Concentric: Studies in Linguistics
30.1 (June 2004): 1-18

Corpus and Grammar: What It Isn't*

Chris C-C Shei
Centre for Applied Language Studies
University of Wales Swansea

Grammar means different things to different people. For generative grammarians, grammar is innate, autonomous and universal. For functionalists, grammar is just one of the many devices which humans employ to communicate their ideas in a social setting. Psycholinguistic studies, neurolinguistic methodologies and the science of biology, may provide the ultimate answer to whether there is an innate core of grammar which can indeed be clearly separated from the influence of performance factors, as generative grammarians claim. Meanwhile, corpus investigation is still a more convincing way to describe and possibly explain how language is actually used by human beings. Many dictionaries have been compiled from the exploration of corpora, many language structures studied and rules of use described. In this paper, a selective and evaluative review of research on generative grammar is offered, mainly from the psycholinguistic point of view. The main claim is that an autonomous syntactic module in the human mind as proposed by generative grammarians is unsustainable, and it is misleading to separate grammar from context in language studies. Although the innate hypothesis can be supported, grammar is constantly shaped by culture and interpersonal interactions. In the end, the place of grammar may not be so pivotal to human language as held by generative grammarians (i.e. the lexicon may be more important in language processing than rules governing sentence formation). A more viable means of approaching the truth of language may be through the investigation of corpora.

Key words: generative grammar, psycholinguistics, aphasia, neuroimaging, corpus linguistics, extended lexical unit

1. Introduction

Grammar is an abstract term, although the degree and kind of abstractness varies for different individuals. For laypeople and language learners, grammar most likely means rules concerning the use of language. For linguistics students, grammar possibly means knowledge of language which can be explored from different perspectives such as phonetics, syntax, semantics, or pragmatics. To the modern theoretical linguists who follow the Chomskyan tradition, grammar is a very abstract and self-contained system residing in the human's genetic codes. For many contemporary grammar book writers, grammar is a set of rules extractable from the actual use of language (i.e. language corpora). When language teachers think of grammar, they will view the subject from a pedagogical point of view, which is more related to corpus linguistics than to theoretical linguistics. For language engineers

* I am very grateful for the two anonymous referees' helpful comments and the English proofreader's very useful corrections and suggestions.

working with machine translation for example, grammar means a set of deterministic rules or probabilistic formulae for computing language data to comprehend or generate natural language. For systemic-functional linguists following the Hallidayan tradition, grammar makes sense only in the context of social functions, while cognitive grammarians, in a somewhat similar vein, are interested in looking at how our minds interact with the environment to produce symbolic links between phonological and semantic structures (Taylor 2002: chapter 2). In sum, grammar means different things to different people, not only to the laymen, but to the academic professionals. As the aforementioned examples show, grammar can be viewed from a number of perspectives. Therefore, when talking about grammar, it is necessary for an individual to identify himself/herself to the subject.

The main purpose of this article is to consider the possibility of investigating grammar from the point of view of empirical data – corpora. However, the emphasis will be on the examination of the so-called mainstream approach to grammar – Chomsky's theories from the psycholinguistic point of view. Thus we will first review generative grammarians' claims of the innateness of language and syntactic autonomy. This will lead to a discussion of neuropsychological evidence to support or refute their views, as well as a discussion from the biological perspective. Finally, corpus linguistics will be introduced and the relationship between corpus-based research and grammar discussed.

2. Generative grammar

Chomsky's Syntactic Structures was published over 40 years ago, and since that time, generative grammar has dominated research in the field of theoretical linguistics. To date, many people working in this tradition still consider themselves representing mainstream syntactic theories (see Borsley 2002 as a recent example of this "mainstream" view).

The primary claims of Chomskyans which I am concerned with in this article are:

- Language is innate.
- Syntax is autonomous.
- Competence can be separated from performance.

(For references of the above, see, for example, Chomsky 1965, Chomsky 1986, Newmeyer 1997, or Radford 1997). That language is to some extent innate is less controversial. The most controversial and, for some, disturbing claims are the second and the third, which led to, respectively, an explosion of research in psycholinguistics

and neurolinguistics to try to confirm or refute the modular view of syntax, and a repercussion from empiricists represented by researchers in the tradition of discourse analysis, functionalism, and later on, corpus linguistics, who generally disagree that grammar can be "taken away" and studied in isolation ("isolation" refers both to grammar being isolated from other components of language and to the linguist being self-isolated from the general language users).

Based on the above three assumptions, the Chomskyans generated a large body of abstract syntactic theories ranging from the initial transformational grammar and the "standard theory" to the government and binding theory, and more recently the minimalist program (see for example Bierwisch 2001 for a concise historical introduction). The related claim of "universal grammar" (UG), and the "parameter-setting" theory, also generated a large body of research in second language acquisition, as researchers tried to prove the existence of UG across languages and their different parameter settings (see Cook 1988 for an introduction to UG).

