

On the scope of linguistics: data, intuitions, corpora

Jacques DURAND

This the prefinal version of an article which appeared in Y. Kawaguchi, M. Minegishi & J. Durand (eds.) *Corpus Analysis and Variation in Linguistics*. Amsterdam: John Benjamins. pp. 25-52. Please refer to the published version for citations.

ABSTRACT

In this paper, I try to establish why corpora play a significant role within modern linguistics and yet occupy a controversial place within the field. After some preliminary remarks (section 1), I look in section 2 at the place of data within Chomskyan linguistics. Thereafter I proceed in section 3 to a partial criticism of the Chomskyan position and examine the light that corpora can throw on various phenomena. I select two areas of investigation: suffixation in *-able* in French morphology and French liaison. In section 4, I show how major changes have affected the field in relation to data and mental representations. The emergence of usage-based grammars, connectionism and other models force us to rethink our attitude, even if we do not have to share all the theoretical presuppositions of such approaches. Recent work by Chomsky and others within evolutionary theory paradoxically reinforce the need to look at data again. In the last section (§5), I try to suggest why intuition is likely to remain an indispensable tool for theory-construction in linguistics. But this should be no excuse for not strengthening our data-bases as bad or insufficient data rarely lead to good theories.

INFORMATION

Name : Jacques DURAND

Professor, Département des Etudes du Monde Anglophone and Director, Cognition, Langues, Langage, Ergonomie, CNRS & Université de Toulouse II. Senior Member of the Institut Universitaire de France.

Recent contributions:

“Mapping French Pronunciation. The PFC project.” In J.-P. Montreuil & C. Nishida (eds).(2006). *New Perspectives on Romance Linguistics*. Vol. 2 : *Phonetics, Phonology and Dialectology*. Selected Papers from the 35th Linguistic Symposium on Romance Languages (LSRL), Austin, Texas, February 2005. Amsterdam : John Benjamins. 2006. 65-82

“Réflexions sur l’analyse des énoncés en contexte. Plaidoyer pour une véritable ouverture théorique.” *Anglophonia (Sigma)* 22 : 21-54, 2007.

With Chantal Lyche, “French liaison in the light of corpus data”. *Journal of French Language Studies*. 18/1: 33-66, 2008.

On the scope of linguistics: data, intuitions, corpora

Jacques DURAND

What is the relationship of linguists's work to their data? Are the data merely entries in their diary that provide some understanding of how they got their eventual insights, or are they the stuff on which the theory stands or falls? This is the question an exegesis of Harris's work must answer. The rationalist today understands the linguist's data as being entries in the linguist's diary, of interest only to historians of linguistics; the real subject is the theory and the models. The empiricist sees the data which is collected by the larger community of linguistic researchers as an integral part of the work of the scientific community; no data, no science. It's not that it's simply unlikely the scientist will stumble upon the right theory without looking carefully at the data; there is no right theory to speak of except insofar as theory is united with data. (John Goldsmith, 2005:724)

1. Preliminary remarks

I would like to take as a starting-point the following observation: corpora play a significant role within modern linguistics and yet their precise place is still a matter of controversy.¹ That corpora occupy a special place in our field seems undeniable given the number of conferences, articles, monographs, book chapters and journals devoted to this very topic. Indeed, there are even researchers who define their work as framed within 'corpus linguistics'. On the other hand, there are still specialists (mainly but not solely within the Chomskyan tradition) who think that recourse to corpora is not *the* correct way of addressing the fundamental issues of linguistics. For instance, in a widely quoted paper entitled "Grammar is grammar and usage is usage", Newmeyer (2003) attempted to rebuff what he calls "anti-Saussurean usage-based models" by "provid[ing] evidence in support of the idea that the mental grammar contributes to language use, but that usage, frequency, and so on are not represented in the grammar itself." In so doing, he restated many of the tenets of Chomskyan linguistics asserting for example that "the mental grammar is only one of many systems that drive usage, since grammars are not actually well designed to meet language users' 'needs'." (2003:682).

¹ This paper reflects a long-standing collaboration with Bernard Laks and Chantal Lyche, including joint presentations we have made on this subject and related ones. They are obviously to be thanked for a number of observations and remarks and are in no way responsible for the weaknesses of this paper. Sincere thanks to John Anderson, Philip Carr, Sylvain Detey, Julien Eychenne, John Goldsmith, Nabil Hathout, Yuji Kawaguchi, Anne Przewozny and Gabor Turcsan. My apologies for not taking everything they told me into account.

While the title of this paper is extremely ambitious, my aim will be more limited. As a practitioner of linguistics with corpora who has a background in Chomskyan linguistics, it seems to me important to be able to stand back and reflect on one's practice and how it relates to the big issue of linguistic theorizing and the role of data, intuitions and corpora. My approach will be partially historical but I hope the reader will forgive me for taking here and there a number of shortcuts.

2. The Chomskyan turn

A good way of understanding the methodological and epistemological opposition to corpora shared by a number of modern linguists is to project ourselves back in time and summarize some of the tenets of Chomskyan linguistics. In *Language and Mind* (1968:2), Chomsky describes his "own feeling of uneasiness as a student at the fact that, so it seemed, the basic problems of the field were solved, and that what remained was to sharpen and improve techniques of linguistic analysis that were reasonably well understood and apply them to a wide range of materials." He even recalls "being told by a distinguished anthropological linguist, in 1953, that he had no intention of working through a vast collection of materials that he had assembled because in a few years it would be possible to program a computer to construct a grammar from a large corpus of data by the use of techniques that were already well formalized." The target of Chomsky's criticism is what he referred to as "structural linguistics" but, as he did point out in *Language and Mind* (1968:19), "structural linguistics" is too sweeping a characterization of a wide range of attitudes and approaches prevalent when he was a student. It would definitely have been more correct to speak of the post-Bloomfieldians. However that may be, it can hardly be denied that very many linguistic practitioners in the United States advocated a bottom-up approach to the discovery of linguistic structure. A good example is the following citation from Harris (1954): "The distributional investigations sketched above are carried out by recording utterances (as stretches of changing sound) and comparing them for partial similarities. We do not ask a speaker whether his language contains certain elements or whether they have certain dependences or substitutabilities. Even though his 'speaking habits' [...] yield regular utterances, they are not sufficiently close to all the distributional details, nor is the speaker sufficiently aware of them. Hence we cannot directly investigate the rules of the "language" via some system of habits or some neurological machine that generates all the utterances of the language. We have to investigate some actual corpus of utterances and derive therefrom such regularities as would have generated these utterances - and would presumably generate other utterances of the language than the ones in our corpus. Statements about distribution are always made on the basis of a corpus of occurring utterances; one hopes that these statements will also apply to other utterances which may occur naturally." (1954[1964]:47)

While acknowledging the fact that 'structural linguistics' allowed us to make significant progress in various areas (the scope of information available to us, the reliability of data, the possibility of studying linguistic relations at an abstract level, the technical precision), Chomsky saw this type of approach as fundamentally misguided: "if we ever are to understand how language is used or acquired, then we must abstract

for separate and independent study a cognitive system, a system of knowledge and belief, that develops in early childhood and that interacts with many other factors to determine the kinds of behaviour that we observe; to introduce a technical term, we must isolate and study the system of linguistic competence that underlies behavior but that is not realized in any direct or simple way in behavior. And that system of behavior is qualitatively different from anything that can be described in terms of the taxonomic methods of structural linguistics, the concepts of S-R [Stimulus-Response, JD] psychology, or the notions developed within the mathematical theory of communication or the theory of simple automata.” (1968:4). But one could, of course, agree with Chomsky that the object of linguistics is the study of cognitive systems internalized by speakers-hearers and disagree on the idea that it cannot be accounted for by general mechanisms available to humans in other domains. As we know, if Chomsky took the view that the language faculty is special, it is among other things because we have a capacity to make infinite use of finite means, a capacity which seems to be specific to humans. As he puts it (2000:3-4): “Human language is based on an elementary property that also seems to be biologically isolated: the property of discrete infinity, which is exhibited in its purest form by the natural numbers 1, 2, 3, [...] Children do not learn this property; unless the mind already possesses the basic principles, no amount of evidence could provide them. Similarly, no child has to learn that there are three and four word sentences, but no three-and-a-half word sentences, and that they go on forever; it is always possible to construct a more complex one, with a definite form and meaning. Such knowledge must come to us “from the original hand of nature,” in David Hume’s (1784/1975: 108, Section 85) phrase, as part of our biological endowment.”

