The Science of Linguistics

Eric Laporte

The Science of Linguistics

Éric Laporte

Éric Laporte is a Professor of Computer Science and Linguistics at Paris-Est Marne-la-Vallée.

Article

No description of any language encompasses both its grammar and its lexicon. Take, for example, the French verb *s'écraser*:

> *L'avion s'est écrasé en mer.*
> The plane crashed at sea.

Do French speakers use the first example transitively?

> *Le pilote a écrasé l'avion en mer.*
> The pilot crashed the plane at sea.

Not frequently. This sort of question, multiplied by the number of similar constructions in French, or in any other language, suggests just why no linguistic description is remotely adequate to the facts. There are simply too

many facts.

Few dictionaries and grammars are fully reliable, even when the facts in question are easily verified. French dictionaries describe nouns referring to families of plants and animals as plural: *cucurbitacées* (Cucurbitaceae). This implies that they are not used in the singular. But they are: *La courgette est une cucurbitacée* (Zucchini belongs to the Cucurbitaceae family). None of the three major English-French dictionaries have an entry or sub-entry for "as yet," which appears only in examples.[1] And traditional grammars tend to ignore some syntactic constructions, especially idioms.

Lélia Picabia suggested that the French adjective *susceptible* (likely) can only be applied to human subjects:[2]

> *Rose est susceptible de devenir mère.*
> Rose might become a mother.

Not so:

> *Le cas est susceptible de se produire.*
> Such a case might happen.

On the other hand, her work *does* provide the data by which her conclusions may be checked. A minority of French predicate adjectives, she argues, are "defined by the constraints [they] impose on the subject and

complement."[3] Her tables provide examples; her conclusions are consistent with her data. This is a step, however small, in the right direction.

Most papers and books in linguistics are otherwise. Catherine Léger, in a paper representative of contemporary linguistics, defines the *effective* adjectives as those French adjectives with sentential complements describing "a subject's relationship—whether causal, potential or other—to the performance of an action."[4] Members of the effective class, she goes on to assert, "all share the property that their complements must be tenseless."[5] Checking a general claim of this sort is difficult. Léger does not provide a list of all the effective adjectives, and describing a subject's relationship to the performance of an action is both vague and fuzzy. She cites *susceptible* as an example. But, consider:

> *Paul est susceptible de tomber.*
> Paul might fall down.
>
> *La couleur est susceptible d'être légèrement différente.*
> The color may be slightly different.

Falling down may represent the performance of an action, but not being slightly different. Léger's definition of an effective adjective yields different results for the same adjective.

Is *susceptible* an effective adjective in French?

There is no way to tell.

The Search for Definitions

BETWEEN 1925 and 1955, American structural linguists introduced distributional analysis into linguistics. In distinguishing countable and uncountable nouns, such as “house” and “milk,” Leonard Bloomfield, to take a prominent example, placed little faith in his own semantic intuitions.[6] He relied on native speakers: *they* determined whether some forms were in use or not. It was not until the 1960s that the methods of distributional analysis were refined by the French linguist Maurice Gross, and his students and collaborators. Lexicon-Grammar was the result. Gross and his students systematically studied the forms that can appear in the subject and complement of the verb *écraser*, in all of its senses.[7] On the basis of distributional differences, they described distinct entries for sentences such as *L’avion s’est écrasé en mer* and *Il a écrasé l’ail*. The pilot *crashed* the plane; but he *crushed* the garlic. Gross rejected *Le pilote a écrasé l’avion en mer*, and so assigned *crash* to a class of intransitive verbs, and *crush* to a class of transitive verbs. Distributional analysis became an experimental protocol.[8]

I discuss Gross’s method further down. Keep reading.

What of corpus linguistics and construction grammars? Corpus linguistics, as the name might suggest, involves the study of a specific body of examples—hence, the corpus. In Henry Kučera and W. Nelson Francis's *Computational Analysis of Present-Day American English*, the corpus contained roughly 500 samples of text, totalling about one million words. With the data at hand, Kučera and Francis did what computational linguists always do: they tested the data. Construction grammars are exercises in generative semantics, the strange glow thrown off by generative syntax in the 1970s. Generative semantics and construction grammar rejected the structuralist approach to language through linguistic form, and chose to focus on meaning instead. Both corpus linguistics and construction grammars make use of distributional analysis, but only as an adjunct to intuition.[9] From the first, Noam Chomsky found operational procedures useless.[10] Most linguists have today abandoned the attempt to collect empirical evidence in a formal and scientific way.

Have they found anything better, I wonder?

Reproducible Subjectivity

WHATEVER their various professional affiliations, all linguists must, in the end, depend on intuition—their intuition and the intuition of their informants. Those who speak English

determine whether certain sentences are good, bad, or both, or neither. Those who do not, have nothing to say. Beyond looking to the users of a language, where would linguists look? The case of countable nouns is again instructive. Bloomfield's procedure for deciding which nouns were countable and which were not was the first to establish a criterion of countability. Could a noun such as "house" occur in the singular without a determiner? A criterion given, he then asked what speakers of the language said. "There is house?" No. "There is milk." Yes. The criterion and its employment work hand in glove. The criterion refines the question; the answers make use of the criterion. Bloomfield chose to classify the countable and the uncountable nouns on the basis of whether or not they required, or took, determiners, because this criterion generated agreement among observers.[11] Examples such as "house-proud," "household," and "house-bound" might suggest that so far as this criterion goes, tune-ups may be required. How about "the house-proud household is house-bound?"

When linguists assume that *susceptible* (likely), but not *digne* (worthy), "describes a subject's relationship ... to the performance of an action," agreement among observers is more precarious.[12] Very often, nothing is reliably observed. The formal procedures of structural and distributional linguistics are a way to avoid this problem by redirecting attention to a narrower and more obvious target.

