CORELA

The Corpus: A Tool among Others

Publié en ligne le 25 novembre 2013

Par Tobias SCHEER

Résumé

Le corpus a toujours été, est et sera un outil précieux qui aide à poursuivre un but. Son statut ontologique d'outil ne changera pas, aussi fabuleux soient la puissance computationnelle, la capacité de stockage, la rapidité d'accès et de transmission, et quelle que soit sa taille. Le corpus est une source de données parmi d'autres (dont, notamment, les jugements de grammaticalité) qui a des avantages et limitations spécifiques dont l'usager doit être conscient – au même titre que pour n'importe quel autre outil.

Noyé dans l'utilitarisme et l'hystérie autour des projets qui de nos jours dominent l'academia, beaucoup croient, ouvertement ou tacitement (ou encore à leur insu), que la recherche qui implique la construction d'un corpus assortie d'une exploitation computationnelle "puissante" est plus sérieuse que celle qui n'en implique pas. D'aucuns même pensent que la finalité d'un projet de recherche peut se réduire à la création d'un corpus, que le corpus produira la science par ses vertus intrinsèques, et qu'il peut donc se substituer au raisonnement et à la dialectique entre données et hypothèse. La même idéologie promeut l'idée qu'aucun énoncé ne peut être scientifique sans être statistiquement significatif. C'est ici que le corpus cesse d'être un outil, i.e. où le système bascule dans la folie. Depuis une décennie ou davantage, c'est ce qui s'est passé : sans avoir rien demandé, le pauvre corpus se retrouve au milieu d'un blizzard et se fait abuser par l'idéologie dominante.

Abstract

The corpus was, is and will be a valuable tool that helps pursuing a goal. Its ontological status as a tool will not change, no matter how fabulous the computational power, storage capacity, access and transmission speed, and whatever the size of the corpus. The corpus is a data source among others (namely grammaticality judgements), which has specific advantages and limitations that the user needs to be aware of – like for any other tool.

Drowned in the ambient utilitarianism and project-hysteria, many people believe, overtly or tacitly (or without being aware that they do), that research which involves the building of a corpus coupled with exploitation by a "powerful" computer programme, is more serious than a competitor which does not. Some even believe that the whole purpose of a research project may be the creation of a corpus, and that the corpus will produce science by itself, i.e. substitute itself to reasoning and the data-expectation dialectic. The same ideology promotes the idea that whatever scientific statement is made, it needs to be statistically relevant. This is where the corpus stops being a tool, i.e. where the system goes mad. And it did on a large scale in the past decade or so. Poor corpora are in the middle of this thunderstorm, and are abundantly abused by the ideology in place.

Mots-clés : hystérie autour des projets, corpus, financement de la recherche, loi de Dürrenmatt., outil, politique de recherche irrationnelle
Keywords : corpus, Dürrenmatt's law., fund raising, irrational research politics, project hysteria, tool

Sommaire

1. Things that ought to be trivial

1

The general idea exposed in the pages below ought to be quite trivial, but may be less so in the current landscape: the corpus was, is and will be a valuable tool that helps pursuing a goal. Its ontological status as a tool will not change, no matter how fabulous the computational power, storage capacity, access and transmission speed, and whatever the size of the corpus. Like all other scientists, linguists have been, are and will be keen to base their reasoning on the best data possible, i.e. data which are reliable, significant, exhaustive, fine-grained etc. The corpus is a data source among others, which has specific properties, i.e. advantages and limitations. The user needs to be aware of these when dealing with corpora. This again is quite trivial a statement, since everybody who inquires into something should be aware of the properties, limitations, and eventual bias-introducing shortcomings of the instrument used. The same is of course true for other sources of evidence such as grammaticality judgements.

2

A related but distinct issue is the fact that "the corpus" is not a monolithic thing: there are many different ways of building corpora, and there are many different ways in which corpora may be used. The result of corpus-based studies is a function of the design properties of the corpus, and of the way the corpus is used. But again, this is all just regular common sense: bad data make hardly a good theory, and bad usage of good data does not either. The corpus is not good or bad *per se*, it can provide some kind of evidence but is unable to produce other relevant information.

3

As in all other scientific inquiry, and especially in the adult (or successful) sciences, advances in understanding how language works are based on the dialectic tension between observation and expectation/theory. It is trivially true that data may and should falsify theories, and hence that better data, i.e. data which are more exhaustive, more fine-grained, more representative etc., are better referees. This is where the technological progress produced by searchable electronic corpora is useful. It is also trivially true, however, that "le point de vue crée l'objet" (Saussure: *the point of view creates the object*). That is, one may stare at a pattern for ages without understanding in which way it makes sense because one is not looking at it through the right lens.