If Chomsky talks of ‘discrete infinity’, it is because one central assumption he makes is that linguistic systems are composed of discrete units characterizable in an algebraic or logico-mathematical way rather than in quantificational terms. In this area, he is definitely the inheritor of a long tradition of work within Saussurean, Sapirian and Bloomfieldian linguistics where numbers were rejected in favour of patterned structure and underlying units indirectly related to surface realizations. As a single example, consider the following statement by Whorf (1956:230-231): “Linguistics is also an experimental science [...] In place of apparatus, linguistics uses and develops TECHNIQUES. Measuring, weighing, and pointer reading devices are seldom needed in linguistics, for quantity and number play little part in the realm of pattern, where there are no variables but, instead, abrupt alternations from one configuration to another.” Utterances might be characterizable in a quantitative manner but our goal as linguists is entirely different as stressed by Chomsky in *Aspects of the Theory of Syntax*: “Linguistic theory is concerned primarily with an ideal speaker-hearer, in a completely homogeneous community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance.” (1965: 3) With a few terminological differences and slight differences in emphasis, this is what is reasserted in more recent work. To paraphrase Chomsky (1995: 18-19), the goal of linguistic theory is the characterization of the language faculty. The latter has an initial state which is genetically determined. During acquisition, the language faculty passes through a series of states until it

reaches, in relatively short time, a stable steady state which will undergo little subsequent change and which appears to be uniform for the species. The theory of the state attained is called its grammar and the theory of the initial state is Universal Grammar (UG). When we say that Jones has the language L, what we mean is that Jones's faculty is in the state L. To distinguish this definition of language from other possible definitions, it will be referred to as "I-language" where "I" is intended to suggest: "internal", "individual" and "intensional". I-language must be sharply separated from E-language, the external manifestation of I-language in the form of utterances, texts, sets of sentences and social conventions.

If I-language involves properties which are biologically isolated, *sui generis*, then there is indeed no reason to expect that utterances taken as physical objects will be necessarily revealing about the underlying functional structures from which they proceed. While part of our biological make-up, the language faculty involves mental representations which are structured in a symbolic way, like propositions. Cognitively, the use of language has to rely on a calculus, a computation on these representations. Given that Chomsky is committed to a mediationalist approach in Huck and Goldsmith's (1996) sense - i.e. linguistic theory must define the link-up between sound and meaning - and that phonology and semantics are seen as interface systems, the calculus on mental representations is of a syntactico-logical type. Syntax is the only truly generative component underlying our ability to make infinite use of finite means. The representations, while causally involved in speech production and perception, do not have to conform to existing models of biological structure. Nor can they be modelled by quantity and numbers. As stressed by Smith, one of Chomsky's most ardent followers: "There are areas in the language sciences where quantificational techniques have a role to play. An obvious example is the setting of the norms for developmental progression of the sort exemplified by Irwin and Wong's work.² Here too it is necessary to take cognizance of the fact that quantification needs to be used insightfully: using group data rather than individual data in such a case obscures rather than enlightens. More importantly, the quantification is parasitic on the results of linguistic theory, it is not determinative of them. In general the numbers game is irrelevant to the linguist both because the nature of his hypotheses do not usually lend themselves to quantitative testing in virtue of being a mental/psychological rather than a physical discipline and more soberingly because we are still a long way from the explicitness of the physicists for whom quantification is a *sine qua non*. Lem describes statistics as 'the rationalist's substitute for demonology': we need to exorcise our demons." (2004:185).

The above remarks may explain why from early on in his work, Chomsky has expressed serious reservations about strengthening the data by operational tests as, for example, in the famous quote from *Aspects of the Theory of Syntax*: "In any event, at a given stage of investigation, one whose concern is for insight and understanding (rather than objectivity as a goal in itself) must ask whether or to what extent a wider range or more exact description of phenomena is relevant to solving the problems at hand. In

² Smith is referring to work on language acquisition reported in Irwin and Wong (1982).

linguistics, it seems to me that sharpening of the data by more objective tests is a matter of small importance for the problems at hand. One who disagrees with this estimate of the present situation in linguistics can justify his belief in the current importance of more objective operational tests by showing how they can lead to new and deeper understanding of linguistic structure” (1965:20-21). Instead of sharpening the data via corpora, for instance, as advocated in much recent work, “the speaker-hearer’s intuition is the ultimate standard that determines the accuracy of any proposed grammar” (ibid: 21). The belief in the centrality and the reliability of intuition is taken for granted by many linguists working in the Chomskyan generative tradition as demonstrated in the following recent quote from Pinker (2007:33-34): “Some people raise an eyebrow at linguists’ practice of treating their own sentence judgments as objective empirical data. The danger is that a linguist’s pet theory could unconsciously warp his or her judgments, but in practice linguistic judgments can go a long way. One of the perquisites of research on basic cognitive processes is that you always have easy access to a specimen of the species you study, namely yourself. When I was a student in a perception lab I asked my advisor when we would stop generating tones to listen to and start doing the research. He corrected me. Listening to tones *was* research, as far as he was concerned, since he was confident that if a sequence sounded a certain way to him, it would sound that way to every other normal member of the species. [...] As a sanity check (and to satisfy journal referees) we would eventually pay students to listen to the sounds and press buttons according to what they heard, but the results always ratified what we could hear with our own ears. I’ve followed the same strategy in psycholinguistics, and in dozens of studies I’ve found that the average ratings from volunteers have always lined up with the original subjective judgments of the linguists.”

Of course, many other things could be said about data, intuitions and theory construction in Chomskyan generative grammar. In particular, the commitment to a Popperian epistemology by the transformationalists combined with the idea that their approach represented a new paradigm à la Kuhn, would require further discussion (Huck and Goldsmith 1996). I nevertheless hope that enough has been said to set the scene for an examination of alternative approaches in the next two sections.

3. Epistemological dissensions and the rise of corpus work

From the very beginning of Chomskyan generative grammar, there have been dissenting voices concerning the nature of data. Most prominent among the critics have been the sociolinguists who have stressed that the idealization taken for granted by the generativists was too strong and that variation was inherent in language-use. Recourse to intuition might ultimately be unavoidable – an issue I return to at several points – but it has become quite clear, contrary to Pinker’s unshakeable confidence, that the judgments made by native speakers are often unreliable, not always shared intersubjectively and not stable over time.

On the question of intuition, I should make my position clear. It seems to me that one of the crucial properties of human language is its capacity for referring to itself, a property I will call reflexivity after Lyons (1977:5-13, 1995:6-11). All human

languages allow their speakers to use language self-referentially.³ The terminology used might not be a technical one (e.g. *Am I speaking too fast?*) but this ability underlies communication between people of different cultures when they first meet and, if their intention is indeed to establish cultural bonds, they will be able to establish correspondences between chunks of linguistic structure and ultimately master other languages than their native one or teach their language to a willing learner. In other words, even in an unwritten language, the speakers will be able to correct learners, formulate paraphrases, differentiate chunks of language or react to what they will judge as nonsense. This ability without doubt underlies the linguist's metalanguage. It does not follow however that judgments of grammaticality are as reliable as we might wish. Smith (2004:7) takes examples such as:

(1) *John speaks fluently English

vs.

