Claimed and Unclaimed Sources of *Corpus Linguistics*

Jacqueline Léon
CNRS, Université Paris 7, France

The promotion of the term *Corpus Linguistics* in the 1990s has marked an important milestone in the attempt of making corpus works a new mainstream discipline within Language Sciences. Since an international conference held in 1991 gathering British, Dutch, Swedish and Norwegian linguists (proceedings in Svartvik (ed) 1992), the researchers of the domain have strengthened their position by the publication of many collective books and text books and the creation in 1996 of an international journal *The International Journal of Corpus Linguistics*. So in 2002, Geoffrey Leech (b. 1936), a leading figure in corpus research, could speak of 'the corpus linguistic world as a well-established research community' (Leech, 2002:167). At the same time as they were giving themselves a name, dating the first occurrence of the term back to 1984 (Aarts and Meijs, 1984), these linguists attempted to provide the new domain with historical legitimacy. It is this issue that I would like to address in my paper in order to see under what conditions, but also at what cost, a common history has been claimed in order to found a new linguistic stream. In particular, it will be shown how the actors have retrospectively built their own history by overstating or forgetting some events, facts or methods.

First it is to be noted that what is called 'Corpus Linguistics' covers various heterogeneous fields ranging from lexicography, descriptive linguistics, applied linguistics – language teaching or Natural Language Processing – to 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 – often including statistical or probability methods but not systematically. Corpus investigations involve inductive method instead of hypothetico-deductive method, meaning that data-driven analyses are preferred to rule-driven ones. The point here is why one should gather such a diversity of approaches under a single term. Besides, why define a specific domain when, since linguistics indisputably remains an empirically based scientific area, any linguist is a potential user of corpora.[1]

In his paper entitled "Corpora and theories of linguistic performance," Leech (1992) promotes computer-based corpus research as a new paradigm of linguistics, denying that it may be regarded as a mere technique or method: '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: 106-107) and lists its main features:

- Focus on linguistic performance, rather than competence;

[1] See for example Fillmore's paper (1992) in one of the corpus linguistics text books, which is an attempt to reconcile intuitive methods (armchair linguistics) with corpus linguistics. See also the French Journal *Corpus* which presents studies using both corpus and hypothetico-deductive methods.

- Focus on linguistic description rather than linguistic universals;
- Focus on quantitative, as well as qualitative models of language;
- Focus on a more empiricist, rather than a rationalist view of scientific inquiry.

Leech insists that each of these features highlights a contrast between the Corpus Linguistics paradigm and the Chomskyan paradigm (and dedicates a long development to it in his paper). In other words, this set of propositions is presented as an anti-Chomskian paradigm.

Let us now examine the common history which is now so largely spread among the researchers and text books that it can be said that a real historiography has been retrospectively set up in the 1990s.

1. The common story

The first versions of the story appeared in Leech's contribution to the studies published in honour of Jan Svartvik (1991), and in two sets of works regarded as significative of the resurgence of the domain claimed at the beginning of the 1990s: Svartik's collective book of 1992 and the special issue of the journal *Computational Linguistic*s about using large corpora published in 1993, introduced by a historical overview (Church and Mercer, 1993). Note that, in the area of Natural Language Processing, the domain, called either *Computational Linguistics Using Large Corpora* or *Corpus-based Natural Language Processing*, had not been dubbed by any specific term at that time.[2]

Leech (1992) gives a simplified version of the story. In the 1940s-50s, corpora were flourishing among American structuralists, for whom 'a corpus of authentically occurring discourse was the thing that the linguist was meant to be studying' (1992: 105). Afterwards corpus linguistics went to sleep for twenty years and only came back in the 1980s with the increasing power of computers and the availability of very large corpora. Chomsky's criticisms of the 1950–60s are put forward to account for the decline of corpus linguistics: '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).[3]

This version of the story takes a more general move with the resurgence of empiricism against rationalism in the 1990s claimed both by Leech (1992) and Church and Mercer (1993): Computer Corpus Linguistics, as well as Corpus-based Natural Language Processing, are claimed to be a rediscovery of empirical and statistical methods popular in the 1950s, in particular the application of Shannon's information theory. After the 1950s empiricism declined, while rationalism became dominant in the areas of linguistics and artificial intelligence, marked by Chomsky's criticism of n-grams in *Syntactic Structures* (1957) and by Minsky and Papert's criticism of

[2] At present, it seems that in Natural Language Processing too the term Corpus Linguistics has been adopted. However other terms have appeared such as "Statistical Natural Language Processing" (Manning and Schütze, 1999) or "Probabilistic Linguistics" (Bod, Hay and Jannedy, 2003).