4

The conclusion, then, is very simple and again trivial: like all other areas of scientific inquiry, linguistics needs to be fed with reliable, significant, representative and if possible exhaustive data. Like all other scientists, the linguist builds generalizations and theories on all data available, whatever their source as long as it is valid. Going along with Feyerabend (1976), any source of evidence is a possible source, and argument will decide whether it should be used or not. Astrological evidence *is* a possible candidate for input data to linguistic reasoning, but it won't pass the filter of argumentation. As far as I can see, there is no conclusive argument that discards the corpus or grammaticality judgements as such. Hence both can and should be used (as much as other sources of evidence) – but when they are, users should be aware of their properties and limitations.

2. The corpus and real-world issues

2.1. When the system goes mad: corpora in the midst of "employability" and "societal demand"

5

Corpora also have a number of very real-world properties these days, since they are relevant in funding competition and decision, as well as for careers and the structuring of scholarly institutions. Drowned in the ambient utilitarianism and project-hysteria, many people believe, overtly or tacitly (or without being aware that they do), that research (and especially a "project") which involves the building of a corpus coupled with exploitation by a "powerful" computer programme (or even better: *surpuissant* in French), is more serious than a competitor which does not. Some even believe that the whole purpose of a research project may be the creation of a corpus, and that the corpus (together with the computational power of the search engine) will produce science by itself, i.e. substitute itself to reasoning and the data-expectation dialectic. Project-based science must produce "deliverables", i.e. real-world objects that one can touch and put on a website, like corpora. Merely advancing understanding and publishing it in peer-reviewed journals is of course not a sound "deliverable". Finally, the same ideology promotes the idea that whatever scientific statement is made, it needs to be statistically relevant: statistics are the ultimate proof in science (ask Einstein...). This is where the corpus stops being a tool, i.e. where the system goes mad. And it did on a large scale in the past decade or so. Poor corpora are in the middle of this thunderstorm, and are abundantly abused by the ideology in place.

6

At least for about a decade, we have lived through a period of intellectual decline where technology and data are confused with science. Rather than stemming from scientists (who however accommodate themselves in the new market-driven environment without too much mourning), this confusion is deliberate, organized and imposed by political decision makers. These do not belong to individual countries, but to larger entities like the European Community and its tentacular sub-organizations, or the so-called Bologna Process (dating back to 1999, currently 47 countries adhering). A common feature of all these is that the acting individuals are elected by nobody: they are anonymous technocrats whose proposals are never discussed in public before they are applied by individual governments or groupings thereof. An example is the OECD (Organization for Economic Cooperation and Development), whose slogan is "better politics for better lives", which has played an important role in what Nicolas Sarkozy called the "piloting of science", i.e. the idea that higher education and research are only legitimate if they respond either to demands of the market or of society. Byzantology is worth nothing because it does not augment the "employability" of students, and it does not help urban mobility, curing Alzheimer disease or making smartphone apps. [1]

7

The transformation of universities into simple institutions that produce the human material needed by "the market" in order to run capitalism is in full swing. The original idea (dating back to Antiquity and the Renaissance when modern Universities were founded) that (higher) education is a necessary ingredient in the development of humans and indispensable in a democratic society where citizens need to make up their mind about general issues when they vote – all this is silently and tacitly trashed. The only relevant issue today is the employability of students, and whether research can be used by industry or satisfies a "societal demand" (French newspeak: *demande sociétale*).

8

The result is fake research: scientists apply for project money pretending that they are interested in one thing when they are interested in another, and then try to spend money and time on the latter as much as they can. Since they still need to pretend that they worked for what they announced, they spend their official work time on what they do not want to do, and do the real research – the one that is motivated by curiosity and the wish to understand how nature works – at night and in their free time. This was exactly the situation of science in the Middle Ages before the State made it a public affair: either scientists had a personal fortune (e.g. Descartes), or they were covered by Church, or they had a day-time job such as librarian, secretary, teacher of children from noble families, accountant etc., and did their science in their free time (Kepler, Copernicus).

9

People who drive scientists to do fake research know very well that in the history of science, typically a discovery was completely useless for industry, "the market" or society by the time it was made. It is only long after the death of the discoverer, sometimes centuries later, that the knowledge is condensed into something that can be sold on the market or does good to society (GPS based on Einstein's theory of relativity is an oft-quoted example). But the decision makers in question put a lot of energy into ignoring this fact, short-lived as they are with a two- or three-year horizon before they need to be reelected, staring at the monthly unemployment statistics and living at the capitalist rhythm of short-term profit. It is the same people who have happily contracted public debts over decades, knowing very well that they cannot be paid back by future generations, but without being bothered by this fact because they will be out of office, retired or dead when the system eventually crashes.