(2) John speaks English fluently

to illustrate the difference between an acceptable and an unacceptable sentence, which grammar has to account for and I agree. But while an **untutored** native speaker presented with these two structures might correct (1) if used by a foreigner they knew well, it is doubtful that they would make sense of questions relating to co-referential structures such as (3a-c):

- (3)
- a. The knowledge that John_i might fail bothered him_j
 - b. The possibility that John_i might fail bothered him_i
 - c. The realization that John_i might fail bothered him_{i/j}

Smith (2004:67), who presents these examples, says that in (3a) *him* cannot (normally) be construed as referring to John, whereas in (3b) *him* typically can be taken to refer to John (although, as before, it could have been used to refer to some unnamed person in the context). In (3c) both possibilities are open. Smith notes that these judgments are subtle and somewhat variable from person to person. My aim here is not to question these intuitions as such but to stress that they are clearly linguists' intuitions. Our own experience in getting students to begin to reason about what is involved in coreference should suffice to convince us of this. But, if these are linguists' judgments, then we can surely see the potential pitfalls. How do we avoid the invention of spurious examples designed to prop up a particular account? Can we be satisfied when disagreements are seen as the result of different dialects (more precisely idiolects)? We are told by specialists defending the strong position that idealization is inevitable: "When Galileo devised the law of uniform acceleration for falling bodies, either by dropping weights from the leaning tower of Pisa or rolling balls down an inclined plane, he ignored the effects of wind resistance and friction" (Smith 2004:10). But one of the characteristics of the so-called hard sciences has precisely been their ability to reintegrate elements that were often discarded in a first round of investigation. Likewise, one would expect linguistics to be able to compare some of the structures which are asserted to be possible or impossible with attested utterances or to test interpretations psycholinguistically. To my knowledge, this has not been done in a systematic way and would be seen by some as mere diversion from real theoretical work.

³ Also called 'reflectiveness', which is one of the design-features listed in Hockett and Altman (1968). Many philosophers, logicians and linguists have drawn attention to this special property of language as a semiotic system. See e.g. Jakobson (1960). In philosophy and logic, reflectiveness is part and parcel of discussions of 'use' vs. 'mention': e.g. Boston is a large city vs. 'Boston' has six letters.

In the light of what has just been asserted, my aim now will be to take two areas from French linguistics and show that sharpening the data is not a luxury but an essential component of theory construction.

3.1 A test-case from morphology

The first example I wish to discuss is drawn from morphology. It is based on work done within my research centre (CLLE-ERSS University of Toulouse and CNRS) and, in particular, I will report on a study of the French suffix *-able* by Hathout, Plénat and Tanguy (2003), which has recently been summarized along with other examples by Hathout, Montermini and Tanguy (2008). As stressed by these authors, the description of word-formation, particularly derivational morphology, needs corpora. Traditionally, linguists have supplemented their own intuitions by recourse to dictionaries and hand-collected examples. But the data has often been sparse. Yet, by its very nature, morphology is perhaps the area most likely to benefit from large amounts of electronic data, due to the ease with which word forms can be gathered and processed by computer programs. In recent years, as far as French is concerned, new sources of information have become available such as the TLFi electronic dictionary (*Trésor de la langue française informatisé*) or text databases such as Frantext. In a matter of seconds, we can obtain extensive lists of words ending in *-able* (taken here as a paradigm example) whereas the same search would have taken months of work not so long ago. In addition to these electronic data-bases, these authors argue that the World Wide Web is an indispensable source of information.

Of course, it is often objected that collections from large text bases, and even more the World Wide Web (hereafter ‘the web’), yield data which is insufficiently controlled and increased noise. However, as far as morphology is concerned, it is arguable that one can exploit raw data which might be unsuitable for other studies. Only large corpora may allow us to witness unusual cases of word-formation or confirm whether some phenomena are rare or totally unattested. For instance, using the web, it can be shown that the prefix *anti-* can be prefixed to simple monomorphemic adjectives, such as in *anti-triste* (‘anti-sad’) or *anti-obèse* (‘anti-obese’), and even to adjectives following the *V-able* scheme, such as *anti-inflammable* (‘non-flammable’). The first case is considered as dubious in Durand (1982) on the basis of intuition and standard dictionaries; the second is presented as theoretically impossible in Fradin (1997:100-101).

Of course, using the web has disadvantages. The authors I rely on discuss these at length. They argue that, with specific search engines and extensive filtering for unreliable data, results close to what is made available in dictionaries is possible, but of course on an unprecedented large scale. For instance, we have to exclude direct transfers from another language (most of the time the author’s mother tongue or the results of machine translation programs!), words coined for stylistic reasons (rhyme, pun, etc.); regionalisms or archaic words; plainly incomprehensible contexts, either from low-quality writing or technical jargon. The tool they use, Webaffix (Hathout and Tanguy 2002), is able to check for the correct target language, typos or bad word divisions. It can also perform more restrictive selections. For instance, only derived words that co-occur with their base forms in a web page are retained. This is an efficient criterion for establishing morphological links between two lexemes. Thus, *copolymérisable* (‘copolymerizable’) will only be considered as a legitimate *-able* adjective form if the inferrable base verb *copolymériser* (in any of its inflected forms) appears in its vicinity. Moreover, the web need not be solely used in a static way. One of the techniques used by my colleagues has involved generating new forms that would be predicted as acceptable or unacceptable on theoretical grounds and

searching the web for their possible presence. In using corpora one does not have to assume that the programs we devise do everything mechanically and we would therefore be dispensed from carefully examining the results of our searches. With these caveats in mind, we can now turn our full attention to the *-able* suffix.

As in English, adjectives derived with *-able* in French have often been regarded as essentially de-verbal adjectives with a passive meaning. In other words, the noun they modify is analysed as corresponding to the direct object or the patient of the base verb, depending on whether the relation is viewed from the point of view of syntax or from that of semantic relations (or deep cases or thematic roles in other terminologies). For instance, in a classical work such as Nyrop (1936 : 84-85), it is asserted that “in the modern language, the *-able* suffix usually has a passive meaning (*désirable*, “qui mérite d’être désiré” [i.e. desirable, ‘which deserves to be desired’, JD], rarely an active sense (*secourable*, “qui secourt” [i.e. helpful, ‘who helps’, JD]” and that, as far as the new formations are concerned, “the passive sense is the only normal one” [my translation, JD]. In other words, what is being claimed is that the basic mechanism for *-able* derivation should be summarised as in (4):

(4) –ABLE derivation : some traditional assumptions

a. *Morpho-phonology*: $[X] \rightarrow [X + /abl\partial /]^4$

b. *Morphosyntax*: $[_V X_V] \rightarrow [_A [_V X] + -ABLE_A]$

Condition : X is a transitive verb and the direct object headnoun it is subcategorized for is the subject selected by the derived adjective in a predicative structure or the noun it can modify within an NP. E.g. we can say *ces choux-fleurs sont mangeables* (these cauliflowers are eatable) or speak of *choux-fleurs mangeables* (eatable cauliflowers) because *manger* (eat) can take *choux-fleurs* as a complement in a sentence such as *Les Français mangent des choux-fleurs* (The French eat cauliflowers).

c. *Semantics* $[_{PRED} X]$ --> “which can be X-ed” (mangeable = “qui peut être mangé”, eatable = “which can be eaten”).

It has however always been known that other examples of *-able* derivation existed. Beside the active sense found in *secourable* (helpful) cited above, one can find cases where the subject of an *-able* adjective in a predicative structure corresponds to a circumstantial (*une piste skiable*, “a slope one can ski on”) or different modal values (including a lack of clear modal interpretation). Thus if somebody is said to be *adorable* in French, it may mean as in English that they are worthy of worship but the most usual sense is “delightful, charming”. The modality does not seem crucial to the sense here as is indeed the case for words like *raisonnable* (reasonable) or *équitable* (equitable). All other examples than the paradigmatic one summarised in (4) have usually been seen as marginal and not requiring explanation.

Using the Webaffix tool, Hathout, Plénat and Tanguy (2003) have constituted a database of some 5000 adjectives in *-able*. Note that complete searches from electronic dictionaries had yielded less than three times this figure : 1641 *-able* adjectives are described in the TLFi and the Robert Electronique. The collected data does confirm that most of the *-able* adjectives have a passive meaning. But the noun they modify can also represent a variety of other participants in the process. The plasticity of *-able* derivation can be illustrated by looking at the possible nouns that a derived adjective such as *pêchable* ‘fishable’ (which does not appear in the TLFi) can modify. We will not be surprised to discover that prominent among the things that are

⁴ Leaving aside variants such as /ibl\partial / or /ybl\partial /.

said to be *pêchable* are fish and various types of seafood. However, places can also be qualified as *pêchable*: i.e. bodies of water (rivers, ponds, streams, etc.) and fishing spots like riverbanks, bridges, dams, etc. Less evidently, depending on whether the fishing season is open or not, whether the weather is pleasant or not, days, seasons and atmospheric conditions can also be characterized as *pêchable* or *impêchable* ('unfishable'). Even more surprisingly, my colleagues have also found contexts where *impêchable* modifies fishing tackle (flies or nylon fishing lines, for instance). Finally, not only the participants in the process, but their properties too can be characterised as *pêchable* or not. Thus they have collected examples where fish size is said to be *pêchable* as in the following extract from the web :

(5) L'ouverture du gisement à la pêche semble incompatible avec sa gestion durable. Compte tenu de la raréfaction des coques de **taille pêchable**. (lit. The opening of the area to fishing seems incompatible with its exploitation over time. Taking into account the attrition of numbers of cockles of **fishable size**.)

