grammatical and ungrammatical sequences scattered throughout the list ; there appears to be no particular relation between order of approximation and grammaticalness. (Chomsky, 1957 : 17).

Note that, in the same chapter of *Syntactic Structures*, he criticizes Hockett's proposition to replace possible sentences by high probable sentences and impossible

[4] According to Zipf's law, empirical data on word frequencies may by represented by an harmonic law: when the words of a text are ranked in order of decreasing frequency, the frequency of a word is inversely proportional to its rank. Benoît Mandelbrot (b. 1924) developed a statistical model which provided a theoretical explanation for Zipf's law.

sentences by low probable sentences. In other words, Chomsky argues against any relationship between probabilities and grammaticality.
Moreover, Chomsky doubts that some sentences, although simple, may be found in any natural corpus. Here again, probability and grammaticality should be distinguished :

(16)
CHOMSKY : … I think 'John ate a sandwich' is a highly unusual sentence that I would be unlikely to say in a lifetime. Just as I would be unlikely to say 'grass is green' or 'birds fly'. These sentences have zero probability. Maybe in talking about probability of sentences you mean grammaticality.
STOCKWELL : You might say 'John is eating a sandwich' but not 'John eats a sandwich'.
CHOMSKY ; Probability has to do with the number of times you find a given item. If we take a sentence like 'John ate a sandwich' I would bet that you would not find it in all the sentences recorded in the Library of Congress.
(Chomsky [1958], 1962 : 180)[5]

Concerning the studies on the statistical properties of language and the use of probabilities, significant variations can be observed in Chomsky's position. When working with George Miller (b. 1920), he seemed more favourable to this kind of methods. In their common paper published in 1963, he agreed that Zipf's law as well as Mandelbrot's work, dealing with probabilities and word length, have to be taken seriously, and their results discussed and verified:

(17)
Miller and Newman (1958) have verified the prediction that the average frequency of words of length *i* is a reciprocal function of their average rank with respect to increasing length. (Miller and Chomsky 1963 : 461).

Conversely, in his review in *Language* of Vitold Belevitch's book untitled *Langage des machines et langage humain*, Chomsky seemed less enthusiastic about probabilistic models, and his appraisal of Mandelbrot's work is ambiguous : while doubting the real significance of Zipf's law, he acknowledged the importance of Mandelbrot's work:

(18)
The real import of Mandelbrot's work for linguistics seems to be that it shows that rank-frequency distributions of the type that Zipf and others have found are consistent with a very wide class of plausible assumptions about linguistic structure, and consequently, that we learn practically nothing about words when we discover this rank-frequency relation. In other words, this way of looking at linguistic data is apparently not a very fruitful one. (Chomsky, 1958 : 102).

[5] This argument has been completed by the contrast between grammatical sentences and meaningful sentences exemplified by the famous ' Colourless green ideas sleep furiously '. The significant point here is that Chomsky refers to very simple sentences.

He concludes in a similar way, doubting of the explanatory significance of statistical studies for linguistics at the same time as he claims their interest:

(19)
Although statistical properties of language and explanatory models for observed uniformities are certainly worth studying, it seems that such investigations have not yet reached the point where they make a significant contribution to the understanding of linguistic processes. (Chomsky, 1958 : 105).

Thus, though he reasserts that statistical properties of language are worth studying, he remains cautious and rather uncommitted. Actually, as far as they do not concern syntax, he does not deny the interest of statistical studies and is inclined to downplay his former criticisms:

(20)
Given the grammar of a language, one can study the use of the language statistically in various ways ; and the development of probabilistic models for the use of language (as distinct from the syntactic structure of language) can be quite rewarding. (Chomsky, 1957 : 17, note 4)

In particular, grammar should remain independent of meaning and probabilities. Thus Chomsky asserts the autonomy of syntax :

(21)
Despite the undeniable interest and importance of semantic and statistical studies of language, they appear to have no direct relevance to the problem of determining or characterizing the set of grammatical utterances. I think that we are forced to conclude that grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure. (Chomsky, 1957 : 17)

1.4. Markov's model

According to Markov's model, revisited by Shannon and Weaver's theory of information, the sentence is conceived as a left-to-right finite state Markov process or verbal chain in which the probability of a word's occurrence is determined by the occurrence of the words preceding it. Chomsky used several arguments against this model.