2.2. Instruments serve a purpose and are theory-dependent

10

The corpus stands in the midst of all this, and its status in research is impacted by the ideology described. That is, corpora and the computational instruments associated follow the law of all cutting-edge technology: there is hype and enthusiasm around its sole technological properties, and there is the naïve, Titanic-based positivist belief that high-tech will produce results by its own. We all know, and history (of science) has shown, that it does not. Advances are made when technology serves a purpose, a hypothesis or a goal: there is no science outside the realm defined by the observation-expectation dialectic.

11

For instance, physicists may put a lot of energy, money, devotion and sophistication into constructing the tools that they need, for example a particle accelerator. They never lose sight of the fact, though, that having built the CERN machine for example serves a purpose and is only one step on a way that may well lead nowhere. In order to know, they need to put their tool to use, and in order to do so, they need to design an experiment that complies with the technical properties of the machine and promises a result: they need a hypothesis, and a theory. And they need to know what they are looking for. Browsing data when you do not know what you are looking for is putting yourself in the aforementioned situation where somebody may stare at a pattern for ages without recognizing its contours. [2]

12

In other words, machines, and more generally instruments, are always designed for a specific purpose and with specific expectations: people want to prove or disprove something, or they want to understand how something works. That is, an intrinsic design property of corpora is the goal that is expected to be achieved with their help. Therefore the instrument is never neutral, and will never produce "raw" data. The myth of the existence of objective, uninterpreted or raw data is typically used in order to discredit a group of people from different theoretical or philosophical quarters, or who use a different methodology (e.g. corpus vs. elicitation, phonetics vs. phonology etc.). The difference between distinct instruments, though, is not that one produces objective, exact and reliable data, while the other is biased – it is only the fact that the bias (i.e. what exactly lies between the observer and the real world) of one party is made explicit, while the one of the other is denied and tried to be kept hidden under the rug.

13

That there is no such thing as a one-to-one blueprint of reality is not only due to the instrument that links the real world to the observer. It is also a fact established in philosophy at least since Kant: humans can observe the real world (thing-in-itself, or noumenon) only through the perception of one of their five senses (and this is true whatever sophisticated aiding machines will be plugged in). The five senses thus stand in the way of a direct perception, and we know for sure that they are not reliable: many established facts such as categorical perception, the McGurk effect or dichotic perception show that the human percept may be dramatically distinct from the signal that has reached his senses. That is, the reality that humans talk about is never the real world itself, but properties thereof reworked and augmented by some mechanism of our cognitive and perceptual apparatus, whose workings we do not understand today.

14

Current quantum physics is entirely based on this: the fact of observing modifies the object observed, to the effect that there is no such thing as an observational fact independent of the observation, and hence of the observer. Another way of putting this confirmation of the Kantian insight is this: "quantum mechanics requires interpretation before it describes the experience of an observer. [...] [T]he behavior of a system after observation is completely different than the usual behavior" (Wikipedia).

15

There is no reason, though, to believe that man will be unable to make advances in the understanding of himself or the world around him. Scientific understanding has always been made by people who were drowned in systems of belief, typically of a religious kind, and therefore had strong expectations and preconceptions. Reason and fact always ended up prevailing, even if it is true that institutional and belief-related brakes may have slowed down the emergence of understanding. Hence there is nothing wrong with investigators being engaged in systems of belief, which may strongly structure the way they proceed in order to know: Feyerabend (1976) explains that any motivation for setting out to discover is a good motivation, and the larger the spectrum, the better for science. One thing that can and ought to be done, though, is to be aware of, and to make explicit, the kind of bias that exists.

2.3. Dürrenmatt's law: technology for technology's sake and ensuing irrational behaviour

16

Friedrich Dürrenmatt's play *The Physicists* is about the law that when knowledge and technology are available, they will always be used no matter what. The physicist Möbius has discovered the "Principle of Universal Discovery". Knowing that its spreading will provoke murder and disease, he hides in a home for the mentally ill. He is spied on by other "patients", though, who work for leading states, and will be unable to keep his knowledge secret. Applied to a current issue, the law predicts for example that the minute man is able to clone man, clones will exist and proliferate. As much as pre-natal selection of humans based on everything that can be detected (or that people believe they can detect: gender, colour of the eyes, size, diseases, homosexuality etc.) is a reality today no matter what laws or ethics committees say, just because the technology necessary is available.