Reproducibility is relevant to technology. Of course it is. The wrong data will inevitably corrupt computer applications. Imperfect dictionaries and

grammars do help with machine translation, but their reliability is certainly a factor in their performance. Linguists interested in language resources for Natural Language Processing (NLP) frequently assess “inter-judge agreements,” and as frequently discover that it is often low.[13] Judgment is more a matter of opinion than data. This situation rarely leads NLP researchers to question the formal basis of their enterprise.[14]

Sociolinguistic and idiolectal variations lead to countless differences in detail. Lexicon-Grammars handle this by comparing independent judgments, the procedure followed by Gross and his colleagues. Between 1968 and 1984, they met regularly in order to classify various French verbs. Most of the lexicon-grammar of French is freely available, and thus exposed to critical evaluation by other linguists.

Psychological bias should never be underestimated. Christian Lehmann notes that such biases may result from prejudice toward a hypothesis, or toward literary norms; “few linguists,” he observes acidly, “have escaped the temptation to dress the data they produce according to the theory they cherish.”[15] Studies that systematically scan a lexicon are less vulnerable to such biases simply because numerous observations are required to validate a hypothesis. Methods of prophylaxis are simple and effective; they involve nothing more than the comparison of independent judgments by several linguists, and the publication of extensive results.

Lexicon-Grammars

IF THERE are obstacles to reproducibility in linguistics, lexicon-grammars address them by requiring observers to be properly trained. Reproducibility is never perfect, but what is?

The level of reproducibility is connected, among other things, to how strongly the observer belongs to a speech community. We all belong to at least one. Sharpness of judgment depends on the skill and training of the observer; reproducibility is enhanced by the kind of extensive practice Gross and his colleagues had, and by collective sessions during which they controlled one another's judgments and analyses.[16] The observer must imagine contexts in which a sequence might make sense and be natural. This ability improved with training.

During the study of an individual linguistic property, hundreds of lexical entries are reviewed.

> Lexical information is represented in the form of *tables*. Each table puts together elements of a given category (for a given language) that share a certain number of *defining features*, which usually concern sub-categorization. These elements form a *class*. These tables are represented as matrices: each row corresponds to a lexical item of the corresponding class, each column lists all features that may be valid or not for the different members of the class; at the intersection of a row and a column, the symbol

> + (resp. −) indicates that the feature corresponding to the column is valid (resp. not valid) for the lexical entry corresponding to the row.[17]

Repetition serves to refine and sharpen definitions and homogenize their encoding. During analysis of entries, LG linguists sometimes notice a problem in the definition of a certain property. Is it one phenomenon at work? Or two? The solution is simple. Either give up the study of this property or redefine it as one of several properties. The more a lexicon-grammar table records reliable judgments, the more it is useful for syntactic parsing and other applications.[18]

Once the tables are published, the results can be checked by other native speakers. By 1985, a large collection of tables of French verbs and predicative nouns had been published together with a few tables of English predicates. Tables of predicates in other languages have been published since then.[19] This work remains unchallenged.

For the moment, there are few debates about reproducibility in linguistics. One exception is Walter Bisang, who suggested several solutions for enhancing reproducibility: “check[ing] each sentence with twenty or thirty informants,”[20] “work[ing] with about ten different lexical forms (multiple lexical variants) that show the same effect,” and “systematic[ally] study[ing] the social basis of variation.”[21] Such practices increase both the number of observers and the targets of observation. They require less care, training and skill than lexicon-grammars. This is not a recommendation in

their favor.

Acceptability

TO BE ACCEPTABLE, a form must be meaningful. When linguists assess the acceptability of a form, they assess the probability that it might be used in some context to convey information. Some forms do not make sense in any context:

> Ideas sleep down.
>
> That ideas sleep down swims.
>
> Ideas sleep down swims.
>
> Sleep down swims.

Acceptability is a simplified form of probability: an unacceptable sequence is unlikely to occur, whether in discourse or anywhere else. Since probabilities belong to a continuous scale, linguistic reality is more complex than anything a binary view of acceptability might suggest. In practice there is no way to measure the probability of any sequence in a language. Where would one start?

Some linguists multiplied levels of acceptability. Starting with the triplet of

acceptable, unacceptable, and unknown; they quickly went on to a quartet,[22] and even a septet.[23]

These proposals are unreasonable, if only because a seven-fold distinction (good, not so good, not really so good, really not so good, not so hot …) is less reliably observable than a distinction based on an old-fashioned, two-fold yes or no. Ellen Gurman Bard et al. experimented with an open scale requiring several informants for each judgment, a solution incompatible with any systematic description of the lexicon.[24] If the ensuing chaos has not been recorded, it may, on the other hand, be imagined.

More concerned by the faculty of language than language itself, Chomsky contrasted acceptability with grammaticality.[25] Grammaticality is laxer. Acceptable sentences must be meaningful, whereas grammatical sentences may be meaningless.

There are several reasons to find acceptability more interesting than grammaticality. Why should grammars account for such nonsense as “Ideas sleep down?” If we are hoping to discover potential computational applications in syntactic parsing or NLP, what is the point in parsing or generating it? To distinguish grammatical from ungrammatical sequences, Chomsky sometimes appeals to such features as prosody or ease of retention.[26] These criteria might point to decreasing grammaticality, from “Ideas sleep down” to “Sleep down swims,” but they are vague. These observational issues obscure Chomsky’s distinction in the case of

meaningless sequences. Grammaticality is less reliably observable than acceptability.

Those who use the term grammatical may, in fact, find the notion of acceptability more relevant: in practice, many sequences that, like “ideas sleep down,” are not in use, and, even though they conform to Chomsky’s criteria of grammaticality,[27] are rejected by native speakers.[28] Witness:

> Karen has not probably left,

which is both widely rejected and never in use.