First, unlike phrase grammar and transformational grammar, finite-state grammar is unable to deal with recursivity:

(22)
If a grammar has no recursive steps … it will be prohibitively complex - it will, in fact, turn out to be little better than a list of strings or of morpheme class sequences in the case of natural languages. If it does have recursive devices, it will produce infinitely many sentences. (Chomsky, 1956, pp.115-116)

Secondly, Chomsky rejected Markov's model as unable to generate the set of grammatical sentences. It will generate non-sentences as well :

(23)
In short, the approach to the analysis of grammaticalness suggested here in terms of a finite state Markov process that produces sentences from left to right, appears to lead to a dead end just as surely as the proposals rejected in §2. If a grammar of this type produces all English sentences, it will produce many non-sentences as well. If it produces only English sentences, we can be sure that there will be an infinite number of true sentences, false sentences, reasonable questions, etc., which it simply will not produce. (Chomsky, 1957, p.24)

However, when discussing Markov model more thoroughly in their paper, Chomsky and Miller (1963) agreed that, though it cannot be implemented on syntax to provide the set of grammatical sentences, it can be applied for lower-level production, such as phonemes, letters and syllables :

(24)
Higher-order approximations to the statistical structure of English have been used to manipulate the apparent meaningfulness of letter and word sequences as a variable in psychological experiments. As k increases, the sequences of symbols take on a more familiar look and - although they remain nonsensical - the fact seems to be empirically established that they become easier to perceive and to remember correctly. …We know that the sequences produced by k-limited Markov sources cannot converge on the set of grammatical utterances as k increases because there are many grammatical sentences that are never uttered and so could not be represented in any estimation of transitional probabilities. (Miller and Chomsky 1963: 429)

In fact Chomsky did not deny statistical studies but excluded them from his realm of interest. His main opposition focalized on finite-state grammars and Markov's model which involve syntactic issues[6].

2. Chomsky's arguments revisited by corpus linguists

2.1. The reconstruction stance

'Corpus' is an ambiguous term : it refers both to a set of data and to a set of methods. In the former sense, it can be said that any linguist is a potential user of corpora, since linguistics, indisputably, remains an empirically based scientific area ; in the latter sense, corpus investigations involve inductive instead of hypothetico-deductive methods, meaning that data-driven analyses are preferred to rule-driven ones. Furthermore they often include statistical or probabilistic methods, but not systematically since corpus research can merely be based on simple concordances.
However, many corpus researchers consider that corpora are not only data or methods but that they have given rise to new theoretical issues. Since the beginning of the 1990s, Corpus Linguistics has been claimed as a new field, even as a new paradigm in language sciences. See for example Leech (1992) :

[6] This is also Abney's position : ' … the inadequacy of Markov models is not that they are statistical, but that they are statistical versions of finite—state automata ! Each of Chomsky's arguments turns on the fact that Markov models are finite—state, not on the fact that they are stochastic. None of his criticisms are applicable to stochastic models generally. ' (Abney, 1996 : 20)

(25)
I wish to argue that computer corpus linguistics defines not just a newly emerging methodology for studying language, but a new research enterprise, and in fact a new philosophical approach to the subject. (Leech, 1992: 106f.)

Geoffrey Leech (b. 1936) proposes the following story: in the 1940-50s, corpora were flourishing among American structuralists. However, because of Chomsky, Corpus Linguistics went to sleep for more than twenty years. Only in the 1980-90s, did corpus research come back with the increasing power of computers and the availability of very large corpora[7].

(26)
The impact of Chomskyan linguistics was to place the methods associated with CCL [Computer Corpus Linguistics] in a backwater, where they were neglected for a quarter of a century (Leech, 1992 :110).

(27)
The discontinuity can be located fairly precisely in the late 1950s. Chomsky had effectively put to flight the corpus linguistics of the earlier generation. (Leech, 1991 :8).

This story appeared in the context of a general claim of the resurgence of empiricism against rationalism in Natural Language Processing in the 1990s (Church and Mercer, 1993). Probabilistic methods first applied to speech recognition, spread to other linguistic fields when *knowledge-based and rule-based methods* have been claimed to give no more results.