As pointed out, in a tongue-in-cheek way, by Hathout, Montermini and Tanguy (2008:77): "Actually, the fishermen seem to be the only participants that cannot be said to be *pêchable*!"

But there is more to this study. It shows that when verbs are not available nominal bases can do as well. So, for example, one can speak of

(6) terrain piscinable, garage boxable, statue muséable, ministre matignonnable

A 'terrain piscinable' is a "piece of land large enough to accommodate a swimming pool"; a 'garage boxable' is "a parking space that can be transformed into a lock-up garage"; a 'statue muséable' is a statue worthy of being exhibited in a museum", a 'ministre matignonnable' is a minister worthy of being appointed as a Prime Minister and reside at Matignon. What Hathout, Plénat and Tanguy conclude is that categorial constraints on the *-able* derivation have a semantic origin: *-able* derivatives usually select verbs as bases because they denote processes, but when a process does not have a specific corresponding verb, a nominal base will do quite nicely. As they put it:

(7) *-ABLE derivation*: « En fin de compte, l'observation des données nouvelles suggère que peut être dit Xable tout élément de la situation qui intervient dans le procès X ou presque, pourvu du moins que cet élément soit conçu comme se prêtant à la survenue de ce processus. » (2003 : 51) (Ultimately, the examination of new data suggests that Xable can be applied to any element of a situation which is totally or partially involved in a process X, provided that the element in question can be seen as suitable for the occurrence of the process. [my translation, JD].)

These authors themselves would agree that the book is not closed by their detailed study of *-able* but they would insist that an appropriate examination of morphological structure cannot be achieved without extensive corpora. The web does contain a variety of language usages, many of which have not been previously taken into account in wide spectrum linguistics studies (except for some popular magazines or paperback novels). This "new" kind of data should not however be seen as a weakness but as a strength. Generative linguists interested in word-structure have appropriately stressed the open nature of the lexicon. The web gives a window on spontaneous, untutored word-coinage arguably more revealing than standard dictionaries which often only reflect normative usage.

3.2 A test-case from phonology

My second example will be taken from phonology and will deal with some aspects of 'liaison', which will be discussed here in a simplified way. French liaison is a sandhi

phenomenon which involves the pronunciation of a final consonant, which is mute in certain contexts, at the boundary between two words (referred to as W1 and W2 here). To take a hackneyed example, the word *petit* in standard French alternates between two main variants [pti] and [ptit] as shown in [4] (leaving aside the possibility of a schwa between the initial /p/ and the /t/):

(8) French liaison

- a. petit écrou (small nut) [ptitekru]
- b. petit cadeau (small present) [ptikado]
- c. il est petit (he is small) [pti]
- d. elle est petite (she is small) [ptit]
- e. petitesse (smallness) [ptites]

An assumption made in classical generative phonology (Schane 1968, Dell 1973/1985) is that the liaison consonant (/t/ in (8)) is part of the underlying phonological representation of the first word (W1) and maintained only if the next word (W2) is vowel-initial. (In (8d) it is hypothesized that there is an underlying schwa corresponding to the written word-final <e>.) This final /t/ is deleted if W1 is utterance-final or if W2 begins with a consonant. Most classical treatments of liaison also assume (a) that the underlying consonant is linked forward, (b) that there is a strong connection with syntax which combines words in a compositional way (in a Fregean sense) and creates the conditions for liaison. Liaison is traditionally said to function in two modes: obligatory (e.g. (8a)) and facultative as in e.g. plural noun + adjective as in e.g. *draps* ([z]) *anglais* (English sheets).

There is a vast and often conflicting literature. Thus the liaison consonant has been treated as word-final, extra-linear, epenthetic, or even as a prefix of the W2 word. The treatments offered have illustrated all theoretical models: rules, principles and parameters, constraints, usage-based schemas, and so forth. The link with syntax has been achieved through boundaries, direct interpretation or via a prosodic coding. I will not dwell on these matters here. The thrust of my discussion will be the observation that many treatments have simply been based on on Fouché (1959), a normative reference book aimed primarily at foreign students and teachers of French, listing over 30 pages of impossible instances of liaison, and claiming to describe “la prononciation soignée [...] des Parisiens cultivés nés vers la fin du XIXe siècle ou plus tard” (i.e. the careful pronunciation of cultivated Parisians born around the end of the 19th century or later). The empirical basis of Fouché’s observations are, however, open to serious questioning as underlined by Morin (2000) and Laks (2002) who demonstrate that so-called standard French (SF) is a hydra invested with multiple definitions.

A simple example of the impact of the normative tradition on contemporary analyses is that of prenominal adjectives. The combination ‘Adj + N’ was assumed by Fouché and many followers (e.g. Delattre 1951 and Léon 1992) to trigger obligatory liaison. Indeed, many specialists ever since Selkirk (1972) have simply discussed *sot ami* as a prime example of liaison and examples like *sot ami* (lit. foolish friend) have been given at nauseam in the technical literature.

Within the PFC project (*Phonologie du français contemporain: usages, variétés et structure*) we have thoroughly investigated the question of liaison. Our surveys involve four activities: the reading of a word-list, that of a passage and two conversations: formal and informal (Durand, Laks and Lyche 2002, Durand and Lyche

2003, Durand 2006). The passage and the two conversations are transcribed orthographically and manually coded for liaison. In recent work (Durand and Lyche 2008) we have attempted to show how the quantitative data throws light on many questions which would simply remain unanswered if one put one's full trust in intuition.

The two conversations fare poorly in providing abundant data on prenominal adjectives for the simple reason that *petit* overwhelmingly represents this particular context and that other adjectives occur only sporadically in a liaison environment. We can nevertheless test the sequence 'Adjective + Noun' in the PFC text read by all our speakers through two expressions: *grand émoi*, *grand honneur*. We assumed that in a reading task, all 100 speakers would link the adjective to the following noun, which is not what we observe: six speakers do not make the liaison and two pronounce a [d] instead of the expected [t]. Four of the six cases concern *grand émoi*, suggesting that the lack of familiarity with the construction impacts on the absence of liaison. Interestingly, all speakers do not treat *grand émoi* and *grand honneur* in a parallel way. Some speakers had a liaison in *grand honneur* but not in *grand émoi*: this seems to correlate with frequency. *Grand honneur* is a familiar phrase where *grand émoi* is not to the same extent. Such observations bring grist to usage-based models rather than models based on the assumption that liaison is compositional (once again in the Fregean sense).

Data on prenominal adjectives present a particular interest due to the theoretical debate concerning their treatment. The current literature opposes a morphological approach (Steriade, 1999; Tranel, 1996, 1999, inter alia) to a phonological one (Féry, 2003). Steriade (1999) argues that masculine and feminine adjective allomorphs are listed in the lexicon, and that a hiatus situation is resolved by lexical conservatism 'a class of grammatical conditions [...] promoting the use of pre-existing familiar expressions or parts of properties of such expressions.' When hiatus occurs between an adjective and a noun, lexical conservatism requires that before inserting new segments that would solve the problem, one should look within the paradigm for possible solutions. Since the feminine allomorph of an adjective usually ends in a consonant, it implies that in a hiatus situation, the masculine allomorph will take the shape of the feminine allomorph.⁵ In the phonological approach, defended by Féry (2003), the proper ranking of syllabification constraints suffices to account for the liaison form of the adjective. Both analyses treat liaison as a means of avoiding hiatus, and both propose to explain the presence of a consonant in examples like *sot ami*, *sot aigle*. We will not dwell here on the numerous examples showing that NO HIATUS must be a low-ranked constraint in French (see Morin 2005), but will instead consider the data the analyses are based upon. Morin (1987) already pointed out the artificial character of *sot ami*, and a search through the entire PFC database should bring a few answers concerning pronominal adjectives. We note that in his elicitation work involving nasal vowels, Sampson (2001) was unable to trigger liaison for adjectives placed in prenominal position and tested without success *fin*, *hautain*, *lointain*, *malin*, *mignon*, *souverain*. He concludes that, outside the usual inventory (*un*, *mon*, *ton*, *son*,... *bon*, *plein*), 'the available evidence suggests that ZERO-liaison may already be established, or be well on the way to becoming established, as the default arrangement' (p.255).