17

Everybody knows that raw extractions from Google cannot be used for linguistic inquiry because of a number of caveats, the most obvious and most invalidating being the fact that there is no control over the identity of those who produced the material: nobody knows what they are native speakers of (or indeed whether they are humans at all: machines translate webpages automatically). Nonetheless, Google-based data are constantly used in the literature, typically preceded by the mention that the author is aware of the caveats. [3] Technology will be put to use just because it exists, no matter whether this is reasonable or not.

18

Everybody (who wants to know) knows that the Shanghai ranking is heavily based on Nobel Prizes, and that there are no Nobel Prizes in many disciplines, typically in the Humanities. [4] Nevertheless, the sole existence of the ranking, and its availability upon a mouse click for people who have no idea about academics but need to distribute money, make the ranking the absolute reference for officials and decision makers, who engage large-scale destructions of the academic landscape on the grounds of what they believe are reliable, objective and measurable facts. France is a case in point: since the presidency of Chirac, the country is engaged into a long-term programme that seeks to create bigger universities by forcing existing universities to fuse (the results are called PRES or Idex). The goal is to "be competitive" internationally, and to mechanically move up in the Shanghai ranking because more Nobel Prizes and more publications under the same roof make better Shanghai scores. The content does not matter, thus, it is the same: the only thing that matters is the Shanghai showcase and its media impact. [5] A ranking that exists will be used, no matter what it measures and what its accuracy.

19

An even more striking (and non-academic) example of Dürrenmatt's law is Klout. Klout is an Internet-based company that promises to measure the social impact that people have in this world ("Klout measures your influence based on your ability to drive action on social networks", Klout webpage, 28 Sept. 2012). A Klout score from 1 to 100 is attributed to every single individual on the planet that the company can get hold of, based on automatic extraction of information from social networks (mainly Facebook and Twitter: "[t]he Klout Score incorporates more than 400 signals from seven different networks", Klout webpage, 28 Sept. 2012). Customers such as head hunters or human resource managers pay in order to access the Klout score of people they may hire, and they do that for exactly the same reasons that lead politicians and ministry-technocrats to push the Shanghai-button: they are incompetent, they have no time and they do not want to bother doing the evaluation themselves – somebody else has done the work already. The thing is that unlike the Shanghai authors who explain how their ranking was built, the Klout algorithm is secret: nobody knows what exactly is counted, how factors are weighted etc. Given Klout's commercial success, visibly this does not prevent supposedly rational people from using the opaque Klout score for making decisions. A ranking that exists will be used, no matter whether it is arbitrary or not.

20

Consider a final example, back to academia: the European Science Foundation has created a Standing Committee for the Humanities, which builds a European Reference Index for the Humanities (ERIH). The purpose of this index is to establish a list of relevant journals for various disciplines, where individual journals are ranked along a three-point scale A, B, C. The authors of the 2007 edition of the index for linguistics introduce the list with the explicit mention that "[a]s they stand, the lists are not a bibliometric tool. The ERIH Steering Committee and the Expert Panels therefore advise against using the lists as the only basis for assessment of individual candidates for positions or promotions or of applicants for research grants." But this is of course *exactly* what happened: the existence of a ranking or a list will automatically lead to its application, no matter what the content, how they were built, whether they are significant or accurate etc.

21

All this is entirely *irrational behaviour* in our supposedly rational, academic world where actors have benefitted from super-high education – but this is how things work, or rather, how humans work. Relevant for our subject, corpora, is that they have had the status of cutting-edge high-tech for some time now, and will continue to have it in the foreseeable future. There is thus reason to be suspicious about the Dürrenmatt-effects associated, which are inescapable.

22

This brings us back to the general line of this paper: the corpus is a tool, nothing more, nothing less. Its sole existence is not a scientific result, and the significance of its contribution to science depends on its design properties, as well as on how it is used.

23

Against this backdrop, the remainder of the article discusses a number of more specific issues related to corpora.

3. More or less direct access to data for different linguistic disciplines

24

Owing to their intrinsic properties, different linguistic disciplines are more or less far removed from data sources. Phonology (and probably non-inflectional morphology) need to construct their object of inquiry much more and much more carefully than syntax (and inflectional morphology) (Scheer, 2004). This is because phonologists can never be sure whether a given alternation is the result of phonological computation, allomorphy, analogy or distinct lexical recordings. Only in the former case is it a valid window on how phonology works.