[3] See the same argument in Leech (1991): 'The discontinuity can be located fairly precisely in the late 1950s. Chomsky had effectively put to flight the corpus linguistics of the earlier generation.' (1991: 8).

Perceptron neuronal networks in 1969.[4] In addition to the increase in computer power and data availability, Church and Mercer give various reasons for the resurgence of interest in these methods in the 1990s. It was first in the area of Speech Recognition that stochastic methods, based on Shannon's model, reappeared in the 1970s when knowledge-based and rule-based methods became unsatisfactory and were given up. This change of method in speech recognition led computational linguistics to adopt probabilistic methods, notably preference-based parsing and lexical preference parsing.[5] Both versions of the story share the fact that they distinguish two corpus periods, the 1950s and the 1990s, and that, in between, corpus linguistics is said to have vanished essentially because of Chomsky.

A slightly different version had been proposed by Leech in 1991, pointing out the apparition of a second intermediary generation of corpus at the beginning of the 1960s: Randolph Quirk's Survey of English Usage (SEU) and Kucera and Francis's Brown corpus, presented as 'the founders of a new school of corpus linguistics, little noticed by the mainstream' (Leech, 1991: 8). In this version, however, it is not mentioned that the SEU predated and influenced the Brown corpus. On the contrary, the Brown corpus is considered the first real computerized corpus. What is more, this pioneer status of the Brown corpus has been taken up by many text books and collections of articles and is now widely shared among the actors of the field.[6] Let us now examine the key features of the story:

- the anteriority of the Brown corpus;
- the discontinuity of corpus design over 30 years;
- Chomsky's arguments against corpora and statistics.

2. The anteriority of the Brown corpus

The pioneer status of the Brown corpus rests on several assumptions: it had no precursor; it was the first computerized corpus and the first freely available corpus; it was supposed to favour general linguistic investigations and not just frequency counts of vocabulary.

The Brown corpus's authors claimed that it had no precursor. The front flap of Kucera and Francis's 1967 book asserts: 'The standard corpus of present-day edited American English prepared at Brown is the first and so far the only such collection of data in English that has been carefully selected by random sampling procedures from a well-defined population and that is completely synchronic, containing samples published in the US during a single calendar year (1961).' No other corpus is ever mentioned in the book, and since its publication it has been referred to as "the Brown corpus". This view was resumed thirty years later: 'The beginning of it all was the making of the Brown corpus, "a standard sample of present-day English for use with

[4] See Rosenblatt, 1958.
[5] See also the recent investigations of language probabilistic properties in Bod and al. (2003).
[6] Several text books have adopted this version since the 1990s. See for example :
Oostdijk and Haan (eds.), 1994; McEnery and Wilson, 1996; Garside, Leech, MacEnery, 1997; Simpson and Swales (eds.), 2001.

digital computers." In the next two decades it was to be followed by a string of successors…' (Svartvik, 1992: 7).

The only acknowledged sources are quantitative analyses of language and literary genre studies, especially word-frequency distribution in various languages. Thus Kucera and Francis's references essentially concern statistical works, namely Yule's and Herdan's works. In this respect, their work was not very original since word-frequency counts were flourishing at the time.[7] It should be added that Kucera was a Slavist of Czech origin, acquainted with the Prague School and its tradition of genre studies; for that matter, one of their references mentions a talk on statistics and genres given in 1966 at Brown University by Lubomir Dolezel (b.1922), one of Kucera's Czech compatriots. Not surprising, then, that the sampling of the corpus rested on genre categories and that the first studies concerned statistical genre studies.

Yet it can be shown that the Brown corpus had other sources: it resulted from a joint idea of one of Firth's pupils, Randolph Quirk, and the American Germanist Freeman Twaddell.[8] Randolph Quirk (b. 1920) decided in 1959 to devise a corpus of both spoken and written British English, the Survey of English Usage (SEU) at University College London (Quirk, 1960). The corpus was planned to be machine-readable, and though it was only computerized in 1989, Quirk took a programming course in order to achieve this purpose in the sixties.[9] The corpus was collected with the ultimate aim of supplying material for the writing of a grammar, the first version of which was published in 1972, *A Grammar of Contemporary English*, co-written by Quirk, Greenbaum, Leech and Svartik.[10]

In various respects the SEU belonged to the British tradition. Besides the fact that Quirk presented his scheme to the Philological Society and that he published his first paper on the SEU in its *Transactions*, the first investigations were for the most part on prosody and phonetics, that is following the phonetic tradition of Daniel Jones and J.R. Firth.