Prenominal adjectives appear in large numbers in the base, but only rarely in a liaison environment, and when they do, the liaison is not categorical. The adjective *gros* will serve to illustrate this point: we record 139 occurrences of *gros* in the base, but

⁵ Steriade (1999) restricts the number of possible liaison consonants to a few, which explains why in a liaison context *grand* is not pronounced with a [d].

only 8 in a liaison context. In 6 instances, the adjective is in its plural form and liaison is realized (*gros [z]ouvrages*). In the other two instances, liaison is present as expected, in the common phrase *gros [z]oeuvre*, but absent in *gros // immeuble*. This particular example shows the strength of the plural marker for liaison, although we should show caution in drawing hasty conclusions. Disyllabic adjectives like *premiers* vary between liaison and no liaison; *grands* links to the following word; and *petits* is pronounced several times with a [t] instead of the expected [z], as in *beaucoup de p(e)tits [t]hotels*. Most interestingly, although so-called elementary adjectives (in the terminology of traditional grammar) do occur regularly as expected, in a prenominal position, they rarely do so in a liaison environment. In other words, speakers seem to talk without difficulty about *un gros type*, *un gros chien*, but not about *un gros homme*, *un gros âne*. The PFC base not only throws doubt on the categorical character of liaison in prenominal adjectives, but it suggests that speakers systematically avoid a situation where they will be compelled to make a decision concerning the presence or not of a liaison (Lyche, 2003).

I will not pursue the question of liaison here (see Durand and Lyche 2008 for an extensive treatment). One objection standardly formulated against corpus studies such as ours is that the observations we make do not represent anybody's grammar but heterogeneous data belonging to various systems. This is to some extent true but on the other hand is the solution to try and guess in one's head how one would say certain sequences? The problem here is that in so doing what the linguist does is to fall prey to normative assumptions partially derived from orthography and school teaching. Given that modern linguistics assumes the priority of the oral medium over the written medium and rejects prescriptivism, it is a rather sad conclusion that what one might simply be formalising artificial systems. It is also important to note that in the PFC approach, we avoid one shortcoming of using the web: information about the speakers is available and our observations can be as wide or as narrow as we wish (e.g. single speakers or all female speakers born before 1950 in a given survey point).

4. The return of distributionalism?

The observations made in section 3 do not as such provide a refutation of the use of intuition but, in so far as the conclusions drawn have any validity, they show that hypotheses cannot simply be matched against judgments made by linguists on their own usage. It is often said that corpora are the equivalent of the telescope in the history of astronomy. Many of the hypotheses concerning the nature of the universe were in place before telescopes were invented. But progressively the observations they made possible proved crucial in the validation or invalidation of various theories. However, it should also be observed that a number of linguists have taken a radical route concerning the nature of data in linguistics. Newmeyer (2003) argues there has been a convergence between two types of approach. On the one hand, it has repeatedly been observed that frequency effects were important in accounting for language structure, language acquisition and language change. On the other hand, new paradigms have become accepted such as connectionism which have been claimed to offer a much more plausible vision of brain processes than the symbolic model advocated within generative grammar and classical cognitivism.

The relevance of quantitative data had begun to emerge within sociolinguistics in the wake of Labov's seminal work (see e.g. Labov 1994, 2001 and the references therein). A concept had even been proposed of variable rules, i.e. generative rules which could be indexed for various parameters - ranging from phonetic to situational (see Chambers and Trudgill 1980: ch9 for an overview). This technique encountered a

great deal of antagonism from a generative point of view. Part of it just reflects methodological assumptions: since variable rules reflected performance data (and even worse group data!) they could simply not be assumed to throw light on individual speakers' competence systems. Moreover, it was argued that they made no sense psycholinguistically: if a rule applies 23 per cent of the time, do we need to keep a tally of all applications to make sure that our behaviour matches the constraints on the application of the rule? But note, in passing, that the critics did not apply the same stringent psycholinguistic requirements on the discrete formal units and the types of computational mechanisms they assumed (e.g. distinctive features, phonemes, morphemes, words, phrases, etc., on the one hand, and e.g. syntagmatic rules and transformations, on the other). The idea that the generalizations made by sociolinguists could be integrated to an abstract model of language competence and language performance, where linguistic and social knowledge would interact, and within which the theoretical probabilities might be matched against actual data was seen by most generative grammarians as uninteresting. Yet it will be recalled that the goal of linguistic theory, even for Chomsky, has also been to explain language use.

In recent years, usage-based models have swept large parts of the field of linguistics. One well-known proponent of this type of approach within what might be called 'core linguistics' is Bybee (2001, 2007) who defends strongly the idea of paying attention to 'tokens' and not only to 'types' and envisaging language as an emergent system which has the properties found in other complex systems in nature. "In complex systems, a small number of mechanisms operate in real time and with repetition lead to the emergence of what appears to be an organized structure, such as a sand dune. However, we know that a sand dune is not fixed in time and space but is ever altering and becoming. So we see that language is also always in a process of becoming – creating, losing, and re-creating structures that are never absolutely fixed, allowing for continued variation and change (Lindblom, MacNeilage, and Studdert-Kennedy 1984; Hopper 1987; Holland 1998). In such theories repetition of actions brings about the formation of structures; thus in language, too, we see that repetition is a necessary component of grammar formation (Haiman 1994). The reason frequency or repetition plays a role in grammar formation is that the mind is sensitive to repetition. This is a domain-general principle; that is, it does not apply just to language but to other cognitive domains as well." (2007: 8). Not surprisingly, this "new" trend has also affected other subfields. As Newmeyer puts it (2003:000), "I am sure that Christopher Manning is right when he writes that '[d]uring the last 15 years, there has been a sea change in natural language processing (NLP), with the majority of the field turning to the use of machine learning methods, particularly probabilistic models learned from richly annotated training data, rather than relying on hand-crafted grammar models' (Manning 2002b:441)". Such work has clear implications for psycholinguistics and various specialists now argue that unsupervised machine learning of grammar lends support for the view that the acquisition of the language faculty can be achieved through general machine learning methods on the basis of a minimal set of initial settings for possible linguistic categories and rules hypotheses (e.g. Lappin 2005, Lappin and Shieber 2007). If this is correct, it offers a rebuttal of the poverty-of-stimulus argument put forward by Chomsky to justify a form of UG including properties not attested elsewhere in the natural world.

Now, the various trends referred to in the previous paragraph do not necessarily provide a coherent picture of language structure. Nevertheless, they provide a strong challenge to the classical Chomskyan view of the language faculty and indeed much that was assumed by structuralists (given that Chomsky is much more the inheritor of

the structuralist tradition that he acknowledges).⁶ My purpose here is not to claim that these alternative views are correct but simply to remind ourselves of their strong presence in the field (as repeatedly emphasized by Newmeyer 2003, 2005). Three types of objection are often formulated against the more data-driven approaches just mentioned.

A first objection consists simply in asking: ‘Where are the results?’. This question and the implied negative answer seem to me extremely unfair. After fifty years or so of modern generative grammar, which subdomain can be considered as settled, even provisionally among specialists? To take an area in which I feel reasonably informed, consider syllable structure. For many, the syllable is as good a candidate as any for the status of phonological universal. Yet, we know that it was (i) present among the structuralists, then (ii) absent in the early generative phase (Chomsky and Halle 1968), (iii) it reappeared strongly in the eighties, (iv) to subside again in some frameworks (e.g. government phonology and its offshoots) and finally (v) made a moderate comeback within Optimality Theory via constraints such as Onset and *No coda. Moreover, there are languages such as Japanese, where the syllable is often argued not to be the most adequate concept for handling basic suprasegmental structure and where the mora is said to capture better generalizations. Finally, I note that even within the camp of syllable supporters, no two specialists agree fully on how it is structured: Does it have a flat structure with perhaps the syllabic selected as head? Does it have an onset and a rhyme, the latter divided into a nucleus and a coda? Does it exhaust all the phonemes a word is composed of or do we have to include extrasyllabic elements (e.g. the initial /s/ in a word like *strap*)? Is it structured in such a way as to reflect sonority structure (e.g. as in Dependency Phonology, Durand 1990) and so on and so forth.