A slightly different version of the story was proposed by Leech in 1991, emphasizing the apparition of a second intermediary generation of corpora at the beginning of the 1960s : Randolph Quirk's Survey of English Usage (SEU) and Kucera and Francis' Brown Corpus, regarded as 'the founders of a new school of Corpus Linguistics, little noticed by the mainstream' (Leech 1991 :8).

The implications of this story are substantial. Chomsky is considered the only one responsible for the disparition of corpus studies for about thirty years. The claim of a revival of Corpus Linguistics in the 1990s implies that there has been a continuity between neo-Bloomfieldian methods and present work, and between neo-Bloomfieldian methods and the second generation of corpora. Thus the claim of continuity implies that Chomsky's critiques aimed just as well at the Brown corpus, generally considered a pioneer, as they aimed at neo-Bloomfieldian methods[8].

Actually, when Leech (1991 :8) recalls Chomsky's view on the inadequacy of natural corpora and Markov's model to found grammaticalness, he does not specify that Chomsky's arguments could not concern the Brown corpus. The first results were

[7] This story is accredited by several corpus linguists. See for example T. McEnery & A. Wilson. Corpus Linguistics. Edinburgh : Edinburgh University Press, 1996. In his review, Stubbs (1997) strongly disagrees with such a use of Chomsky's arguments.

[8] On this point see Léon (2005).

published in 1967 (Kucera and Francis, 1967) while Chomsky's critiques date from the 1950s and early 1960s. Moreover it consisted mainly of word frequency counts carried out on sampled texts ; its aim was to compare frequency counts between genres, and to test general statistical models on vocabulary. No idea there of taxonomy, distributional classes or discovery procedures.

Therefore, the notion of corpus at work in the Brown corpus did not match the American structuralist approach. Chomsky's arguments against corpora were not directed at this type of research since his position was that statistical studies could be valuable provided that they did not deal with syntax. Thus there is no continuity of methods between the Neo-Bloomfieldians and the second generation of corpus allegated by corpus linguists. Appealing to Chomsky's arguments seems here quite misleading.

It should be added that one of the Brown corpus's author partially agreed with Chomsky on the use of statistical studies. When using a Markov model in the comparative phonological study of Russian, Czech and German (Kucera and Monroe 1968), Kucera (b. 1925) agreed with Chomsky that this type of model could only be applied to lower-level units and not to syntax and sentences.

Besides, Chomsky was not the only one criticizing discovery procedures and empiricist methods in the 1950s-60s. The debate was a vivid one at that time. In particular, Yehoshua Bar-Hillel (1915-1975), who promoted his own operational syntax based on categorial grammar, was strongly opposed to these methods, especially in the field of Machine Translation. Bar-Hillel was doubting of the empirical method, in particular data-driven grammars which consist of 'deriving syntactical rules from a huge number of observations, that is examples that occur in some actual text, rather than testing rules with concocted counter-examples' (Bar-Hillel, 1960 : 110).

When examining the arguments used to invalid Chomsky's critiques in order to promote Corpus Linguistics as a 'new philosophical approach', it can be seen that they are more technical than theoretical. Corpus linguists claim a theoretical program, systematically opposed to the Chomskyan model : performance against competence, linguistic description against universals, use of quantitative methods in addition to qualitative methods, empiricist approach against rationalist approach. Yet, for Leech, the revival of Corpus linguistics is strongly associated with the increasing power of computers : 'The new master is the computer' he says in his 1992 article (p.105).

The term 'Computer Corpus Linguistics' coined by Leech in 1992, is above all a technical field focused on practical issues. It refers to Corpus Linguistics as belonging to Natural Language Processing. And the arguments he opposed Chomsky's criticisms are of practical nature : that is large computer-based corpora enabling investigations on vast amounts of lexical and syntactic phenomena.