Freeman Twaddell (1906–1982), a specialist of German literature and phonology, created two departments at Brown University in Providence (Rhode Island), one of linguistics and one of Slavic studies in the early 1960s. In 1955, he invited Henry Kucera (b.1925) and in 1962 Nelson Francis (1910–2002) to join these departments and to participate in his corpus project.[11] The corpus was assembled at

[7] A glimpse of the extent of the area can be caught from Pierre Guiraud's *Bibliographie critique de la statistique linguistique*, published in 1954, where more than 240 references are exclusively dedicated to word counts.

[8] It should be noted that Firth's filiation is never mentioned by the Brown corpus researchers, whereas Firth is considered one of the precursors of corpus linguistics by M.A.K. Halliday, John Sinclair and their followers, and several recent corpus studies mention the notion of collocation as one of the touchstones of corpus linguistics (Léon, in preparation).

[9] The data were recordings amounting to 30,000 words (over three hours) of spontaneous English speech in the form of discussions between a total of 31 educated British adults (Crystal and Quirk, 1964). By 1991, over 200 publications had used material from the Survey Corpus, either in its original slip form or in its later computerized form (Altenberg, 1991).

[10] However, this grammar cannot be said strictly speaking to be "data-driven". As Sinclair (1991) points out, occasional reference is given to the SEU, and only a few examples had been extracted from the corpus.

[11] Kucera was a specialist of Czech phonology and became a computational linguist in order to achieve the comparison of the phonological similarities of three languages: Russian, Czech and German

Brown University during 1963-64; its analysis was performed during 1965-66 and the results published in 1967.

The connections between the SEU and the Brown corpus were strong and the anteriority of the SEU cannot be denied. During a scholarship in the United States in 1951, Quirk met Twaddell at Brown University and followed the teaching of Charles Fries at Ann Arbor. The planning of the SEU therefore owes much to Fries's empirical method of working on the syntax of spoken language (Quirk, 2002). In 1962, Nelson Francis obtained a scholarship from the Ford foundation to work on the SEU at UCL. Finally, in 1963, Randolph Quirk was present at the original conference that agreed on the specifications of the Brown corpus and where the major decisions were made (Kucera and Francis, 1967: xx).

Let us move on with the claim concerning the linguistic aims of the Brown corpus. Its ambition was to be a 'reservoir of linguistic usage in a form (computer tape) that makes it relatively easy to extract exhaustively all available specimens of a given word or describable grammatical item' (Twaddell's foreword in Kucera and Francis, 1967: v).

In the event, when we study Kucera and Francis's book, it is entirely focused on statistical studies. As they point out in their introduction: 'The main objective of this book is the presentation of lexical and statistical data about the Corpus' (Kucera and Francis, 1967: xvii); 'The bulk of the book [pp. 5-274] comprises two frequency lists of the words in the Corpus, the first in descending order of frequency and the second alphabetical. The rest [pp. 275-430] comprises a number of tables and graphs resulting from various counts, calculations, and studies and two essays analyzing some of the results' (Kucera and Francis, 1967: xxi-xxii). The book was in fact devoted to the distribution of frequent words, as well as word length and sentence length distribution, carefully carried out on sampled texts and genres. It was only later that the Brown corpus was used for grammatical investigations, once Quirk and his colleagues had already published several studies on grammar and prosody based on the exploration of the SEU (Quirk and Mulholland, 1964; Quirk and al., 1964; Quirk and Crystal, 1966).

Thus, at the end of the 1960s, the Brown corpus was used for exactly the same purpose as were quantitative data at that time, that is, statistical studies of vocabulary. Looking at Altenberg's bibliography (1991) listing the publications using English computer corpora, it can be seen that by 1970 more papers (10) using the SEU for grammatical or prosodic investigations had been published than using the Brown corpus for grammatical investigations (4 papers).[12]

The Brown corpus's main quality rested on its method of sampling. Its even sample size of 2,000 words helped to make the corpus statistically sound and was an

(Kucera, 1963; Kucera and Monroe, 1968). In particular, he had to write programs to deal with his phonological frequency data. He taught computational linguistics to his colleague Nelson Francis who had been trained in philology and dialectology (Francis, 1998; Kucera, 1998).