A second objection made by Newmeyer (2003) is that the current sea-change towards usage implies an abandonment of the competence/performance distinction. This is probably true for a number of specialists who will welcome this (e.g. Laks 2008 if I correctly understand his position), but as clearly argued by Clark (2005), this assumption is unwarranted. It is possible to defend the idea that grammar incorporates probabilistic information (cf. the concept of ‘stochastic grammar’) and is a mental object (i.e. internalized in the brain of an individual). Thus Clark defends the idea that a realistic mental grammar must make room for the insights of the inherent variability tradition. As he puts it: “The inherent variability tradition includes the variable rule approach (see Labov 1969 and the references in Paolillo 2002), classification and regression tree analysis (Ernestus & Baayen 2003), analogical modeling of language (Skousen 1989), generalized linear models (see the references in Manning 2003), various versions of optimality theory (e.g. stochastic optimality theory (Boersma 1998, Clark 2004), partial ordering (Anttila 1997), floating constraints (Nagy & Reynolds 1997), extensions of head-driven phrase structure grammar (Bender 2001) and extensions of the principles and parameters framework (Yang 2004). A guiding assumption is that mental grammar accommodates and generates variation, and includes

⁶ On this see Anderson 2005 who argues that in fact Chomsky is the arch-structuralist. If Anderson’s argument is correct, then the title of this section and some of the arguments presented here need radical rethinking.

a quantitative, noncategorical and nondeterministic component (Weinreich et al. 1968, Bender, 2001)’’.

A third objection is that the new paradigm requires us to accept connectionism and that connectionism is not secure enough or developed enough to provide a foundation for all work in linguistics. But not all linguists defending stochastic grammar are necessarily committed to connectionism (in fact Lappin and Shieber 2007 explicitly separate the two issues). In any case, connectionism models come in many different varieties. The most extreme varieties have often been referred to as ‘eliminationist connectionist models’. Among the tenets of these approaches is the idea that the basic analytic concepts of generative and most other linguistic theories are simple artefacts which can and should be eliminated in some sense and that the dynamic numerical computations embodied in neural nets should replace all symbolic approaches. There is however a problem in adopting such a strong stance. Connectionism is neither quantitatively nor qualitatively fully realistic from a neurobiological point of view. Firstly, the current techniques of dynamic calculations and the size, complexity and architecture of the networks effectively manipulable are vastly inferior to the capacities of the human brain. Secondly, our knowledge of neural architectures, synaptic processes and more generally of the neurophysiology and neurochemistry of the brain is still in its infancy, despite the very real progress which has been achieved. Thirdly, while one might disagree with the particular modular view advocated by Chomsky, and in particular the ‘radical autonomy’ of syntax, it is difficult to deny that there are psycholinguistic examples (e.g. dissociations) which support specialisation or a some form of modularity in the present state of knowledge. Finally, in the systemic approaches which connectionism takes as a point of departure, the quantitative complexity factors are in fact functional qualitative factors. It is often such factors which lead systems to diverge. From this angle, and in terms of complexity and therefore also of functional properties, neural network and neuromimetic networks are within orders of magnitude which are not fully comparable. Indeed, it cannot be excluded that the typical properties of each of them might be substantially modified as sizes increase.

The view advocated in Durand and Laks (2003), if one wants to have recourse to connectionism, is to see this framework as allowing us to explore functional and cognitive processes in what is, in our current state of knowledge, a plausible framework from the point of view of neurobiology. Thus, the neuromimetic metaphor is only a first approximation. We do not see connectionism as a realistic model (except by fiat) but as an interesting way of simulating a number of higher cognitive processes (whether linguistic or not) such as reasoning, drawing inferences, or categorising. The position taken by Laks (1996) is that connectionism offers an intermediate level of modelling. This level seems interesting from the point of view of cognition. The reason is that, if the physical and neuronal level is ultimately the *causal* level, it is not (partially for the complexity reasons raised above) analysable or penetrable as such. This has always posed severe problems for strictly physicalist approaches. To solve them, a better strategy is to analyse neurophysiological causality and cognitive processes in general at two levels: on the one hand, the level of the concrete implementation of these processes which is solely neurophysical and, on the other hand, the description, analysis and understanding of these same processes which are, whether we like it or not, constructed on the basis of a symbolic and discursive vocabulary and belong therefore to a quite distinct symbolic level.

In relation to some of the ideas presented above, it is interesting to note the evolution of Chomsky’s thinking. In a widely quoted article on the evolution of

language, Hauser, Chomsky et Fitch (2002) try to tease out what may be special about language and what might plausibly have evolved from other systems or faculties.⁷ They establish a distinction between the faculty of language in a broad sense (FLB) and the faculty of language in a narrow sense (FLN). For them, the key component of FLN is a computational system (narrow syntax) that generates internal representations and maps them into the sensory-motor interface by the phonological system, and into the conceptual-intentional system by the (formal) semantic system. Minimally, the key property of FLN is recursion, which they attributed to narrow syntax in the conception the outline. FLN takes a finite set of elements and yields a potentially infinite array of discrete expressions. This capacity of FLN yields discrete infinity (a property that also characterized the natural numbers). Each of these discrete expressions is passed on to the sensory-motor and conceptual intentional systems, which process this information in the use of language' (2002:1571). But as noted in Laks (2008) this poses a severe epistemological problem in that it can be argued that this position opens the door for a return to a neo-behaviorist approach wherein data are no longer marginal (i.e. the data is not a mere trigger for the unfolding of an innate language faculty but part and parcel of an interactional scenario between humans and their social and linguistic environment). As Laks puts : "Voici donc une bonne nouvelle pour conclure: pour Chomsky lui-même, *la phonologie est une science des usages et il ne saurait donc, ni en droit ni en fait, y avoir d'autre phonologie que de corpus!* (This is excellent news leading to the following conclusion : for Chomsky himself, phonology is a science of usage and there cannot be, neither de jure nor de facto, a phonology which is not corpus-based! [my translation, JD]).

The conclusion drawn by Laks is of course provocative but it does underline to what extent there has been a sea-change in linguistics. Distributionalism, which seemed dead and buried, has made a strong comeback within the world of linguistics. As argued within this section, a strong emphasis on corpora and distributional properties of corpora need not be connected with an anti-cognitivist stance. There needn't be an irreconcilable opposition between the rationalist view and the empiricist view described in the quotation by John Goldsmith placed as an epigraph to this chapter. But, if for epistemologically inclined linguists such a choice has to be made, then it is clear that much contemporary work, including the research on French reported here, argues for the data as not being mere entries in the linguist's diary but the stuff on which theories stand and fall and which must be united with them.

5. Which way forward?

In the foregoing sections, a range of approaches have been presented. All point to a rehabilitation of authentic data such as the ones gathered in modern corpus approaches as part of an investigation of language structure as well as language processing. I should make it clear that for me the judgments made by linguists are indispensable. In that sense, I concur with Chomsky's stance. Earlier on, examples from co-reference used by Smith (2004) were mentioned. One would agree, as stressed by all linguists, that for familiar simple examples such as those in (9):

⁷ See too Fitch, Hauser and Chomsky (2005) for further clarifications.