25

For example, the bare existence of the two words *electri[k]* and *electri[s]ity* in English does not allow us to conclude that there is a phonological computation relating k and s (Halle, 2005, Green, 2007:175ff). The alternation may be due to a grammatical (but non-phonological) computation, i.e. allomorphy, to non- or para-grammatical activity (analogy), or to no computation at all in case *electricity* is morphologically non-complex, i.e. a single lexical recording.

26

Setting idioms aside, syntax does not have this grievance: every sentence that is uttered is the result of online syntactic computation. [6]

4. Data are an artefact, not a natural object: they are *always* constructed

27

Data are the result of a human construction, not a thing that is found in nature. As was shown above, this applies to all scientific inquiry in a broad, kantian sense and is an essential of current physics.

28

This being said, it also structures much more narrowly the everyday work of the linguist: when the PFC corpus (Laks, 2011) is coded for, say, schwa, there are numerous cases where the value of a sound cannot be determined, even when the number of transcribers is multiplied. In case it is decided that a sound is a schwa, whether or not it is coded as such depends on a number of further decisions, since it may also represent a transitional sound in word-final position, rather than a vowel that is linguistically relevant. Hence it is the linguist, not the real world, who decides which real-world item is knighted a piece of linguistically relevant data, i.e. has the right to impact linguistic reasoning.

29

The same applies to the *electric* - *electricity* example: before anything can be analyzed at all, a decision needs to be made regarding the question whether or not both items entertain a derivational relationship in phonology. This is not anything that may be decided by a corpus or real-world properties of the items in question. Only reasoning and (theoretical) assumptions can show the way. Regarding the specific issue of drawing a red line between the four mechanisms at hand (phonological computation, allomorphic computation, analogy, independent lexical recordings), no criterion is in sight that would allow the linguist to make a firm decision in all cases. In the 70s, phonologists attempted to define such a criterion, called the evaluation metric (or measure), without success (e.g. Kiparsky, 1974, Campbell, 1981, Goyvaerts, 1981).

5. Limitations: relevant information that the corpus cannot provide

30

Different data sources have their specific strengths and limitations. There is a large body of literature on what grammaticality judgements can and cannot do, how they may be biased, how they should or should not be used etc. (e.g. Botha, 1981, see the discussion in Durand, 2009). Dangers and limitations of grammaticality judgements and elicitation are due to the fact that they are partly the result of conscious activity, which produces the following caveats:

31

1) impact of normative elements;

32

2) impact of sociological parameters;

33

3) the fact that a good informant needs to be tutored before he is able to inform.

34

Corpora are in the same situation. Below is a (non exhaustive) list of things that they cannot, and will never be able to do.

5.1. Corpora cannot attest the absence of something

35

A defining property of corpora is the fact that they are finite. Hence they can assert the presence of X, but not its absence in a language: by definition, there is life outside the corpus that the corpus is blind to. The fact that X does not occur in a corpus, however multi-billion-item it is, does not mean that X is agrammatical, or irrelevant.

36

This especially impacts fields of inquiry such as syntax where the number of well-formed items is infinite, and hence where most of what grammar can generate will never be attested. Relevant for the study of grammar is what is attestable, not what is actually attested. Grammaticality judgements fill the gap: they can check the non-attested space.

5.2. Corpora can only record performance

37

Corpora can only record performance: they will never be able to provide direct access to competence. If it is true that what linguists are after is competence, and that performance is but a shadow on a Platonian cave wall that needs to be interpreted in order for the real object to be discovered, corpora can only do the first step of inquiry. By contrast, grammaticality judgements open a direct window on competence.

38

It is well-known, for example, that performance produces a lot of irrelevant noise, i.e. attested items that must not be used as input data to reasoning (e.g. Sampson, 1978). The string "that want cat" is not well-formed in English, but may perhaps be attested. All linguists will immediately discard it from the set of input data to reasoning, and their decision will be based on prior knowledge, i.e. their intuition as native speakers. In other words, producing valid and significant input data based on a corpus requires the linguist in charge to work hand in hand with grammaticality judgements.

5.3. Liaison and h-aspiré: corpora cannot detect emphasis

39

A specific example of what corpora cannot do comes from liaison properties of h-aspiré words (Encrevé & Scheer 2005). The generalization is that h-aspiré words may produce a glottal stop if preceded by a C-final word, as under a. No glottal stop is possible after V-final words b, or with words that do not have a h-aspiré c.

40

(1) liaison and h-aspiré

41

a. quelle [ʔ] housse

42

quel [ʔ] hêtre

43

b. une jolie *[ʔ] housse

44

un joli *[ʔ] hêtre

45

c. quelle *[ʔ] armoire

46

quel *[ʔ] homme.