Likewise, Leech opposes a practical argument to Chomsky's criticism of Markov's model and their inadequacy to account for syntactic structures. For Leech, finite-state grammars (with probabilities assigned to state transitions) prove to be the most successful system in automatic grammatical tagging. Markov algorithms associated with large corpora are better tools than ruled-driven algorithms to deal with Natural Language Processing, as far as they tolerate a certain amount of errors - which is implied by the term ' robust '. This argument refers more to Natural Language Engineering than to linguistic theory :

(28)
One thing in favour of probabilistic language processing systems is that they are eminently *robust*. They are fallible, but they work ; they produce a more or less accurate result, even on unrestricted input data, in a way that outperforms most rule-driven language modelling systems. (Leech, 1991 : 18)

2.2. The British empiricist tradition inherited from Firth

Let us know examine the other stance on Chomsky's arguments against corpora, stemming from the neo-Firthians and represented by M.A.K. Halliday (b. 1925), John Sinclair (b. 1933) and their followers.

It should be said that these stances coincide with the two British traditions noted by Stubbs (1993). According to him, the Firth-Halliday-Sinclair line of development differs from the Quirk-Leech Corpus Linguistics on two main points. Quirk's grammars (Quirk and al. 1972, 1985), despite the avaibility of the Survey of English Usage, are essentially based on invented data studies, while Sinclair's works on lexicography are always based on attested data. The second difference concerns corpora construction. While the Brown Corpus and his immediate followers, the LOB and the LUND corpora, are using sampling methods, Sinclair's Cobuild only includes whole texts which alone can be dealt with by inductive methods safely[9].

It should be added that Quirk (1968) rejected Firth's filiation very early when he regarded Henry Sweet as the true pioneer of lexicographical grammar instead of Firth.

Researchers belonging to the Firth-Halliday-Sinclair trend sometimes think that advances in Corpus Linguistics are not so great as is claimed by the other stance, and that the most important issue of corpora is the avaibility of data. From this view, ‘ Corpus linguistics ’ is far from being a new linguistics[10]. See Kennedy (1998) :

(29)
Although there have been spectacular advances in the development and use of electronic corpora, the essential nature of text-based linguistic studies has not necessarily changed as much as is sometimes suggested. Corpus Linguistics did not begin with the development of computers but there is no doubt that computers have given Corpus Linguistics a huge boost by reducing much of the drudgery of text-based linguistic description and vastly increasing the size of the databases used for analysis. (Kennedy, 1998 : 2)

In the second stance, instead of historical reconstruction, there is a strong tradition of empirical linguistics. Several Chomskyan issues have been discussed by the neo-Firthians since the 1950s and alternative solutions proposed : grammaticalness vs.

[9] However it should be said that the two groups are not explicitely opposed. In fact they sometimes meet in the same publications (see Svartvik 1992 for example).

[10] This view is also shared by corpus researchers coming from Computational Linguistics: ‘Potentially these corpora enable a range and scope of research opportunities unmatched by earlier corpus projects. In practice, though, researchers have not always fully exploited this potential. First of all, relatively few studies have exploited the machine-readable character of these corpora.’ (Biber et Finegan, 1991 : 209)]

acceptability, wellformedness vs. naturalness, linguistic creativity, probabilistic linguistics, lexicogrammar.

• Grammaticality and acceptability

As the 1950s, British researchers of the London School challenged the distinction grammaticalness / acceptability proposed by Chomsky and the relevance of frequency of use. When working for his PhD in the 1950s, Halliday investigated the frequencies of syntactic classes in a Chinese dialect. His procedure was intended to distinguish whether non-occurrence or low occurrence in a corpus was due to chance or whether it was evidence for the ungrammaticality or rarity of a structure in the language (Halliday 1959 : 58).

Others in the 1960-70s designed experiments to investigate the relationship between frequency and acceptability judgements (see for example Quirk and Svartvik, 1966, Greenbaum 1976). Some of them continued to develop Firth's legacy in fields like systemic analysis or lexicography and finally corpus-based research. They addressed theoretical issues which seems more challenging for Chomskyan views.

• Probabilistic hypothesis and lexis-grammar continuum

As far as corpora are concerned, Halliday's position rests on two main assumptions:

1) the linguistic system is inherently probabilistic : 'frequency in text is the instantiation of probability in the grammar' (Halliday, 1991 : 30 ; Halliday, 1992 : 66).

2) there is no fundamental difference between lexis and grammar : ' I have always seen lexicogrammar as a unified phenomenon, a single level of 'wording' of which lexis is the 'most delicate' resolution ' (Halliday, 1991 : 31).