Differential Assessment of Meaning

INTRODUCED by Gross, the technique of differential assessment of meaning, or DAM, is essential to the observation of practically any formal property.[29] Bloomfield’s criterion for count nouns implicitly involves DAM, because in employing it, he checked for unexpected changes in meaning. “Sloth kept me from getting up” points to an uncountable noun. I may be slothful, but I cannot count the sloths. Unless, of course, I am thinking of sloths, as in “the zoo’s sloth is pregnant.” A suitable interpretation of sloth and sloths must distinguish cases where meaning changes. DAM explains why introspective distributional analysis is not

obsolete at a time when technology and the availability of huge corpora instantly tells the linguist that “sloth” occurs without a determiner in the singular. Technology facilitates the practice of distributional analysis by extracting examples, but does not assess meaning changes.

Some form of DAM is essential to sound linguistic practice because the method allows linguists reliably to apply distributional and transformational tests. If, in the case of “collect,”

> Karl collects waste in the markets,

is understood as “Karl does the collection of waste in the markets,” then “Karl makes a collection of waste in the markets,” if understandable, has a different, slightly more puzzling meaning. In “Karl collects waste in the markets,” the direct object accepts a definite determiner: “Karl collects the waste in the markets.” Now, in most contexts,

> Karl collects pictures,

does *not* mean “Karl does the collection of pictures,” but rather “Karl makes a collection of pictures,” and the direct object of “Karl collects pictures” does not accept a definite determiner. “Karl collects the pictures” paraphrases “does the” but not “makes a.”

These observations lead to a distinction between at least two

“collect”/“collection” pairs, with distinct meanings, since they respond differently to the same two criteria. Depending on the pair, translations into French are different: *ramasser* (collect) for *collects waste* versus *faire une collection* (collect) for *collects pictures*.

Gross used introspection on a large scale in studying the syntax and lexicology of English and French, and developed methodological precautions to make his observations reproducible. He chose to rely on binary acceptability and differential assessment of meaning because these types of empirical observation involve reproducible introspection. Introspective procedures involving artificial sequences can produce authentic empirical evidence if the observers are rigorous, and if they participate in collective sessions while selecting carefully the questions to be answered.

Lexicon-Grammar studies contributed massively to the understanding of support verbs.[30] These are verbs that lend a helping hand: In “Bob paid or drew or gave special attention to the details of his own funeral,” the verbs “paid,” “drew,” and “gave” provide inflectional and aspectual information. It is “attention” and “detail” that do the heavy lifting.[31]

These methods have brought descriptive linguistics closer to an empirical discipline. Few linguists learned from this work, perhaps because few linguists were aware of it. Supporters of most trends either dispense with introspection, or use it without any noticeable precautions.[32]

Introspection and Corpora

THE PROCEDURES of empirical linguistics rely on introspection or corpus observation. Both are useful, since they give access to two aspects of language reality and use. Corpora are important for forms that might otherwise go unnoticed, while introspection is needed to distinguish rare forms from those that are not in use. But the empirical task of collecting introspective data must follow systematically controlled procedures.

Corpus exploration is in the twenty-first century easier, more efficient, and scientifically safer than introspection. It may even provide evidence of acceptability, as in the case of “there was house.” This form’s absence from a large corpus is evidence that it does not occur in English. There is no other explanation. The three words are common, and “there was *N*,” where *N* stands for a noun, is a frequent sentence pattern. Still, with less frequent words and syntactic constructions, introspection remains essential. If a large corpus contains only three occurrences of the verb “honorificabilize,” all of them in the active voice, is this evidence that the verb does not have a passive voice? It is a question that no corpus can properly answer. If the form does not turn up in one place, it may turn up in another. To determine whether “honorificabilize” admits the passive, a more decisive empirical procedure is needed. The linguist must produce

artificial forms. The resort to made-up forms is inevitable, because, as Chomsky notes, “the set of grammatical sentences cannot be identified with any corpus of utterances obtained by the linguist in his field work.”[33]

DAM also requires introspection. Limits of variability can be discovered only by observing unacceptable sequences and comparing their meanings. Witness the perfectly ordinary

> *L’avion s’est écrasé en mer,*

descending ignominiously to

> *Le pilote a écrasé l’avion en mer.*

The procedures of structural linguistics require introspection when differential assessments of meaning are needed. At a time when efficient corpus exploration was not available, Gross and his colleagues used introspection. They began to use corpora as soon as the tools were available. But they continued to practice controlled introspection to test each construction systematically. Some linguists reject the idea that introspection might provide empirical evidence because, in the words of Geoffrey Sampson, it is “flatly contradictory to describe ‘intuitions’ as ‘empirical’ data.”[34] This is a criticism never applied to physicists, who rely in the end on their eyes or who distinguish between theories by taste.

Most observers, Lehmann remarks, do not, and, presumably, cannot, master every last sociolect.[35] True enough. There is always the possibility that a bizarre construction may be found in some tucked-away community of idiosyncratic native speakers prepared defiantly to accept “there is house” on the grounds that it is just like “there is fire.” In advancing these criticisms, Sampson and Lehmann have overlooked Gross’s methodological contributions.[36] The long-standing polarization of linguists between generative and corpus linguistics has encouraged the view that if generative linguistics fails properly to make use of introspection, then introspection cannot be used at all.

Reproducible and Non-Reproducible

IF A SEMANTIC property is obscure or difficult to observe, it may yet be correlated with formal properties that are not. Semantically gradable adjectives, such as “young,” often combine with “very.” Non-gradable adjectives, such as “dead,” do not. Cynthia may be very young, but while John can be almost dead, he cannot be very dead. Either he is or he isn’t. Adjectives such as “gorgeous” do not feel clearly gradable. Cynthia is so gorgeous, but is she *very* gorgeous? There is a difference in definiteness between the semantic and distributional analysis of certain adjectives, and semantic and distributional properties pose distinct methodological problems.