47

However, all asterisked forms do in fact exist and are attested – but this is only when the nouns have an emphatic meaning as in contrastive focus (indicated by upper case) under below.

48

(2) quelle [ʔ] ARMOIRE

49

quel [ʔ] HOMME

50 une jolie [ʔ] HOUSSE

51 un joli [ʔ] HEROS

52 une jolie [ʔ] ARMOIRE

53 un joli [ʔ] HOMME.

54 The simple attestation of items with a glottal stop in a corpus will never produce this generalization, however large and sophisticated the corpus. This is because the corpus cannot make a difference between emphatic and non-emphatic meaning. For this we need a human (and his intuitions) who decides (e.g. by coding a corpus for this property).

6. There is no datum vs. exemplum – there is just good and bad empirical work

55 Bernard Laks has argued for a distinction between two kinds of data, the *datum* and the *exemplum*. He holds that generative linguistics, broadly speaking, is an ill-inspired exemplum-interlude ("armchair linguistics") in serious scientific endeavour. Serious work in linguistics was always based on datum, and the field has blessedly returned to this perspective since the turn of the 21st century (Laks 2008, Laks Ms 2011, Laks & Calderone Ms 2012). According to Laks the watershed line is Zellig Harris' *Methods in Structural Linguistics* (Harris 1951): this is when serious datum-linguistics was replaced by untrustworthy exemplum-armchair-generativism.

56 Opposing datum and exemplum does not make sense. Conceptually, there is nothing to be opposed: exemplum is the step in the construction of knowledge that logically follows the acquisition of the datum, and is based on it. Empirically, there is serious empirical work after 1951, and non-serious empirical work before 1951.

57 The meaning of the word *example* is explicit by itself: it does not refer to just a few items of evidence (as opposed to a large empirical record on the datum side), as Laks implies. Trivially, examples are *exemplary*: they surely refer to only a few items of evidence, but the author who quotes them takes on the responsibility that these items are representative of the full empirical record. If this promise is not brought home, the author has done a bad job – but this does not tell us anything about whether or not quoting examples is a good or a bad thing to do.

58 Examples exist in order not to drown the audience in a useless and never ending flow of repetitive data: a few representatives of each significant class or pattern are shown. Examples are logically based on a larger pool of data, and they suppose an *analysis* over this pool: first patterns need to be identified, then their relevance needs to be established. The data pool by itself may be amorphous, but examples are not: they

are the result of reasoning, of analysis and of theory. Examples enhance the work of everybody: of the analyst, who knows where the problems lie and what needs to be accounted for; of the audience, which is given the same information by means of a few items. All sciences of all times have always reduced data sets to a few significant examples.

59

Since data are always constructed (see section 4), building knowledge involves three mappings:

60

1) input: real-world items, output: data;

61

2) input: data, output: examples (patterns);

62

3) input: examples (patterns), output: theory.

63

Hence there is no difference between practice A which is not serious because it bases theories on a few pieces of data only, and practice B which is serious because it builds on the full empirical record. There is only a difference between solid and non-solid empirical work. And, secondarily, there is a difference between work that discusses relevant pieces of data that have been cautiously chosen and represent whatever is significant, and work that reviews endless streams of amorphous data.

64

Needless to say that it is also not the case that no solid empirical work was done by generative linguists, or after 1951: making such a claim is being unkind to thousands of linguists who have filled up endless notepads while doing fieldwork, or who have built extensive databases that try to be exhaustive in a specific area.

65

Conversely, it is not true either that there was no non-solid empirical work before 1951. A famous case in point is a 1942 paper by Martin Joos (Joos, 1942), which reports the existence of a "dialect B" in Canadian English regarding a phenomenon called Canadian Raising. Joos' article is three pages long and was published in *Language*; it is based on a few words collected, as the author says, in a highschool classroom. Joos' data have made an important career, since they were uncritically quoted, taken over and spread by generativists: in 1989 they reappear as Bromberger & Halle's (1989) key witness showing that phonological computation executes instructions in a chronological order (ordered rules).

66

The trouble is that there is no evidence independent from Joos' three pages that dialect B has ever existed: in the 1970s, Canadian dialectologists could not find any trace of it. Kaye (1990) therefore concludes that either all speakers of this dialect died out naturally before the age of 40, or that using this particular rule order is lethal. Dialect B is thus a case where a whole field was taken hostage by 1) a structuralist who did bad empirical work before 1951, and 2) generativists who gullibly repeated bad data without checking them.