These options opposed Halliday to Chomsky very early. During the discussion which followed Chomsky's talk at the 9th Congress of linguists in 1962, Halliday acknowledged the interest of *grammaticalness* on condition that it is expressed in terms of degree and not of exclusivity between well-formed and ill-formed sentences, and that it is completed by *lexicalness* (see Chomsky, 1964 : 989).

In a paradigmatic interpretation, lexis and grammar form a continuum : at one end is the grammar, described as general choices, such as 'polarity : positive / negative' 'mood : indicative (declarative /interrogative) / imperative', 'transitivity : material / mental / relational', while at the other end is the lexis, with highly specific but open-ended choices. Lexis is open-ended, while grammar contains closed classes.

Both assumptions, statistical properties and complementarity between lexis and grammar are connected. Given the notion of lexicogrammar, it does not make sense to accept relative frequency in lexis on one hand and deny its validity in grammar on the other hand. Thus Halliday advocated that Zipf's law should be generalized to syntax. Besides, probabilities can be interpreted in terms of Shannon and Weaver's Information Theory so that frequency information from the corpus can be used to set the probability profile of any grammatical system[11].

[11] Hypothesis on probabilistic properties of language has been taken up by several linguists. See Bod and al. (2003).

Sinclair took up the issue of loose boundaries between lexis and grammar taken by assuming that speakers use ready-made linguistic forms, or prepackaged chunks, rather than isolated words in rule-governed sequences. Sinclair (1991 : 109) speaks of two complementary 'principles', namely the 'open-choice principle' when speakers choose words in rule-governed sequences, compatible with Chomsky's views and generally adopted by linguists, and the 'idiom principle' when they choose semi-preconstructed sequences, such as idioms, phrasal verbs or collocations.

• Collocations and linguistic creativity

Firth's view on linguistic creativity was expressed in strongly empiricist words :

(30)

> Language, like personality, is a *binder of time*, of the past and future in 'the present'. On the one hand there is habit, custom, tradition, and on the other innovation, creation. Every time you speak you create anew, and what you create is a function of your language and of your personality. From that activity you may make abstraction of the constituents of the context, and consider them in their mutual relations. In the process of speaking there is pattern and structure actively maintained by the body which is itself an organized structure maintaining the pattern of life. (Firth, 1957 [1948] : 142).

It was criticized by Chomskyans as early as the late 1960s, when Langendoen, in his dissertation supervised by Chomsky, said that Firth's approach was based on the 'opinion that language is not 'creative' and that a person is totally constrained essentially to say what he does by the given social situation.'(Langendoen, 1968 :3).

For Neo-Firthians, language in use remains a balance between routine and creation, and transmits the culture (Stubbs, 1993). For them, Sinclair's idiom principle is an issue that questions Chomsky's linguistic creativity. The use of high frequencies of preconstructed segments, such as collocations, give new relevance to memory in language learning and production. They reintroduce probabilities as a language property. Counter to Chomsky's view, Kennedy (1998) claims that the use of partially lexicalized elements does not restrict the innovative property of language[12]. There is no reason why many sentences cannot be treated as partially lexicalized rather than purely syntactically generated.

A similar argument has been put forward by historians of linguistics, such as Joseph (2003), to show that Chomsky's conception of infinite linguistic creativity obliges him to reject any ' collocational ' model while for Sinclair and his followers, collocations do not involve a lack of creativity.

• well-formedness and naturalness

Excerpt (16) above exemplified Chomsky's argument against corpora relying on the fact that well-formed and (above all) simple sentences may never occur in any natural

[12] This argument has also been taken up by computational linguists: 'They [collocations] also have theoretical interest : to the extent that most of language use is people reusing phrases and constructions that they have heard, this serves to de-emphasize the Chomskyan focus on the creativity of language use, and to give more strength to something like a Hallidayan approach that considers language to be inseparable from its pragmatic and social context'. (Manning et Schütze, 2002 :29f) .

corpus. Similar examples seem to have been used *a contrario* by Sinclair (1984) when he contrasts well-formedness and naturalness. A simple sentence, such as 'Prince Charles is now a husband' can be syntactically well-formed and yet native speakers may still feel that it is unnatural. 'Well-formedness and naturalness are independent variables' (Sinclair, 1984 : 95). Sinclair suggests that naturalness will always be probabilistic and therefore distinct from well-formedness, which is absolute ; the textual evidence for naturalness is probabilistic.