When formal and semantic features are correlated, there may well be a causal nexus between them. The fact that an adjective is semantically gradable explains the fact that the adjective does not admit adverbs of degree. Empirical evidence about distributional properties is essential to the verification of such hypotheses.

In some cases, predicates denoting several entities require plural agreement, as in “John collects paintings,” but not, “John collects a painting.” It is the semantic feature of the verb “to collect” that is the cause of the formal feature expressed in plural agreement. As usual, focusing on the formal feature results in superior reproducibility. The intuition governing ascriptions of causality are less reliable. In “his son-in-law married them in their wheelchairs,” the object of “marry” denotes a set of two people, but it can, nevertheless, occur in the singular: “His son-in-law married him in his wheelchair.” The semantic feature does not always cause the formal feature.

Is it really a cause?

Systematic Description of the Lexicon

LEXICAL coverage is a matter of how much of the vocabulary of a language a research project includes in its study. The size of a

lexicon makes coverage difficult, and for the most obvious of reasons: it is enormous.

Lexicon-Grammars provide evidence that chaos prevails in large portions of any lexicon.[37] This raises an unavoidable difficulty. Grammar involves generalizations from a lexicon, and a study encompassing large lexical coverage is the only way to indicate whether a grammatical feature is general. There is nothing to be done about this. Language is unbelievably complex. Still, it is worth noting that the lexicon-grammar of French outperforms in the number of its entries all other major NLP dictionaries in French or English: FrameNet, VerbNet, ComLex, and Meaning-Text.

Bias and Objectivity

IN LINGUISTICS, *objectivity* means that the linguist and his informants are distinct. The linguist listens; his informants talk. Questions of reproducibility are different. Psychological biases are a problem in any case, the more so when "the speaker involved is ... the theorist, so that theory and data are simultaneously produced by the same person at the same time."[38]

Corpus linguists are particularly keen to establish that

> data points that are coded are not made-up, their frequency distributions are

> based on natural data, and these data points force them to include inconvenient or highly unlikely examples that armchair linguists may "overlook."[39]

Generative linguists are plausible targets of such criticism, since they freely resort to introspection and, as Steven Abney observes, inconvenient observations may always be dismissed as a matter of performance.[40]

Neurolinguists and psycholinguists are scrupulous about subjectivity. Preferred sources of empirical evidence are experiments that do not allow a participant to be both a subject and an observer. These very reasonable scruples need not remain the sole possession of neurolinguists or psycholinguists. Linguistic protocols can ensure that subjective results are also reproducible. This is what happens with reproducible acceptability judgments. No neurolinguistic experiment can determine whether the semantic difference between "He made a joke" and "He joked," is the same as the semantic difference between "He reported a joke" and "He joked." This distinction requires the aid of a native speaker.

It is possible to imagine psycholinguistic experiments that aid in the construction of both grammars and their lexicons. Subjects might be asked to distinguish parts of speech. What would be the practical implications? Daunting. Some words are assigned to different parts of speech in different contexts. The word "record" can be both a verb and a noun. There is "Let me record this," but there is also "This is a record," as well as, "Let me get

this on the record." No experiment could rely on a single subject for a given word. In order for twenty informants to pass judgments on 220,000 lemmata,[41] a single experiment would need to be repeated four million times.[42]

And parts of speech are the easy example. Does a given lexical entry, with a given sense, enter into a given syntactic construction? The practical problems are roughly the same as with parts of speech, but on a much larger scale. There are hundreds of syntactic constructions in a language and tens of thousands of words. The compatibility of a syntactic construction with a word may be predictable or unpredictable. For each combination, a trial requires an acceptability judgment. Some sequences might be extracted from corpora, but not all: sequences with rare words or rare syntactic constructions would be more difficult to find.

Objectivity is sometimes incompatible with a systematic study of individual lexical entries and syntactic constructions. Lexicon-Grammar studies have demonstrated, on the other hand, that the requirement of reproducibility *is* compatible with the description of a language on a large scale.

Are Corpora Necessary?

CORPUS linguistics sets a high standard of rigor with respect to factual observations: the linguist studies facts on the ground. Authenticity and objectivity are paramount. If an authentic procedure is one based on the facts, we are returned to our point of departure—the need for empirical evidence. This does not yet seem an advance. If an authentic procedure is one free from manipulation, but excludes introspection, the requirement is counter-productive. If I want to determine whether *Le pilote a écrasé l'avion en mer* can be used in the same sense as *L'avion s'est écrasé en mer*, and if nothing like the second sentence occurs in my corpus, I am bound to cobble together something like it, and judge its acceptability by introspection.

Objectivity is similar. Taken in its strict sense, it excludes introspection. It is better to understand objectivity as one way, among others, of ensuring that observations are reproducible.

Confrontation with Reality

THE LEXICON-GRAMMAR approach to a natural language explicitly aims at large and fine-scaled lexical and syntactic coverage. The results are usually presented in a table, and they show unexpected differences between lexical entries, unexpectedly complex syntactic behavior, and unanticipated discrepancies between form and meaning.