67

Trubetzkoy's (1939) *Grundzüge* is another famous case of exemplum-based reasoning, by a structuralist and before 1951. The author almost exclusively quotes second hand evidence from languages that he does not know and has never worked on, and he typically does not quote a few, but *zero* words or items: vocalic systems are reported on the basis of descriptive literature without quoting a single word of the language in question (e.g. p.111f for the Central Chinese dialect of Siang-tang). Trubetzkoy did the best he could: he used the data that were available to him (often extracted from the anthropological literature), and he used only those that he judged reliable (discussion is often provided regarding this issue). He may have been, and surely was, wrong on a number of occasions, when his sources turned out not to be waterproof. This way of browsing a large number of languages (210 are mentioned in the language index) is often found in generative work of the past 15 years or so, where several hundreds of languages are browsed, typically in Ph.Ds (e.g. Kirchner, 1998).

7. What is corpus linguistics?

68

Noam Chomsky has triggered a polemic regarding corpus linguistics: according to him, there is no such thing.

> *"Corpus linguistics doesn't mean anything. It's like saying suppose a physicist decides, suppose physics and chemistry decide that instead of relying on experiments, what they're going to do is take videotapes of things happening in the world and they'll collect huge videotapes of everything that's happening and from that maybe they'll come up with some generalizations or insights. Well, you know, sciences don't do this. But maybe they're wrong. Maybe the sciences should just collect lots and lots of data and try to develop the results from them. Well if someone wants to try that, fine. They're not going to get much support in the chemistry or physics or biology department. But if they feel like trying it, well, it's a free country, try that. We'll judge it by the results that come out. So if results come from study of massive data, rather like videotaping what's happening outside the window, fine-look at the results. I don't pay much attention to it. I don't see much in the way of results. My judgment, if you like, is that we learn more about language by following the standard method of the sciences. The standard method of the sciences is not to accumulate huge masses of unanalyzed data and to try to draw some generalization from them. The modern sciences, at least since Galileo, have been strikingly different. What they have sought to do was to construct refined experiments which ask, which try to answer specific questions that arise within a theoretical context as an approach to understanding the world." Noam Chomsky in an interview by Andor (2004).*

69

Chomsky uses what according to him could be an analogue in physics in order to show that collecting raw data is worth nothing: would it cross the mind of any physicist to film how leaves fall down from a tree for days or months if the goal is to understand how and why they turn while falling? Rather, recording data supposes:

70

1) to know what one is looking for, i.e. to design an experiment;

71

2) to have a working hypothesis;

72

3) to further analyze the data once they are acquired.

73

In Chomsky's mind, corpus linguistics is something where only step one of the production of knowledge is done (see section 6): real world items are collected, but they are not established as datum, there is no exemplum, no analysis, no expectation, no experiment design. Certainly Chomsky's reaction is also due to the term *corpus linguistics,* which may be interpreted as the linguistics of corpora. A less ambiguous label would be *corpus-based linguistics,* indicating that there is linguistic activity beyond the corpus.

Bibliographie

References followed by the mention WEB can be downloaded at http://www.unice.fr/scheer.

ANDOR, J. (2004). "The master and his performance: An Interview with Noam Chomsky." *Intercultural Pragmatics.* 1: 93-111.

BOTHA, Rudolf (1981). *The Conduct of Linguistic Inquiry. A Systematic Introduction to the Methodology of Generative Grammar*. The Hague, Paris, New York: Mouton.

BROMBERGER, Sylvain & HALLE, Morris (1989). "Why Phonology Is Different". *Linguistic Inquiry.* 20: 51-70.

BYBEE, Joan (2005). "La liaison: effets de fréquence et constructions". *Langages*. 125: 24-37.

COLLINI, Stefan (2012). *What Are Universities For?* New York: Penguin.

DURAND, Jacques (2009). "On the scope of linguistics: data, intuitions, corpora". *Corpus and Variation in Linguistic Description and Language Education*. Y. Kawaguchi, M. Minegishi & Durand J. (eds) 25-52. Amsterdam: Benjamins.

ENCREVÉ, Pierre & SCHEER, Tobias (2005). "L'association n'est pas automatique". 7e colloque annuel du GDR 1954 Phonologie, Aix-en-Provence 2-4 June. WEB.

FEYERABEND, Paul (1976). *Wider den Methodenzwang*. Frankfurt am Main 1986: Suhrkamp.

GREEN, Anthony D. (2007). *Phonology Limited*. Potsdam: Universitätsverlag Potsdam.