• corpora versus intuition

Sinclair (1991) criticizes the sole recourse to intuition unable to deal with language use. First, properties such as grammaticality do not exist for lexis. Besides, in large texts, the meaning of the most frequent words is not the meaning given by intuition. Language use seems to delexicalize the most frequent words by reducing their distinctive contribution to meaning.

However, some arguments seem rather dubious. According to Stubbs (1995), native speakers may be able to give examples of collocation or to judge their likelihood, but they cannot document them, that is give accurate estimates of their frequency. They are very poor at estimating large numbers. This argument cannot be said to infirm the recourse to intuition. As categories in a grammar, frequency counts should be regarded as parts of the analysis, so that they are not directly accessible to native speakers' intuition.

Some authors close to Sinclair, such as Kennedy, assume a mixed position associating intuition and corpus work. In some respects, they agree with Chomsky. Kennedy acknowledges that corpora are not able to account for some aspects of language, such as the distinction between possible and impossible. Unlike many corpus linguists, Kennedy does not seem to advocate the sole recourse to attested data and acknowledges that an element not occurring in a corpus does not mean that this element does not exist. Conversely the occurrence of an element in a corpus does not establish its grammaticality :

(31)

The use of both introspection and corpus-based analysis can contribute to linguistic analysis and description. Corpora cannot tell us everything about how a language works. For example, they cannot be used as a basis for stating what structures or processes are not possible … The fact that an item or structure does not appear in even the largest corpus does not necessarily mean that it cannot occur, but could suggest the corpus might be inadequate or the item infrequent. Neither does the fact that a construction occurs in a corpus necessarily establish its grammaticality. … Whether utterances which involve phonetic or syntactic reductions such as *where you going ?, wannanother one ?* or *Good that you got here early* have to be accounted for grammatically will probably depend in the final analysis on frequency of occurrence and intuitive judgments as to what is 'normal'. (Kennedy, 1998 : 271f.)

Conclusion

The attempts of Corpus Linguistics to become an autonomous field within sciences of language encounter substantial difficulties. Unifying under a unique name all the domains where corpora are used in linguistics requires an epistemological stance which is hard to take for granted : institute a practical object and a set of methods in the place of a theoretical object. This stance induces its supporters to indulge in doubtful means of legitimization, such as the creation of a more or less credible history. In particular the fact that Chomsky would have stopped corpora during the 1960s makes statistical studies of vocabulary appear to be the heirs of Neo-Bloomfieldians and the main target of Chomsky's criticism, which is false. In this respect, it was the London School that was the object of Chomsky's attacks at the same time as Neo-Bloomfieldians, not the Brown Corpus.[13]

The proposals put forward to define a new linguistic paradigm do not belong to the sole Corpus Linguistics in so far as they look very similar to those advanced by functionalists to dissociate themselves from the generativist model and from structuralism in general[14]. Functionalist are more favourable to a continuist stance than to a radical opposition. They claim a continuity between rationalism and empiricism, so that data provided by use or by statistical methods, and those provided by intuition are complementary. They claim a weak version of innateness and a continuum between universalism and relativism. They question the relevance of a strict distinction between competence and performance, between the speaker's grammatical knowledge and his knowledge of the use of grammar. The use of corpora and statistical methods are only aspects among others of the functionalist approach which remains widely favourable to large scale empirical data.

Corpora are used by all the trends of language sciences, whatever their theoretical options. They do not define a new paradigm. Large computerized corpora make new data available to every linguist. Therefore we cannot but agree with Sinclair (1991) and Halliday (1992) when they claim that new technological means, such as instrumentation in phonetics and later computerized corpora, provided linguistics at last with really significant data.

This is not the first time that such an attempt to legitimate computer applications has been undertaken. Academical, financial and sometimes industrial issues are at stake. See for instance the history of Machine Translation and Natural Language Processing more generally.