The goal of enumerating all of the linguistic instances that are relevant to a given phenomenon was new in the 1970s, and even today, no other linguistic approach has gone as far. Generative grammars have never implemented any systematic description of both the grammar and the lexicon of a natural language. A great deal has been lost. The verb "irritate" can be interpreted in a physical and a psychological sense. Is it more accurately described with a single entry or with two? The answer suggested by Lexicon-Grammar is that it depends.[43] Are both senses compatible with sentential subjects? A corpus may help in answering this, but only if all the senses and syntactic constructions are well represented. Corpus studies triggered revolutionary improvement in lexicography, but did not dramatically change the way in which NLP dictionaries were constructed. Corpus linguistics may contribute to the lexical coverage of dictionaries by providing a list of unaccounted forms, but it is insufficient by itself to turn these forms into a list of lexical entries together with a formal representation of their properties.[44] Such work requires a further confrontation with linguistic reality.[45]

Idioms, Granularity, and the Intersection

THEN THERE are idioms. Gross and his colleagues have inventoried the various senses of French verbs. Their focus was on full verbs such

- as *passer* (pass; drop in; go through; hand over), but the inventory also produced a list of verbal idioms, such as *passer en revue* (review). There are surprisingly many such verbal idioms; and they figure prominently with respect to syntactic constructions, distributional properties, and transformational properties.[46]

Lexicon-Grammars divide each polysemous word into a finite number of lexical entries: the French verb *écraser* is represented by a "crush" sense, a "crash" sense, and fourteen others. This operation separates the semantic field of a word into discrete parts. It is a prerequisite for the formalization of lexical properties. In a formal system, each property must be a property of *something*, and properties vary according to senses. Most lexicographical and lexicological traditions also separate lexical entries from one another. A word can be separated into lexical entries of higher or lower granularity, depending on how finely semantic distinctions are taken into account. For example, the sixteen-entry description of *écraser* separates a concrete "crush" sense from a concrete "squeeze" sense: *Il a écrasé l'ail* (He crushed the garlic) but *Tu m'écrases le pied* (You are stepping on my foot). A less fine-grained description might merge these entries, and there is no ultimately satisfactory level of granularity. Each description defines its level in an arbitrary way.

In a lexicon-grammar, every distinction with any reproducible property is formalized unless the distinction is merely a matter of syntactic transformation. It also should be strictly correlated with at least one

reproducibly observable property. The concrete “crush” sense of *écraser* is compatible with a prepositional complement that denotes the resulting state of the crushed object:

> *Il a écrasé l’ail en purée.*
> He crushed the garlic into paste.

The concrete “squeeze” sense is not:

> *Tu m’écrases le pied en N.*
> You are stepping on my foot into pulp, you fat fool.

Once this property has been encoded, it supports the separation, even though it was initially suggested only by intuition. If the description were more fine-grained, it would represent semantic distinctions not supported by reproducible observation; if it were more coarse-grained, it would erroneously assign the formal property to the other sense.[47]

The study of grammatical properties identifies and lists properties for which reliable systematic encoding is possible. Lexicon-Grammars are based on two lists: lexical entries and grammatical properties. As such, this model is a simplified view of linguistic reality, but it makes possible cross-tables that combine entries with properties. The tabular layout is natural, clear and readable for descriptive work. Once tables are ready, they can be

translated into other formats for computer processing. Specialists in automatic syntactic parsing use formats where each lexical entry is represented by a formula that explicitly states its positive or negative properties. The constructions in which the entry does not appear are left out. The only alternative to such fine-grained encoding is the use of generalization-based rules, which are more compact than tables. Many computational linguists are more familiar with rules than with tables. Tables are better. If the rules are checked before use, this requires fine-grained encoded resources such as tables. If they are not, they are only approximations and may produce the wrong results.

Many of my examples spotlight phenomena that belong to syntax and to the lexicon. By definition, Lexicon-Grammar studies their intersection. Linguists have been aware of this intersection at least since Edward Sapir's observation that all grammars leak.[48] Consider "book"/"books," "ox"/"oxen," "sheep"/"sheep," and "goose"/"geese." No set of grammatical rules encompasses this degree of irregularity. Lexicon-grammars have shown that the grey zone between the syntax of a language and its lexicon is enormous.

Understanding syntax and semantics often requires taking the lexicon into account. Take the following problem. In order to formalize the meaning of "John has a flu" and "John has a wart," which formal structure should be adopted? *Have(John, flu)* and *Have(John, wart)*? In logical terms, "have" is functioning in these sentences as one two-place predicate: *H(x,y)*. There

are two arguments in “John” and “flu” (or “wart”), but only one predicate in “have.”

Or is it better to represent the logical structure of “John has a flu” in terms of a one-place predicate? *Flu(John)* or *Wart(John)*? Jacques Labelle makes the interesting point that some disease nouns have a second argument (“John has a wart on his hand”); others (such as flu) do not.[49] Predicate structures are markedly different; and so are ancillary logical structures. John has a flu, if expressed as *Have(John, flu)* implies that there is something that John has, but expressed as *Flu(John)*, it implies only that John is sick as a dog. Labelle noticed the difference between these nouns *only* when he listed them and registered their properties.

Conversely, lexicology involves syntax. Differences in syntax are immensely useful in separating senses, as in the example of *écraser*, which can mean “crush” or “squeeze.” The intersection between syntax and the lexicon is particularly relevant to syntactic analysis and language technology. Chomsky’s assertion that the more the lexicon is studied, the less syntax is, and vice-versa, has persuaded most generative grammarians and some linguists that it is one thing or the other—the lexicon or syntax.[50]

The widespread impression among linguists that syntactical studies are somehow more scientific than lexical studies has deterred them from studying the lexicon. Syntax is where the theories are, and where would we be without the theories? “Picking up shells on the beach,” some

linguists scoff.[51] *That* is where we would be.

Didn't Hubble discover the galaxies after patiently observing individual stars?