- (9) (i) John_i said he_i would buy the book
 s(ii) *He_i said John_i would buy the book
 (iii) After he_i came, John_i said he would buy the book

a co-reference relation, as specified by the indices, is possible in (i) and (iii), but not in (ii). Work of great complexity has been devoted to this issue on the basis of initial insights such as these. It does not follow that the **only** way to proceed is to invent examples from artificially constructed sentences (even if this will also be crucial). Classical cognitivists in the Chomskyan school tend to argue that the goal of linguistics for them bears little relation to what is done within the all-embracing, data-driven approaches mentioned in the previous section. They will point out that they are interested in the faculty of language in the narrow sense and that this requires working with constructed data which will test the limits of the possible and the impossible. Again, the argument has some force. But we have seen that what is included in the faculty of language in the narrow sense may well not have a single interpretation and that Chomsky himself appears to have retreated to an extremely minimalist interpretation (in his technical sense and the informal sense). My own position is that what counts as ‘core linguistics’ in the Chomskyan tradition is too narrow. I am convinced that combinatorics and recursion are central characteristics of human language. But many specialists have stressed that these properties are not unique not language and that they are not the only ones to take into account. In the ‘énonciativiste’ (enunciativist) tradition defended in France (see e.g. Benveniste 1966, 1974, Culioli 1992, 1995, Durand 2000), the deictic coordinates of speech (speaker/hearer coordinates, temporal indices, spatial anchoring) are for instance seen as constitutive of the language faculty and not as pragmatic external factors. The way utterances are constructed reflects these deictic coordinates and the interactional unfolding of communicative events. If so, paying attention to the actual structuring of utterances in attested corpora is a prerequisite to achieving observational and descriptive adequacy. Without these, explanatory adequacy is unlikely to be attained. In French, spoken utterances diverge quite considerably from examples invented by linguists which, as observed in section 3, often reflect normative written conventions (as particularly stressed in the work of Blanche-Benveniste and her collaborators, e.g. Blanche-Benveniste 2006).

One can still wonder why ‘intuition’ has to remain central to linguistic work. I venture that this is linked to ‘reflexivity’ (or reflectiveness) which is also a distinctive property of language. As stressed in section 2 (see too note 3), reflexivity is the ability to use language to speak about language and is omnipresent in everyday use (cf. an expression such as ‘as I am saying’ or a question such as ‘am I speaking too quickly?’) It is difficult to establish what is at the root of reflexivity. Is it that communication includes feedback loops inducing modifications to what is being uttered? If so, animal systems allow for modulations of signals without creating reflexivity. Is it linked to the fact that reference in human language is not fixed (one can lie)? Is it a consequence of the enunciative structuring of speech as the French would call it? Is it a by-product of consciousness which after decades of work trying to make it disappear is still with us (see Searle 2004). I have no precise answer to this question. But the fact is that self-reference is a central property of language and that the system can also modify itself as it goes along (not just in the creation of words but also in subtle quasi-quotation self-modifying mechanisms). This seems to me neither properly rule-governed nor rule-breaking to use Chomsky’s traditional distinction.⁸ All this makes human language a

⁸ Some of the questions raised in this conclusion are dealt with in Notari (2008) in a most interesting way.

very special type of semiotic system which cannot be fully investigated as an object out there in the phenomenal world. At the same time, we can see that reflexivity places restrictions on how far we can go in investigating language. The work of the logician Gödel (1931) is often cited as having profound consequences for the relationship between statements in a formalized language like that of arithmetics and meta-statements about the object language. Thus, “if arithmetic is consistent its consistency cannot be established by any meta-mathematical reasoning that can be represented within the formalism of arithmetic” (Nagel and Newman 1971: 96). Whether the results have any consequence for the formalization of human language I am not sure. Nagel and Newman (1971) warn us that “Gödel’s proof should not be construed as an invitation to despair or as an excuse for mystery-mongering” (1971: 101). Analogically, nevertheless, I would like to suggest that Gödel’s work indirectly emphasizes the complexity of studying a system such as human language which allows the intermixing of levels of reference. The formal study of language must be pursued in all areas from phonetics to pragmatics. But there may well be limits to our understanding of language and the mind. If so, this should lead to more tolerance within linguistics. At any rate, to go back to the main topic of our discussion, what has been argued here is that the data relevant for linguistic theory do not lie out there waiting for the lucky analyst to discover them, even through bold hypotheses. The data have to be patiently constructed and gathered experimentally. Whatever the limitations of the strategy advocated here, I think it represents an important advance over the reiteration of data which is nobody’s data – the linguistic Frankenstein’s monsters dubbed ‘standard languages’ in their written normative instantiations.

References

- Anderson, J.M. 2005. “Structuralism and Autonomy”. *Historiographica Linguistica* 32(1-2).117-148.
- Anttila, S. 1997. “Statistical methods and linguistics”. *The Balancing Act: Combining symbolic and statistical approaches to language*, J. Klavans and P. Resnik (eds). Cambridge: MIT Press. 1-26.
- Bender, E.M. 2001. *Syntactic Variation and Linguistic Competence. The case of AAVE copula absence*. Stanford, CA: Stanford University dissertation.
- Benveniste, E. 1966. *Problèmes de linguistique générale*. Tome 1. Paris: Gallimard.
- Benveniste, E. 1974. *Problèmes de linguistique générale*. Tome 2. Paris: Gallimard.
- Blanche-Benveniste, C. 2006. The Case of French Language. *Spoken Language Corpus and Linguistic Informatics*, Y. Kawaguchi, S Zaima and T. Takagaki (eds). Amsterdam/Philadelphia: John Benjamins. 35-66.
- Boersma, P. 1998. *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. The Hague: Holland Academic Graphics.
- Bybee, J. 2001. *Phonology and Language Use*. Cambridge: Cambridge University Press.

- Bybee, J. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.
- Chambers, J.K. and P. Trudgill. 1980. *Dialectology*. Cambridge: Cambridge University Press.
- Chomsky, N. 1965. *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.
- Chomsky, N. 1968. *Language and Mind*. New York: Harcourt, Brace & World.
- Chomsky, N. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.
- Chomsky, N. 2000. *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.
- Clark, B. 2004. *A Stochastic Optimality Theory Approach to Syntactic Change*. Stanford, CA: Stanford University Dissertation.
- Clark, B. 2005. On stochastic grammar. *Language* 81.207-217.
- Culioli, A. 1990. *Pour une linguistique de l'énonciation. Opérations et représentations*. Paris: Ophrys.
- Culioli, A. 1995. *Cognition and Representations in Linguistic Theory*. Amsterdam: John Benjamins.
- Delattre, P. 1951. *Principes de phonétique française à l'usage des étudiants anglo-américains*. Middlebury College.
- Delattre, P. 1966. *Studies in French and Comparative Phonetics*. The Hague: Mouton.
- Dell, F. (1973/1985). *Les règles et les sons*. 2nd edition, 1985. Paris: Hermann.
- Durand, J. 1982. "A propos du préfixe *anti-* et de la parasyntèse en français". *Essex Occasional Papers* 25.1-34.
- Durand, J. 2000. "French Linguistics and énonciation: Meanings, utterances and representational gaps", *Currents in Contemporary French Intellectual Life*, C. Flood & N. Hewlett (eds). London: MacMillan. 76-95.
- Durand, J. 2006. "Mapping French Pronunciation. The PFC project". *New Perspectives on Romance Linguistics*. Vol. 2: *Phonetics, Phonology and Dialectology*, J.-P. Montreuil and C. Nishida (eds). Amsterdam: John Benjamins. 65-82.
- Durand, J., B. Laks and C. Lyche. 2002. "La phonologie du français contemporain: usages, variétés et structure". *Romanistische Korpuslinguistik - Korpora und gesprochene Sprache/Romance Corpus Linguistics - Corpora and Spoken Language*, C. Pusch and W. Raible (eds).Tübingen: Gunter Narr Verlag. 93-106.
- Durand, J. and C. Lyche. 2003. "Le projet 'Phonologie du Français Contemporain' (PFC) et sa méthodologie". *Corpus et variation en phonologie du français : méthodes et analyses*. Toulouse: Presses Universitaires du Mirail. E. Delais and J. Durand . 212-276.