[12] Note that Altenberg lists the SEU among computer corpora even though it was not computerized at that time.

advantage for comparison.[13] However, as Sinclair (1991) pointed out, the divisions of the corpus into genres, settled on intuitive criteria, were less reliable. Moreover, following Gellerstam (1992) commenting on the Brown corpus, it can be said that the sampling method was more suited to producing quantitative results than conducting general linguistic investigations.[14]

The next claim concerns the anteriority of the Brown corpus as the first computerized corpus. This again has to be mitigated. There was at least one computerized predecessor; the computerization of the Trésor de la Langue Française had begun before the Brown corpus was planned. In 1957 a conference about 'Lexicologie et lexicographie françaises et romanes' took place in Strasbourg to study the faisability of a Dictionary of modern and contemporary French (1789–1960) based on a computerized Thesaurus or "Trésor de la Langue Française" (later called TLF), that is a corpus of 1350 literary or technical books written from 1789 to 1960. Subsequently, the Center of TLF was created by the CNRS (Centre National de la Recherche Scientifique) in 1960 (CNRS, 1961).[15]

Thus the anteriority of the Brown corpus cannot be settled on the computational level. However it is true, that unlike the TLF, the Brown corpus was made available immediately, and a tape provided to any researcher who asked for it.[16]

There is another predecessor which has been completely forgotten by corpus historiography, namely the Rand corpus. This corpus was created by the Rand Corporation group of Machine Translation in Los Angeles, led by David Hays (1928-1995) from 1960 to 1969. They developed empirical methods in Machine Translation as early as 1949. The method was to derive a dictionary and a grammar from the corpus and to test them on a new sample of the corpus to expand them. This method of building data-driven grammars and dictionaries was much more ambitious than the frequency explorations of the Brown corpus. Thus as early as 1959, the Rand corpus, comprising more than 200 Russian articles in Physics and Mathematics and more than two hundred thousand running words, was made available in the form of punched cards for the use of researchers.[17]

To return to the Brown corpus, we can see that it was not the first available computerized corpus, and that data-driven grammars date back to the Rand corpus.

[13] The corpus, of more than one million words, comprised 500 samples of 2000 words each. Fifteen categories (genres) were represented, from sports, scientific journals, and popular fiction to philosophical discussion.

[14] 'By looking at sampling principles, you can see that the focus was on obtaining quantitative data (frequencies of words, constructions, morphemes, graphemes) rather than on compiling a range of corpora useful for different purposes. An important point was to make the corpus so diversified that no individual text could possibly distort the frequency figures.' (Gellerstam, 1992: 152).

[15] In 1970, when the editorial work on the dictionary actually began, the computerized corpus comprised more than 80 millions of running words. It was completed by indexes, frequency counts and concordances (Martin, 2000).

[16] Leech (1991) mentions the TLF as a pioneer corpus of written texts, but not in connection with the second wave of corpora which he says appeared in the 1970s.

[17] It should be remembered that empirical methods had been sternly criticized by Machine Translation researchers themselves, first of all by Yehoshua Bar-Hillel (1915–1975) in his report of 1960 which initiated the decline of Machine Translation (Bar-Hillel, 1960).

However, it can be understood that the Machine Translation filiation, made infamous by the ALPAC report of 1966, has been forgotten.[18]

3. The discontinuity of corpora design over 30 years

Now let us examine another point: the claimed discontinuity of corpus linguistics during 20-30 years, between the corpus work of the American structuralists in the 1950s and the revival of corpora in the 1990s. Actually the accounts diverge on this point. Leech (1992) asserts there was a break of twenty years, while in his version of 1991 he mentions the appearance of a second generation of corpora in the early 1960s. Svartvik (1992) points out that the Brown corpus of the late 1960s was followed by a string of successors in the 1970s.

In fact the publication of many text books on computerized corpora during the 1970s-80s attests to this continuity. In particular, Aarts and Meijs (1984) comprises papers about the successors of both the Survey of English Usage and the Brown corpus which had been developed in the 1970s in collaboration with Swedish, Norwegian and Dutch universities. Let us only mention the first most famous ones: in 1975, the London-Lund Corpus of Spoken English (LLC); in 1978, the Lancaster-Oslo-Bergen Corpus of British English (LOB); in 1987, the Collins Birmingham University International Language Database (COBUILD) which is also a metonymic name for the Collins COBUILD English Language Dictionary published in 1987.

To conclude this point, it can be said that there was no break between the 1960s and the 1980s or 1990s in the production of corpora. Incidentally, the text books published before the 1990s never mention the priority of the Brown corpus, nor any attempt made by Chomsky to stop corpus development.

4. Chomsky's arguments against corpora and statistics

This remark leads to our last point. According to current corpus linguists, Chomsky stopped corpus linguistics in the 1950s, so that the pioneering Brown corpus appeared in a very hostile context:

> 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).

> 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).