HALLE, Morris (2005). "Palatalization/velar softening: what it is and what it tells us about the nature of language". *Linguistic Inquiry*. 36: 23-41.

HARRIS, Zellig (1951). *Methods in Structural Linguistics*. (Edition 1960 entitled *Structural Linguistics*.) Chicago & London: University of Chicago Press.

HATHOUT, Nabil, MONTERMINI, Fabio & TANGUY, Ludovic (2008). "Extensive data for morphology: using the World Wide Web". *Journal of French Language Studies.* 18: 67-85.

HATHOUT, Nabil, NAMER, Fiammetta, PLÉNAT Marc & TANGUY, Ludovic (2009). *La collecte et l'utilisation des données en morphologie. Aperçus de morphologie du français.*Bernard Fradin, Françoise Kerleroux & Marc Plénat (eds). 267-287. Saint-Denis: PUV.

HUNDT, Marianne, NESSELHAUF, Nadja & BIEWER, Carolin (eds). (2007). *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

JOOS, Martin (1942). "A phonological dilemma in Canadian English". *Language*. 18: 141-144.

KAYE, Jonathan 1990. "What ever happened to dialect B?" *Grammar in Progress: GLOW Essays for Henk van Riemsdijk*. Joan Mascaró & Marina Nespor (eds). 259-263. Dordrecht: Foris.

KIRCHNER, Robert (1998). An effort-based approach to consonant lenition. Ph.D dissertation, University of California at Los Angeles.

LAKS, Bernard (2008). "Pour une phonologie de corpus". *Journal of French Language Studies.* 18: 3-32.

LAKS, Bernard Ms (2011). Pourquoi y a-t-il de la variation plutôt que rien ?

LAKS, Bernard (ed.) (2011). "Phonologie du Français Contemporain".*Langue Française.* 169.

LAKS, Bernard & CALDERONE, Basilio Ms (2012). French liaison and the lexical repository.

SAMPSON, G. (1978). "Linguistic universals as evidence for empiricism". *Journal of Linguistics*. 14:

183-206

SCHEER, Tobias (2004). "En quoi la phonologie est vraiment différente?" *Corpus*. 3: 5-84. WEB.

TRUBETZKOY, Nikolai Sergeyevich (1939). *Grundzüge der Phonologie*. 6th edition 1977, Göttingen: Vandenhoeck & Ruprecht.

Notes

[1] See Collini (2012) on the idea that science, and especially the humanities, need to be immediately useful to society.

[2] Note that serendipity, which has produced a number of scientific discoveries, does not undermine this point. Louis Pasteur put it this way: "luck favours the prepared mind" ("*Dans les champs de l'observation, le hasard ne favorise que les esprits préparés*").

[3] Cases in point are Hathout *et al.* (2008) and Hathout *et al.* (2009); Hundt *et al.* (2007) provides an overview of the landscape. Of course, identifying material on Google and then testing it with native speakers is a perfectly regular strategy of investigation, and there is no objection. It is only when statistics are directly made on Google data that there is no way to control for the caveats. A standard response is that caveat-created noise will lean out statistically and may be detected by this means. Or that the volume of this noise is so small that it won't have any significant impact on the result. I have never seen a case where these assertions are checked against the data, the reason being that 1) separating noise from non-noise statistically is not an easy thing to do and 2) even if this were done, the result could not be compared to the real amount of noise, which is unknown and cannot be calculated.

[4] Except for economics and literature, but the latter is for people who *produce* literature, not for academics who write *about* it.

[5] Exactly the same logic is applied in higher education: French government wants higher "success rates" (ratio between students inscribed and students who are granted a degree), and State funding of Universities is partly correlated to this "indicator of success" (benchmarking imported from capitalism). Universities are thus under pressure (an economic "incentive") to distribute more degrees. This can be achieved either by trying to provide better instruction and thereby modifying the content of the students' minds, or simply by lowering the standards: students need a degree, they pay for it, so give it to them no matter what their skills and competences. Guess which way Universities go…

[6] Though it is of course true that there are also lexically stored sentences, and that their number is subject to debate. See e.g. Bybee (2005) on this issue.

Pour citer cet article

Tobias Scheer (2013). "The Corpus: A Tool among Others". *CORELA - Numéros thématiques | Statut et utilisation des corpus en linguistique*.
[En ligne] Publié en ligne le 25 novembre 2013.
URL : http://corela.edel.univ-poitiers.fr/index.php?id=2952
Consulté le 11/12/2013.

A propos des auteurs

Tobias Scheer
University of Nice - Sophia Antipolis, CNRS 7320
scheer@unice.fr