- Durand, J. and C. Lyche. (2008) "French Liaison in the Light of Corpus Data". *Journal of French Language Studies* 18(1).33-66.
- Ernestus, M. and R. Harald Baayen. 2003. "Predicting the Unpredictable: Interpreting neutralized segments in Dutch". *Language* 79.5-38.
- Féry, C. 2003. "Liaison and Syllable Structure in French". Postdam, Ms.
- Fitch, W.T., M.D. Hauser and N. Chomsky. 2005. "The Evolution of the Language Faculty: Clarifications and implications". *Cognition* 97.179-210.
- Fouché, P. 1959. *Traité de prononciation française*. Paris: Klincksieck.
- Fradin, B. 1997. "Esquisse d'une sémantique de la préfixation en *anti*". *Recherches linguistiques de Vincennes* 26.87-112.
- Goldsmith, J. 2005. Review of *The Legacy of Zellig Harris: Language and information in the 21st century*. Vol. 1. *Philosophy of Science, Syntax and Semantics*, edited by B. Nevin. *Language* 81(3).719-736.
- Gödel, K. 1931. "Übe formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme". *Monatshefte für Mathematik und Physik* 38.173-98.
- Haiman, J. 1994. "Ritualization and the Development of Language". *Perspectives on grammaticalization*, W. Pagliuca (ed.). Amsterdam: John Benjamins. 3-28.
- Harris, Z.S. 1954[1964]. "Distributional structure". *Word* 10(2-3).146-162. (Reprinted in *The Structure of Language. Readings in the philosophy of language*, Fodor, J.A. and J.J. Katz (eds)(1964). Englewood Cliffs, New Jersey. 33-49.)
- Hathout, N., F. Montermini, L. Tanguy. 2008. "Extensive Data for Morphology: Using the World Wide Web." *Journal of French Language Studies* 18(1).67-85.
- Hathout, N., M. Plénat and L. Tanguy. "Enquête sur les dérivés en -able." *Cahiers de grammaire* 28.49-90.
- Hauser, M.D., N. Chomsky, and W.T. Fitch. 2002. "The Faculty of Language: What is it, who has it and how did it evolve?" *Science* 298:1569-1579.
- Holland, J.H. 1998. *Emergence: From chaos to order*. New York: Basic Books.
- Hockett, C.F. and S.A. Altman. 1968. "A note on design features". *Animal Communication: Techniques of Study and Results of Research*, T.A. Sebeok (ed). Bloomington: Indiana University Press.
- Hopper, P. 1987. "Emergent Grammar." *Berkeley Linguistics Society* 13:139-157.
- Huck, G.J. and J. Goldsmith. *Ideology and Linguistic Theory: Noam Chomsky and the deep structure debates*. London: Routledge.
- Lindblom, B., P. MacNeilage and M. Studdert-Kennedy. 1984. "Self-organizing Processes and the Explanation of Phonological Universals". *Explanations for Language Universals*, B. Butterworth, B. Comrie and O. Dahl (eds). New York: Mouton.

- Hume, D. 1740[1978]. *A Treatise of Human Nature*. Edited by L.A. Selby-Bigge. (Second edition revised by P.H. Nidditch, 1978.) Oxford: Clarendon Press.
- Irwin, J.V. and S.P. Wong (eds). 1982. *Phonological Development in Children: 18 to 72 months*. Carbondale: Southern Illinois University Press.
- Jakobson, R. 1960. "Closing Statements: Linguistics and poetics." *Style in Language*, T.A. Sebeok (ed.). Cambridge, Mass.: MIT Press.
- Labov, W. 1969. "Contraction, Deletion, and Inherent Variability of the English Copula". *Language* 45:715-762.
- Labov, W. 1994. *Principles of Linguistic Change. Internal factors*. Oxford: Blackwell.
- Labov, W. 2001. *Principles of Linguistic Change. Social factors*. Oxford: Blackwell.
- Laks, B. 2002. "Description de l'oral et variation: La phonologie et la norme". *L'information grammaticale* 94.5-11.
- Laks, B. 2005. "La liaison et l'illusion". *Langages* 158.101-126.
- Laks, B. 2008. "Pour une phonologie de corpus". *Journal of French Language Studies* 18. 3-32.
- Lappin, S. 2005. "Machine Learning and the Cognitive Basis of Natural Language", *Computational Linguistics in the Netherlands 2004*, T. van der Wouden et al. (eds). Utrecht: LOT. 1-11.
- Lappin, S. and S.M. Shieber. 2007. "Machine Learning Theory and Practice as a Source of Insight into Universal Grammar. *Journal of Linguistics*, 43(2).393-427.
- Léon, P.R. (1992). *Phonétisme et prononciations du français (avec des travaux d'application et leurs corrigés)*. Paris : Nathan.
- Lyons, J. 1977. *Semantics*. (2 vols.) Cambridge: Cambridge University Press.
- Lyons, J. 1995. *Linguistic Semantics. An introduction*. Cambridge: Cambridge University Press.
- Manning, C. 2002. Review of *Beyond Grammar: An experience-based theory of language* by Rens Bod. *Journal of Linguistics* 3.441-442.
- Manning, C. 2003. "Probabilistic syntax," *Probabilistic Linguistics*, R. Bod, J. Hay and S. Jannedy (eds). Cambridge, MA: MIT Press. 289-341.
- Morin, Y.-C. 1987. "French data and phonological theory". *Linguistics* 25. 815-843.
- Morin, Y.-C. 2000. "Le français de référence et les normes de prononciation". *Cahiers de l'Institut de linguistique de Louvain* 26(1).91-135.
- Morin, Y.-C. (2005). "La liaison relève-t-elle d'une tendance à éviter les hiatus. Réflexions sur son évolution historique ". *Langages* 158.8-23.
- Nagel, E. and J.R. Newman. 1959. *Gödel's Proof*. London: Routledge and Kegan Paul.

Nagy, N. and B. Reynolds. 1997. "Optimality Theory and Variable Word-Final Deletion in Faeter". *Language Variation and Change* 9.37-55.

Newmeyer, F. J. 2003. "Grammar is Grammar and Usage is Usage". *Language* 79(4).682-797.

Newmeyer, F. J. 2005. "A reply to the critiques of 'Grammar is grammar and usage is usage'". *Language* 81(1).229-36.

Notari, C. 2008. *Métaphore de l'ordinateur et linguistique cognitive*. Unpublished PhD dissertation. University of Toulouse II.

Nyrop, K. 1936. *Grammaire historique de la langue française*. Vol. 3: *Formation des mots*. Second edition. (First edition, 1908.) Copenhagen/Paris: Nordiske Forlag.

Paolillo, J.C. 2002. *Analyzing Linguistic Variation: Statistical Models and Methods*, CSLI Publications.

Pinker, S. 2007. *The Stuff of Thought*. London: Allen Lane.

Sampson, R. 2001. "Liaison, Nasal Vowels and Productivity". *Journal of French Language Studies* 11.241-58.

Schane, S. 1968. *French Phonology and Morphology*. Cambridge, Mass.: MIT Press.

Searle, J.R. 2004. *Mind. A brief introduction*. Oxford: Oxford University Press.

Selkirk, E. 1972. *The Phrase Phonology of English and French*. Ph.D. dissertation, MIT. (1980, New York: Garland.)

Skousen, R. 1989. *Analogical Modeling of Language*. Dordrecht: Kluwer.

Smith, N. 1989. *The Twitter Machine. Reflections on language*. Oxford: Blackwell.

Smith, N. 2004. *Chomsky. Ideas and ideals*. Second edition. Cambridge: Cambridge University Press.

Steriade, D. 1999. "Lexical Conservatism in French Adjectival Liaison". *Formal Perspectives in Romance Linguistics*. B. Bullock, M. Authier and L. Reed (eds). Amsterdam: John Benjamins. 243-270.

Tranel, B. (1996). "French Liaison and Elision Revisited: a unified account within Optimality Theory". *Aspects of Romance Linguistics*. C. Parodi, C. Quicoli, M. Saltarelli and M.L. Zubizarreta (eds). Washington, DC : Georgetown University Press. 433-455.

Tranel, B. (1999). Suppletion and OT: On the issue of the syntax/phonology interaction. *Proceedings of the Sixteenth West Coast Conference on Formal Linguistics*. E. Curtis, J. Lyle and G. Webster (eds). Stanford: CSLI. 415-429.

Weinreich, U., W. Labov and M. Herzog. 1968. "Empirical Foundations for a Theory of Language Change". *Directions for historical linguistics: A symposium*. W. P. Lehmann and Y. Malkeil (eds). Austin: University of Texas Press. 95-188.

Whorf, B.L. 1956. *Language, Thought and Reality*. Selected Writings of Benjamin Lee Whorf, edited by J.B. Carroll. Cambridge, Mass. The MIT Press.

Yang, C. D. 2004. "Universal Grammar, Statistics or Both?" *Trends in Cognitive Science* 8(10). 451-456.