[18] Since the ALPAC report which put an end to MT research in the USA and in the rest of world, Machine Translation has been included in Computational Linguistics, and the first MT experiments, innovating though they might be, have hardly been mentioned (see Léon, 1999; Cori & Léon, 2002).

> The Brown Corpus was significant not only because it was the first computer corpus compiled for linguistic research, but also because it was compiled in the face of massive indifference if not outright hostility from those who espoused the conventional wisdom of the new and increasingly dominant paradigm in US linguistics led by Noam Chomsky. (Kennedy, 1998: 19).

These statements involve two assumptions. First, that the early computerized corpora, and particularly the Brown corpus, should be regarded as the revival or the continuation of American Structuralists' conception of corpora. Second, that Chomsky's criticisms equally concerned every statistical model and were virulent enough to stop any form of corpus research.

Let us examine the first assumption. In post-Bloomfieldian linguistics, a theory should aim at a systematic taxonomy of linguistic elements (distributional classes) from a corpus of observed data through discovery procedures. The critics of structuralism, most notably Chomsky, argued that these procedures yielded no more than a static inventory of signs, devoid of any significance and not allowing any theoretical explanation. The description obtained by this method was limited to the data which had been collected and led to no insight into the nature of language.[19]

Now it can be seen that the idea of corpus at work in the Brown corpus does not match the American structuralist approach. The authors and users of the Brown corpus conceived of the corpus as a set of observed utterances from which frequency counts could be investigated. No idea here of taxonomy, nor of discovery procedures. Whereas statistics were used by American structuralists to predict which sentences belonged to the corpus and which ones did not, the Brown corpus users aimed to compare frequency counts between genres, or to test general statistical models on vocabulary.

In fact, the links of the Brown corpus with American structuralism were rather loose. As Falk (2003) has shown, none of Twaddell's contemporaries included him among the structuralist linguists.[20] What is more, Twaddell does not seem to have conceived of the Brown corpus as a taxonomic machine. In his foreword to Kucera and Francis's book, he points out the advantages and disadvantages of the use of a corpus in linguistics without claiming strong theoretical views:

> The advantages and disadvantages of basing linguistic statements on a specific corpus are familiar. A corpus protects against gross lapses of recall to which introspection is liable. Statistical statements depending on subjective judgments are unreliable in the extreme. On the other hand, a corpus of manageable size will underrepresent some units and structures that introspection can supply and specify adequately. (Twaddell's foreword in Kucera and Francis, 1967: v)

[19] Recall that Hockett's views on corpora, as early as 1948, involved the notions of infiniteness of language and of projection (Hockett, 1948).

[20] In his book dedicated to the evaluation of post-Bloomfieldian achievements and Chomsky's work, Hockett ranks Twaddell among those 'whose training had been, for better or worse, somewhat freer of Bloomfield's influence' (1968: 18).

As to Quirk, he devoted more discussion to theoretical issues and to the divergences between Chomsky and the post-Bloomfieldians. In particular, Quirk and Svartvik (1966) took part in the debate, widespread among the linguists and psychologists of those days, on Chomsky's approach of competence versus performance, and grammaticality versus acceptability. Though Quirk and Svartvik supported performance and acceptability against competence and grammaticalness, they proposed an experiment for establishing degrees of acceptability in English sentences, which they used later in their grammar (Quirk and al., 1972).

Now, let us briefly examine the early arguments put forward by Chomsky against corpora in "Three Models of grammar" (1956), taken up in *Syntactic Structures* (1957), in the discussion between Chomsky and American structuralists which took place at the University of Texas (1958 [1962]), and finally in Miller and Chomsky (1963).

Note however that ten years separated the publication of *Syntactic Structure* and the Brown corpus and that Chomsky's early criticisms of the use of corpora in syntax could not concern this corpus, nor any computerized corpus since they did not exist yet at that time. Actually, Chomsky did not attack word frequency counts strictly speaking, since his main criticisms concern the use of probabilities, essentially Markov's model and Shannon's information theory, in syntactic analysis. As information theory and discovery procedures were widely debated in the 1950-60s, Chomsky's criticisms were not isolated and their reach and strength should not be overstated.

Chomsky argues that any particular corpus of utterances obtained by linguists in their fieldwork cannot be identified as the set of grammatical sentences, inasmuch as the notion of grammaticality involves those of projection, infiniteness and ideal speaker:

> 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)

When invited by Archibald A. Hill (1902–1992) in 1958 to present his linguistic model at the University of Texas, Chomsky (1962) addresses this issue a little differently, arguing that any natural corpus is skewed. If generated, it will produce non-sentences or conversely be incomplete and not provide every grammatical sentence.[21] In addition the description would be reduced to a mere list without any explanatory hypothesis.