The Armchair Linguist

THE ARMCHAIR linguist has been a staple of controversy for more than fifty years. It is a long time for anyone to have remained seated. The story is worth recounting. Noam Chomsky, whether sitting or standing, has been the dominant figure among generative grammarians for as long as there has been anything like generative linguistics. Chomsky has little use for corpora, and even less use for the formal procedures developed by American structural linguistics. Generative grammar limits itself to facts that "reflect a regular grammatical process of the language."[52] Everything else is assigned to the outer darkness of the lexicon.[53] Chaos is so much the standard in *any* language that large portions of the lexicon are simply excluded from generative studies. It goes without saying that the success of these ideas among linguists compromised the quality of their empirical data. Disagreement began in the 1970s, and became exasperation, whereupon Charles Fillmore's armchair linguist made his appearance, a linguist careless in observation, and indifferent to complexity.[54] Call it the counterrevolution; whether counter or not, the ensuing movement served

to rehabilitate corpus linguistics. But corpus linguists did little to rehabilitate either the formal procedures of empirical observation or the systematic studies of a lexicon.

Gross stressed formal procedures of empirical observation and systematic lexical studies. Generative grammar rejects, and corpus linguistics overlooks, both.

C'est dommage.

1. Sylviane Granger and Marie-Aude Lefer, "Enriching the Phraseological Coverage of High-Frequency Adverbs in English–French Bilingual Dictionaries," in *Advances in Corpus-Based Contrastive Linguistics: Studies in Honour of Stig Johansson*, eds. Karin Aijmer and Bengt Altenberg (Amsterdam: John Benjamins, 2013), 164. ↩

2. Lélia Picabia, *Les constructions adjectivales en français (Adjectival Constructions in French)* (Geneva: Droz, 1978), 114. ↩

3. Lélia Picabia, *Les constructions adjectivales en français (Adjectival Constructions in French)* (Geneva: Droz, 1978), 107. ↩

4. Catherine Léger, "Sentential Complementation of Adjectives in French," in eds. Patricia Cabredo Hofherr and Ora Matushansky, *Adjectives: Formal Analysis in Syntax and Semantics* (Amsterdam: John Benjamins, 2010), 285. ↩

5. Catherine Léger, “Sentential Complementation of Adjectives in French,” in eds. Patricia Cabredo Hofherr and Ora Matushansky, *Adjectives: Formal Analysis in Syntax and Semantics* (Amsterdam: John Benjamins, 2010), 286. ↩

6. Leonard Bloomfield, *Language* (New York: Holt, Rinehart, and Winston, 1933), 205. ↩

7. See Jean-Paul Boons, Alain Guillet, and Christian Leclère, *La structure des phrases simples en français: constructions intransitives (The Structure of Simple Sentences in French: Intransitive Constructions)* (Geneva: Droz, 1976), 378; Alain Guillet and Christian Leclère, *La structure des phrases simples en français: Les constructions transitives locatives (The Structure of Simple Sentences in French: Transitive Locative Constructions)* (Geneva: Droz, 1992), 446. ↩

8. Operational procedures based on distributional analysis were also adopted by Naomi Sager’s Linguistic String Project (LSP), which lasted from 1960 to 2005. ↩

9. Nick Enfield, “Review of the Book Constructions at Work: The Nature of Generalization in Language by Adele E. Goldberg,” *Linguistic Typology* 12 (2008): 157. Even when they use these methods, linguists adopting them have little use for, and rarely acknowledge, structural and distributional linguistics. ↩

10. “When an operational procedure is proposed, it must be tested for adequacy … by measuring it against the standard provided by the tacit knowledge that it attempts to specify and describe.” Noam Chomsky, *Aspects of the Theory of Syntax* (Cambridge, MA: MIT Press, 1965), 19. ↩

11. Other criteria are possible. It would be useful to develop a single criterion by which “Man comes home and finds house on fire with dog inside” is understood in a way that makes “man,” “house,” and “dog” uncountable. ↩

12. Catherine Léger, “Sentential Complementation of Adjectives in French,” in eds. Patricia Cabredo Hofherr and Ora Matushansky, *Adjectives: Formal Analysis in Syntax and*

Semantics (Amsterdam: John Benjamins, 2010), 285–86. ↩

13. See Paola Merlo and Suzanne Stevenson, “Establishing the Upper Bound and Inter-Judge Agreement of a Verb Classification Task,” in Proceedings of LREC, (European Language Resources Association (ELRA), 2000): 1,659–64. ↩

14. As Joan Bresnan said, for example:

    > [In the 2000s] I began to realize that we theoretical linguists had no privileged way of distinguishing the possible formal patterns of a language from the merely probable. Many of the kinds of sentences reported by theorists to be ungrammatical are actually used quite grammatically in rare contexts. Authentic examples can be found in very large collections of language use, such as the World Wide Web.

    Joan Bresnan, “A Voyage into Uncertainty,” in *Thinking Reed: Centennial Essays by Graduates of Reed College*, eds. Roger Porter and Robert Reynolds (Portland, OR: Reed College, 2011), 74. ↩

15. Christian Lehmann, “Data in Linguistics,” *The Linguistic Review* 21, no. 3–4 (2004): 294–95. ↩

16. Maurice Gross, “Methods and Tactics in the Construction of a Lexicon-Grammar,” in *Linguistics in the Morning Calm 2: Selected Papers from SICOL 1986* (Seoul: Hanshin, 1988), 177–197. ↩

17. Elsa Tolone and Benoît Sagot, “Using Lexical-Grammar Tables for French Verbs in a Large-Coverage Parser,” *Human Language Technology. Challenges for Computer Science and Linguistics: Lecture Notes in Computer Science* 6,562 (2011): 183. ↩