Jacqueline Léon
Laboratoire d'histoire des théories linguistiques
CNRS, Université Paris 7
Case 7034
2, place Jussieu
75005 Paris
France
Jacqueline.leon@linguist.jussieu.fr

References:

[13] 'Modern linguistics has been largely concerned with observational adequacy. In particular, this is true of post-Bloomfieldian American linguistics … and apparently, of the London school of Firth, with its emphasis on the ad hoc character of linguistic description.' [note 9 : Firth J.R. and al. 1957, *Studies in Linguistic Analysis*, Oxford, England.] (Chomsky, 1964 :924).

[14] See Noonan's article (1999) who locates the functionnalist model - which he names 'West Coast Functionalism' - with respect to the structuralist model and the formalist model, by examining the various features which they have in common and those which differ. Functionalists see in the generativist paradigm one form of the structuralist model.

Abney, Steven
1996. 'Statistical Methods and Linguistics' In J. Klavans and Ph. Resnik (eds) *The Balancing Act*. Cambridge : The MIT Press.

Aijmer Karin & Bengt Altenberg, 1991, *English Corpus Linguistics : Studies in Honour of Jan Svartvik*, London & New York : Longman.

Bar-Hillel, Yehoshua
1960. 'The present Status of Automatic Translation of Languages' *Advances in Computers* vol.1, F.C. Alt ed. Academic Press, N.Y., London: 91-141.

Biber, Douglas and Finegan, Edward
1991. 'On the exploitation of computerized corpora in variation studies' In Aijmer & Altenberg (eds*) English Corpus Linguistics : Studies in Honour of Jan Svartvik*, London & New York : Longman: 204-220.

Biber, Douglas, Conrad, Susan and Randi Reppen
1998. *Corpus Linguistics : investigating language structure and use*. Cambridge : CUP

Bod, Rens, Hay Jennifer, and Jannedy, Stefanie (eds)
2003. *Probabilistic Linguistics*, Cambridge (Mass.), London (England) : The MIT Press.

Bourdeau, Michel
1979. *Chomsky et la critique des théories behavioristes du langage*, Doctorat de 3e cycle, Université Paris 1.

Chomsky, Noam
1956. Three models for the description of language. In : *IRE Transactions on Information Theory* IT-2, 113-124.

Chomsky, Noam
1957. *Syntactic Structures*. The Hague : Mouton.

Chomsky Noam
1958. review of Vitold Belevitch *Langage des machines et langage humain* 1956, *Language* 34-1 :99-105.

Chomsky, Noam
1959. review of B.F. Skinner *Verbal Behavior* 1957 New York : Appleton-Century-Crofts., *Language* 35-1 :26-59.

Chomsky Noam 1964a « Transformational Approach to Syntax [1958] » *The structure of language. Readings in the philosophy of language* Ed. by Jerry A. Fodor, Jerrold J. Katz. New jersey: Prentice Hall, 211-245.

#Chomsky, Noam
1964b. 'The Logical Basis of Linguistic Theory' *Proceedings of the 9th International Congress of Linguists*, 1962, ed. by Horace G. Lunt, 914-1008. The Hague: Mouton.

Chomsky Noam
1965. *Aspects of the Theory of Syntax*, Cambridge :the MIT.

# Chomsky Noam and Miller George

1963. « Finitary Models of Language Users », in : *Handbook of Mathematical Psychology*, ed. by Robert D. Luce, Robert R.. Bush, Eugene Galanter, Vol. II, New York, Wiley :419-491.

Church, Kenneth, Robert L. Mercer
1993. 'Introduction to the special Issue on Computational Linguistics Using Large Corpora'. *Computational Linguistics* 19 : 1-24.

Firth, John Rupert
1957 [1948]. 'The semantics of linguistic science', in : *Papers in Linguistics (1934-1951),* Oxford, Oxford University Press : 139-147.

Goodman Nelson
1955. *Fact, Fiction and Forecast*, Indianapolis and New York: The Bobbs-Merrill Company Inc.

# Greenbaum Sidney, 1976, Syntactic Frequency and Acceptability, *Lingua* 40 99-113.

Halliday, M.A.K.
1959. *The language of the Chinese 'Secret History of the Mongols'*. Oxford : The Philological Society.