Moreover, Chomsky claims that grammaticality cannot be identified with high statistical approximation, and criticizes descriptivists' suggestion of replacing possible sentences by highly probable sentences and impossible sentences by low probability

[21] This argument has been reported by Leech (1991, 1992), as well as Chomsky's issues developed since 1965, performance / competence and I-language / E-language, which we do not address in this paper.

sentences (Chomsky, 1957). As a matter of fact, grammaticality, which is what the grammar can account for must be distinguished from acceptability, the judgement made by native speakers.[22]

This argument is linked to the claim that English is not a finite state language and to the rejection of Markov's model as unable to isolate the set of all grammatical sentences.[23] Resuming Chomsky's stand taken in *Syntactic Structures*, Chomsky and Miller (1963) put forward detailed psychological and statistical arguments against the idea that grammar would be a Markov chain and that probabilities could be applied to syntactic structures, in particular because of the recursivity and discontinuity of natural language.

In addition it should be said that Chomsky's criticism essentially concerned the taxonomic view of corpus and discovery procedures, and not statistical methods in general. Actually, as far as they did not handle syntactic structures, Chomsky did not dispute the interest of statistics and probabilistic models in the study of language. Note that the following arguments made by Chomsky have not been mentioned by corpus linguists:

> 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. (Chomsky, 1957: 17)

> 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. Cf. B. Mandelbrot, "Structure formelle des textes et communication: deux études" *Word 10.*1-27 (1954); H. A. Simon, "On a class of skew distribution functions" Biometrika 42.425-40 (1955). (Chomsky, 1957 note 4: 17).

Miller and Chomsky (1963) agreed that Zipf's law as well as Mandelbrot's work, dealing with probabilities and word length in a text, had to be taken seriously, and their results discussed and verified: 'Miller and Newman (1958) have verified the prediction that the average frequency of words of length *i* is a reciprocal function of

[22] Later, Sidney Greenbaum (1976), one of the authors of *A Grammar of Contemporary English*, attempted to demonstrate that the acceptability of syntactic structures is influenced by their frequency of use.

[23] Named after Andrej A. Markov (1856–1922), who studied poetry as stochastic sequences of characters, a Markov chain is a sequence of random values whose probabilities at a certain time interval depends upon the value of the number at the previous time. Claude E. Shannon (1916–2001) used a Markov chain to create a probabilistic model of the sequences of letters in a piece of English text (Shannon, 1948). A Markov model of order *n* predicts that each letter occurs with a fixed probability, but that probability can depend on the previous *n* consecutive letters (*n*-gram). Since the 1950s Markov models have been used in Machine Translation and Natural Language Processing to disambiguate graphic units.

their average rank with respect to increasing length' (Miller and Chomsky, 1963: 461).[24]

Miller and Chomsky discussed Markov models more thoroughly in their paper, and agreed that, though they cannot be implemented on syntax to provide the set of grammatical sentences, they can be applied to lower-level production, such as phonemes, letters and syllables:

> 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)

Note that Kucera was concerned with Information Theory (Kucera, 1963) and used a Markov model in a comparative phonological study of Russian, Czech and German (Kucera and Monroe, 1968). He agreed with Chomsky's view that this type of model could only be applied to lower-level units and not to syntax and sentences.

To conclude this point, it has been shown that the Brown corpus could neither be a descendant of the taxonomy-oriented methods advocated by the post-Bloomfieldians nor the real target of Chomsky's criticisms. Concerning statistics and probabilities, Chomsky found certain types of statistical works quite valuable so far as they do not deal with syntax. Therefore, corpora like the Brown corpus, dedicated to word frequency counts, were not Chomsky's concern.

Conclusion

It has been shown that retrospective construction of a history aiming to legitimize Corpus Linguistics as an autonomous discipline, rests on a fair number of assertions and omissions. The Brown corpus has been presented as the key precursor by omitting pioneer works in Machine Translation or in computerized corpora in the area of dictionary making. Legitimization has been achieved by placing Corpus Linguistics at the heart of a revival of 1950s empiricism, in particular by making it a follower of the post-Bloomfieldian tradition. Still it was seen that the Brown corpus was submitted to

[24] George Kingsley Zipf (1902–1950) was a behaviorist linguist at Harvard and the founder of 'Dynamic Philology'. Zipf's law has been much used in statistical studies of vocabulary. 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 (Zipf, 1949).
Benoît Mandelbrot (b. 1924), a French mathematician, developed a statistical model which provided a theoretical explanation for Zipf's law (Mandelbrot, 1954).

objectives quite different from any taxonomic machinery, in so far as its main concern was word frequency counts or statistical model testing.