18. For an example, see the following table:

    > Figure 1. *Sample of verb class 33*

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: les figé	Nég	<ENT>	N0 V	N0 être V -ant	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	N0hum V W sur ce point	[extrap]	<OPT>
+	–	–	<E>	–	–	–	–	renaître	+	+	–	+	–	–	+	–	–	Max renaît au bonheur de vivre
+	–	–	se	+	–	–	–	rendre	+	–	+	+	+	–	+	+	+	Max s'est rendu à mon opinion
+	–	–	Se	+	–	–	–	rendre	+	–	+	–	–	–	–	–	–	Le caporal s'est rendu à l'ennemi
+	–	–	<E>	–	–	–	–	renoncer	–	–	+	+	–	–	+	–	–	Max renonce à son héritage

The Lexicon-Grammar methodology consists in establishing a taxonomy of syntactic-semantic classes whose lexical items share some syntactic features. For instance, class 33 contains verbs that enter the construction with one indirect complement introduced by preposition *à*. Each class is represented by a table that includes all lexical items of the class. If a verb has two meanings, it is divided into two lexical items. In the verb class 33 … *se rendre* has two meanings and therefore two lexical items:

1. *Jean s'est rendu à mon opinion* (John finally accepted my opinion)
2. *Vercingetorix s'est rendu à César* (Vercingetorix surrendered to Caesar)

Each feature of the table of classes is associated with a set of operations that combine linguistic objects together; for instance, when feature N0 =: Nhum is true for a given entry, an object defining a human noun phrase is added to the distribution of N0 (i.e., the argument 0 of the predicate). If the feature is assigned true for a given lexical entry, the associated operations are activated. Ppv stands for positive predictive value.

A selection of features is applied to all entries and their linguistic validity is

> checked. At the intersection of a row corresponding to a lexical item and a column corresponding to a feature, the cell is set to '+' if it is valid or '-' if is not. For instance, one meaning of se rendre (to accept) accepts a non human nominal complement in its canonical sentence: its feature N1 =: N-hum value is true ('+') while it is false ('-') for the other (to surrender).

Matthieu Constant and Elsa Tolone, "A Generic Tool to Generate a Lexicon for NLP from Lexicon-Grammar Tables," in *Actes du "27e Colloque international sur le lexique et la grammaire" (L'Aquila, 10-13 septembre 2008). Seconde partie,* ed. Michele De Gioia, (Roma: Aracne, 2010), 79–93. ↩

19. 61 tables for simple verbs have been developed for the French language, as well as 59 tables for predicative nouns, 65 tables for idiomatic expressions (mostly verbal), and 32 tables for (simple and idiomatic) adverbs. See Elsa Tolone and Benoît Sagot, "Using Lexical-Grammar Tables for French Verbs in a Large-Coverage Parser," *Human Language Technology. Challenges for Computer Science and Linguistics: Lecture Notes in Computer Science* 6,562 (2011): 183. ↩

20. Walter Bisang, "Variation and Reproducibility in Linguistics," in *Linguistic Universals and Language Variation*, ed. Peter Siemund, (Berlin/New York: Walter de Gruyter, 2011), 251. ↩

21. Walter Bisang, "Variation and Reproducibility in Linguistics," in *Linguistic Universals and Language Variation*, ed. Peter Siemund, (Berlin/New York: Walter de Gruyter, 2011), 254. ↩

22. Georgia Green, *Semantics and Syntactic Regularity* (London: Cambridge University Press, 1973). ↩

23. Adriana Belletti and Luigi Rizzi, "Psych-Verbs and θ-Theory," *Natural Language and Linguistic Theory* 6, no. 3 (1988): 291–352. ↩

24. Ellen Gurman Bard, Dan Robertson, and Antonella Sorace, "Magnitude Estimation of

Linguistic Acceptability," *Language* 72, no. 1 (1996): 32–68. ↩

25. Noam Chomsky, *Syntactic Structures* (The Hague/Paris: Mouton, 1957), 15. ↩

26. Noam Chomsky, *Syntactic Structures* (The Hague/Paris: Mouton, 1957), 16, 35–36. ↩

27. Carson Schütze, *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology* (Chicago: University of Chicago Press, 1996). Schütze provides a detailed discussion of how inconsistently Chomsky has used the terms grammaticality and acceptability. ↩

28. Thomas Ernst, "Speaker-Oriented Adverbs," *Natural Language & Linguistic Theory* 27, no. 3 (2009): 497–544. ↩

29. Maurice Gross wrote:

    > When we compare meanings as we did between *La décoratrice enjolive la vitrine de minijupes claires* (The decorator ornaments the shop window with bright miniskirts), and *La décoratrice enjolive la vitrine de sa minijupe claire* (The decorator ornaments the shop window with her bright miniskirt), or between the two interpretations of *Pierre amuse Paul* (Peter entertains Paul; Paul finds Peter funny), we will say we proceed to differential assessment of meaning. The distinction has a parallel in measurement of physical quantities. For instance, the absolute measurement of a temperature or weight is relatively coarse, whereas differential measurement of the same variables is quite sharp.

    Maurice Gross, *Méthodes en syntaxe: régime des constructions complétives (Syntax Methods: The System of Complements Constructions)* (Paris: Hermann, 1975): 32. ↩

30. Maurice Gross, "On the Failure of Generative Grammar," *Language* 55, no. 4 (1979): 868. ↩

31. Éric Laporte, Elisabete Ranchhod, and Anastasia Yannacopoulou, “Syntactic Variation of Support Verb Constructions,” *Lingvisticae Investigationes* 31, no. 2 (2008): 173–85. ↩

32. Most of the early lexicon-grammar studies were in French, but this was still an international language for linguistics in the 1970s. ↩

33. Noam Chomsky, *Syntactic Structures* (The Hague/Paris: Mouton, 1957), 15. ↩

34. Geoffrey Sampson, “What was Transformational Grammar? A Review of: Noam Chomsky, The Logical Structure of Linguistic Theory,” *Lingua* 48 (1979): 370. ↩