Halliday M.A.K
1991. 'Corpus Studies and Probabilistic Grammar' In Aijmer & Altenberg (eds) *English Corpus Linguistics : Studies in Honour of Jan Svartvik*, London: Longman: 30-43.

Halliday M.A.K
1992. 'Language as System and Language as Instance : the Corpus as a theoretical Construct' *Directions in Corpus Linguistics* ed. by J. Svartvik, 61-77, Berlin : Mouton de Gruyter.

Hockett, Charles F.
1948 [1957]. 'A note on structure' In : Joos, Martin (ed.) *Readings in Linguistics I. The Development of Descriptive Linguistics in America 1925-56*. 3rd Edition. Chicago and London : The University of Chicago Press : 279-280.

Hockett, Charles F.
1954, 'Two models of grammatical Description', *Word* 10: 210-234.

Joseph, John E. 2003 'Rethinking linguistic creativity' In *Rethinking Linguistics*, Haylay Davis Talbot Taylor (eds) London and New York : Routledge Curzon

Kennedy, Graeme
1998. *An Introduction to Corpus Linguistics*. London and New York : Longman.

Kucera, Henry and W. Nelson Francis
1967. *Computational Analysis of Present Day American English*. Providence: Brown University Press.

Kucera, Henry and George K. Monroe
1968. 'A comparative quantitative phonology of Russian, Czech and German' New York : American Elsevier Publ.

Langendoen, D. Terence 1968.

The London School of Linguistics: a study of the Linguistic Theories of B.Malinowski and J.R. Firth. Research Monograph n°46, Cambridge Massachussetts : the MIT Press.

Leech, Geoffrey
1991. 'The state of the Art in Corpus Linguistics', in *English Corpus Linguistics : Studies in Honour of Jan Svartvik*, Aijmer & Altenberg (eds.) London & New York : Longman: 8-29.

Leech, Geoffrey
1992. 'Corpora and theories of linguistic performance', in Svartvik Jan (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*, Berlin, New York Mouton de Gruyter: 105 -122

Léon Jacqueline
2005. 'Claimed and unclaimed sources of Corpus Linguistics' *The Henry Sweet Society Bulletin of History of Linguistic Ideas* n°44 :34-48.

Miller George A., Newman E.B. 1958 Tests of a statistical explanation of the rank-frequency relation for words in written English. *American Journal of Psychology*, 1958, 71 : 209-258.

Noonan, Michael,
1999. 'Non-structuralist Syntax', in Darnell Michael et al. (eds.), *Functionalism and Formalism in Linguistics*, Amsterdam, Philadelphia : John Benjamins Publishing Company :11-32.

# Paul, Hermann, 1886, Prinzipien der Sprachgeschichte Second edition (1886) Translated into English by H.A. Strong London : Longmans, Green and co. 1890.

Quirk, Randolph and Jan Svartvik
1966. *Investigating linguistic acceptability*. The Hague : Mouton.

Quirk, Randolph
1968 *Essays on the English Language: Medieval and modern*. London: Longman

# Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik. 1972. *A grammar of Contemporary English*. London : Longman.

# Quirk, Randolph, Sidney Greenbaum Geoffrey Leech, Jan Svartvik,1985 *A comprehensive grammar of the English language*. London : Longman.

# Shannon Claude E. / Weaver Warren
1949
The Mathematical Theory of Communication Urbana: University of Illinois Press

Sinclair, John
1984. 'Naturalness in Language', In J. Aarts and W. Meijs (eds) *Corpus Linguistics : recent developments in the use of Computer corpora in English Language Research*, Amsterdam : Rodopi, 203-210.

Sinclair, John
1991. *Corpus, concordance, collocation*, Oxford : Oxford University Press.

Stubbs, Michael
1993, 'British Traditions in Text Analysis – From Firth to Sinclair', In M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology. In Honour of John Sinclair,* Amsterdam : John Benjamins, 1-36.

Stubbs, Michael

1995, 'Collocations and semantic profiles : On the cause of the trouble with the Quantitative studies' *Functions of Language* 2 (1) : 1-33

Stubbs, Michael, 1997. Review of Tony McEnery & Andrew Wilson. Corpus Linguistics. Edinburgh : Edinburgh University Press, 1996, *International Journal of Corpus Linguistics* Vol 2-2 : 296-302.