Chomsky's arguments against the post-Bloomfieldians have been used to explain an alleged hiatus of corpus production which did not really occur. In fact, it can be assumed that there was no discontinuity between the present annotated corpora and vocabulary count corpora which were flourishing throughout the early twentieth century. On the other hand, Chomsky's criticism of corpora and statistical methods did not concern vocabulary counts. Rather, he seemed to find the use of Markov and word statistics models quite valuable, as far as they did not deal with syntax. Recall too that Kucera, one of the Brown corpus's authors, agreed with Chomsky on this point, using Markov's model in a comparative study of phonemes.

Connecting the first computerized corpora to the American tradition alone, in addition to being unfounded, has serious consequences. By ignoring the strong empiricist British filiation inherited from Firth's work, Corpus Linguistics has been deprived of a real precursor. Instead, two restrospective constructions were forged at the moment when NLP was technologically ready to invest in the field of corpora: a theoretical anti-precursor, one of the most famous theoretical linguist, i.e Chomsky; and a technical precursor, in fact a product, the Brown corpus.

It remains to be explained why the use of corpora, which has undeniably seen an unprecedented technological development and is valuable in most linguistic areas, absolutely needs be built up as an autonomous discipline. Another important issue is how to appraise the real impact of the increasing power of computers and the availability of linguistic data. Although corpus linguists invoke these technological developments as revolutionary for linguistic research, their real significance has as yet hardly been evaluated.[25]

References

Altenberg, Bengt. 1991. 'A Bibliography of Publications Relating to English Computer Corpora'. In: *English Computer Corpora: Selected Papers and Research Guide*. Ed. by Stig Johansson and Anna-Brita Stenstrom. Berlin: Mouton. 355–96.

Aarts, Jan & Willem Meijs. 1984. *Corpus Linguistics: recent developments in the use of Computer corpora in English Language Research*. Amsterdam: Rodopi.

Bar-Hillel, Yehoshua. 1960. 'The Present Status of Automatic Translation of Language' In: *Advances in Computers vol.1*. Ed. by Franz L. Alt. New York & London: Academic Press. 91–141.

Baayen, Harald. 2003. 'Probabilistic Approaches to Morphology' In: *Probabilistic Linguistics*. Ed. by Bod Rens, Jennifer Hay and Stefanie Jannedy. Cambridge & London: MIT Press. 229–287.

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

[25] It seems a little difficult to follow Baayen (2003) when he claims that formal languages were developed in the 1960s instead of corpus investigations because at that time computers had computer power but no memory capacities. It seems exaggerated to claim that technological developments alone are responsible for theoretical orientations in language sciences.

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

-----. 1957. *Syntactic Structures*. The Hague: Mouton.

-----. 1962. 'Transformational Approach to Syntax.' *Third Texas Conference on Problems of Linguistic Analysis in English May 9-12, 1958, Studies in American English*. Austin: The University of Texas. 124–158

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

Cori, Marcel and Jacqueline Léon. 2002. 'La constitution du TAL. Etude historique des dénominations et des concepts'. *Traitement Automatique des Langues* 43-3: 21–55.

Falk, Julia. 2003. 'Turn to the history of linguistics. Noam Chomsky and Charles Hockett in the 1960s'. *Historiographia Linguistica.* 30: 129–185.

Fillmore, Charles J. 1992. ''Corpus linguistics' or 'Computer-aided armchair linguistics''. In: *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*. Ed. by Jan Svartvik. Berlin, New York: Mouton de Gruyter. 35 –60.

Francis, W. Nelson. 1998. 'A Pilgrim Progress: From philology to linguistics'. In: *First Person Singular III*. Ed. by E. F. K. Konrad. Amsterdam, Philadelphia: Benjamins. 59–70.

Garside, Roger, Geoffrey Leech, and Tony MacEnery. 1997. *Corpus Annotation, Linguistic Information from Computer Text Corpora*. London and New York: Longman.

Gellerstam, Martin. 1992. 'Modern Swedish Text Corpora'. In: *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*. Ed. by Jan Svartvik. Berlin & New York: Mouton de Gruyter. l49–163.

Greenbaum, Sidney. 1976. 'Syntactic Frequency and Acceptability'. *Lingua* 40: 99–113.

Guiraud, Pierre. 1954. *Bibliographie critique de la statistique linguistique*. Utrecht & Anvers: Editions Spectrum.