35. Christian Lehmann, “Data in Linguistics,” *The Linguistic Review* 21, no. 3–4 (2004): 294–95. ↩

36. Maurice Gross, *Méthodes en syntaxe: régime des constructions complétives (Syntax Methods: The System of Complements Constructions)*, (Paris: Hermann, 1975). ↩

37. For English examples, see Morris Salkoff, “Bees are Swarming in the Garden: A Systematic Synchronic Study of Productivity,” *Language* 59, no. 2 (1983): 288–346; Morris Salkoff, “Verbs with a Sentential Subject. A Lexical Examination of a Sub-Set of Psych Verbs,” *Lingvisticae Investigationes* 25, no. 1 (2002): 97–147. ↩

38. William Labov, “Some Observations on the Foundation of Linguistics,” (1987). ↩

39. Stephan Gries, “Methodological and Interdisciplinary Stance in Corpus Linguistics,” in *Perspectives on Corpus Linguistics*, eds. Vander Viana, Sonia Zyngier, and Geoff Barnbrook (Amsterdam: John Benjamins, 2011): 87. ↩

40. Steven Abney, “Data-Intensive Experimental Linguistics,” *Linguistic Issues in Language Technology* 6, no. 2 (2011): 9. ↩

41. For a definition, see *Wikipedia*, “Lemma (Morphology)”:

> In morphology and lexicography, a lemma (plural *lemmas* or *lemmata*) is the

> canonical form, dictionary form, or citation form of a set of words (headword). In English, for example, *run*, *runs*, *ran* and *running* are forms of the same lexeme, with *run* as the lemma. *Lexeme*, in this context, refers to the set of all the forms that have the same meaning, and *lemma* refers to the particular form that is chosen by convention to represent the lexeme. In lexicography, this unit is usually also the *citation form* or headword by which it is indexed. … Lemmas or word stems are used often in corpus linguistics for determining word frequency. In such usage the specific definition of "lemma" is flexible depending on the task it is being used for.

↩

42. In reality, not even Balota et al. do that in their objectivity-aware work on the lexicon of English for experimental psycho- and neurolinguistics. See David Balota, Melvin Yap, Michael Cortese, Keith Hutchison, Brett Kessler, Bjorn Loftis, James Neely, Douglas Nelson, Greg Simpson, and Rebecca Treiman, "The English Lexicon Project," *Behavior Research Methods* 39, no. 3 (2007): 445–59. Their dictionary does provide parts of speech of the words, and they do not report the source of the information, but they do not claim that this source is free of subjectivity. They probably trusted grammarians or lexicographers. ↩

43. Morris Salkoff, "Verbs with a Sentential Subject. A Lexical Examination of a Sub-set of Psych Verbs," *Lingvisticae Investigationes* 25, no. 1 (2002): 97–147. ↩

44. Many computational linguists hope that statistical and probabilistic models might be able to compensate for deficient dictionaries and grammars in NLP systems. See Steven Abney, "Data-Intensive Experimental Linguistics," *Linguistic Issues in Language Technology* 6, no. 2 (2011): 11–12. Such models do help in some respects, but there is no proof that they are able automatically to deal with the meanings and formal variations in all relevant NLP systems. ↩

45. According to Paul Hopper, confrontation with linguistic reality reveals the existence of

unexpected cases of low consistency. However, his methodological contribution to confronting such aspects of language does not go beyond recommending a psychological attitude:

> [Y]ou accept and incorporate into your working method, your theory, the assumption that language is not especially uniform, and you make the ragged, messy nature of your data and their unpredictable margins visible and public … we should be prepared to question [consistency] when the competing demands of accuracy and integrity militate against it.

Paul Hopper, "The Ideal of Consistency in Thinking about Language," *Southwest Journal of Linguistics* 19, no. 1 (2000): 2. ↩

46. "In French, for instance, Maurice Gross indexed 26,000 verbal idioms and 12,000 adverbial idioms." Éric Laporte, "In Memoriam Maurice Gross." (Paper presented at the Language and Technology Conference (LTC), Poznan, Poland, 2005): 3. ↩

47. Éric Laporte, "Defining a Verb Taxonomy by a Decision Tree," in *Autour des verbes: Constructions et interprétations (Around Verbs: Constructions and Interpretations)*, ed. Kozué Ogata, (Amsterdam: John Benjamins, 2013), 102. ↩

48. Edward Sapir, *Language: An Introduction to the Study of Speech* (New York: Harcourt Brace, 1921), 38. ↩

49. Jacques Labelle, "Grammaire des noms de maladie (The Grammar of Disease Names)," in *Langue Française* 69 (Paris: Larousse, 1986), 109. ↩

50. Noam Chomsky, "Remarks on Nominalization," in *Readings in English Transformational Grammar*, eds. Roderick Jacobs and Peter Rosenbaum (Waltham, MA: Ginn, 1970), 185. ↩

51. Brian Silver, *The Ascent of Science*, (Oxford: Oxford University Press, 1998), 16. ↩

52. John J. McCarthy and Alan S. Prince, "Foot and Word in Prosodic Morphology: the Arabic

Broken Plural," *Natural Language and Linguistic Theory* 8, no. 2 (1990): 267. ↩

53. For Chomsky, the lexicon is made of language-specific idiosyncrasies that need not be integrated into grammatical theories, which are only designed to account for core phenomena. Noam Chomsky, *The Minimalist Program* (Cambridge, MA: MIT Press, 1995). ↩

54. Charles Fillmore, "Corpus Linguistics or Computer-aided Armchair Linguistics," in *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991*, ed. Jan Svartvik (Berlin: Mouton de Gruyter, 1992), 35. ↩