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

-----. 1968. *The State of the Art*. The Hague: Mouton.

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

Kucera, Henry. 1963. 'Entropy, redundancy and functional load in Russian and Czech'. In: *American contribution to the fifth international congress of Slavists, Sofia*. The Hague: Mouton & Co. 291–318.

-----. 1998. 'A Linguistic Round Trip: From practice to theory and back'. In: *First Person Singular III*. Ed. by E. F. K. Koerner. Amsterdam, Philadelphia: Benjamins. 81–96.

Kucera, Henry and George K. Monroe. 1968. *A Comparative Quantitative Phonology of Russian, Czech and German*. New York: American Elsevier Publ.

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

Leech, Geoffrey. 1991. 'The State of the Art in Corpus Linguistics'. In: *English Corpus Linguistics: Studies in Honour of Jan Svartvik.* Ed. by Aijmer Karin & Bengt Altenberg, London & New York: Longman. 8–29.

-----. 1992. Corpora and theories of linguistic performance. In: *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991* Ed. by Jan Svartvik. Berlin, New York: Mouton de Gruyter. 105–122.

-----. 2002. 'Geoffrey Leech' In: *Linguistics in Britain: personal histories.* Ed. by Keith Brown & Vivien Law. Oxford: Publications of the Philological Society: 155–169.

Léon, Jacqueline. 1999. 'La mécanisation du dictionnaire dans les premières expériences de traduction automatique (1948–1960)'. In: *History of Linguistics 1996 vol. II.* Ed. by David Cram, Andrew Linn and Elke Nowak. Amsterdam & Philadelphia: John Benjamins. 331–340.

Lexicologie et lexicographie françaises et romanes. Orientations et exigences actuelles. 12-16 nov. 1957. 1961. Paris: Ed. du CNRS.

McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics.* Edinburgh: Edinburgh University Press.

Mandelbrot, Benoît. 1954. Structure formelle des textes et communication. *Word* 10 (3): 1–27.

Manning, Christopher and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts & London: MIT Press.

Martin, Robert. 2000. 'Le Trésor de la langue française'. In: *Histoire de la Langue Française (1945–2000)* Ed. by Gerald Antoine & Bernard Cerquiglini. Paris: CNRS Editions. 969–979.

Miller, George A. and Noam Chomsky. 1963. 'Finitary Models of Language Users'. In: *Handbook of Mathematical Psychology Vol. II.* Ed. by R. Duncan Luce, Robert R. Bush and Eugene Galanter. New York: Wiley. 419–491.

Oostdijk, Nelleke and Pieter de Haan (eds.). 1994. *Corpus-based Research into Language*. Amsterdam, Atlanta: Rodopi.

Quirk, Randolph. 1960. 'Towards a description of English Usage'. *Transactions of the Philological Society.* 40–61.

-----. 2002. Randolph Quirk. In: *Linguistics in Britain: personal histories.* Ed. by Keith Brown and Vivien Law. Oxford: Publications of the Philological Society. 239–248.

Quirk, Randolph & David Crystal. 1966. 'On scales of contrast in connected English Speech'. In: *Memory of J.R. Firth* Ed. by Charles E. Bazell et al. London: Longman.

Quirk, Randolph, A.P. Duckworth, Jan Rusiecki, Jan Svartvik & A.Colin. 1964. 'Studies in the correspondence of prosodic to grammatical features'. In: *Proceedings of the Ninth International Congress of Linguists.* The Hague: Mouton: 679–691.

Quirk, Randolph & Joan Mulholland. 1964. 'Complex prepositions and related sequences'. *English Studies* 45: 64–73.

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

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

Rosenblatt, Frank. 1958. The Perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65: 386–408.

Shannon, Claude E. 1948. 'A Mathematical Theory of Communication'. *The Bell System Technical Journal* 27: 379–423, 623–656.

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

Simpson Rita C. and John M. Swales (Eds.). 2001. *Corpus linguistics in North America: Selections from the 1999 Symposium*. Ann Arbor: University of Michigan Press.

Svartvik Jan (ed.). 1992. *Directions in Corpus Linguistics. Proceedings of Nobel Symposium, 4-8 August 1991*. Berlin & New York: Mouton de Gruyter.

Zipf, George Kingsley. 1949. *Human Behaviour and the Principle of Least-Effort*. Cambridge, Massachusetts: Addison-Wesley

Contact Details: jacqueline.leon@linguist.jussieu.fr

Jacqueline Léon
Université Paris 7
UMR7597, Histoire des Théories Linguistiques
case 7034
2, place Jussieu
75005 Paris
France