It is interesting to observe the change of Chomskyan claims about syntactic structures from the initial very complicated system of multi-level representations and inter-level transformations to the recent idea of a minimal core of language. Although universal grammar is the basic rationale of Chomskyan syntactic theories, there is no conclusive support for these theories in languages other than English, where Chomskyan theories originated. Therefore, a more feasible approach is to make more general claims about language (Bates 2003). In a way, this lessens the role played by syntax in language processing and implicitly supports a more "distributed" view of language and a more isomorphic view of language to other forms of cognition. In linguistic theories, this could mean that performance factors are more connected to competence than Chomsky originally thought. In neurolinguistics, this complies with a number of recent studies showing that syntactic processing cannot be modularised to a single area in the brain (Moro et al. 2001, Kaan & Swaab 2002). Aitchison (1998: 109) made the same observation when she said that "Chomsky's increasingly broad and general claims about language bring him closer to people he disagrees with, …." In the next section, I will discuss how psycholinguistic research has failed to support the Chomskyan view of the autonomy of syntax.

3. Psycholinguistics

Since Broca's seminal contribution to aphasic studies, aphasia has become one of the two main instruments for investigating the localisation issue of the brain (i.e. whether a particular region in the brain is uniquely connected to a language module or function). The other main instruments used in studies that aim to relate region of the

brain to function are the evolving neuroimaging methodologies, for example, functional magnetic resonance imaging (fMRI). The advantage to neuroimaging is that it can be employed with both health individuals and those experiencing medical problems.

Initially, the now classical model of Wernicke-Lichtheim in the 1880's seemed to offer encouraging thoughts regarding the mapping of areas of brain lesion to types of aphasia (which implied a relationship between cortical regions and language modules or functions). Figure 1 provides an easy-to-understand summary of the classical findings that have been compiled from various sources (see, for example, Dingwall 1998 for more information):

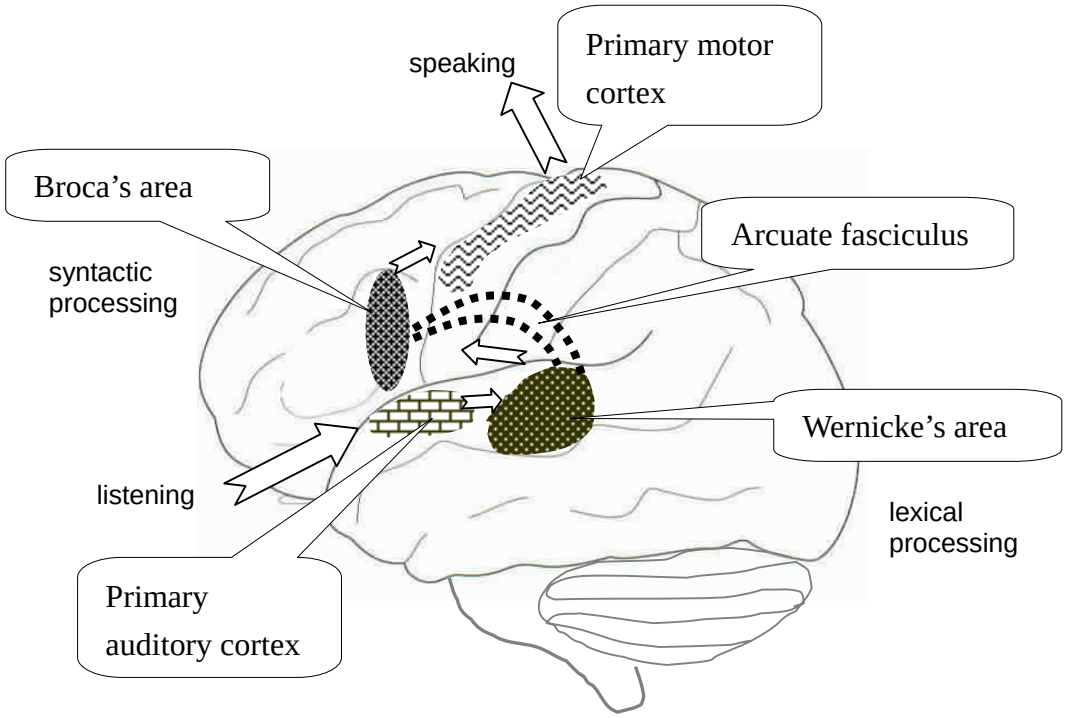

Figure 1: Classical model of language processing derived from aphasia studies

The two most important parts of the brain marked in Figure 1 are: Broca's area, which was strongly associated with syntactic processing because Broca's aphasics typically suffered from poor sentential structure (especially telegraphic speech devoid of function words) although meaning was relatively intact, and Wernicke's area, which was more associated with semantics since Wernicke's aphasics produced relatively well-structured utterances which were, nevertheless, mostly meaningless. According to the Wernicke-Lichtheim's model, distinct types of aphasia will result from injury to different parts of the brain identified in the model.

Recent experimental findings suggest, however, that patients suffering from

so-called Broca's aphasia do not necessarily have lesions in Broca's area. Conversely, injury to Broca's area does not necessarily lead to symptoms associated with Broca's aphasia. (Dingwall 1998: 92-94). The close association between language modules and cortical regions can thus no longer be supported.

Syntactic modularity is important for generative grammarians, because their entire research methodology depends on it. They have to justify the preclusion of any other language modules by establishing that syntax is autonomous and oblivious to other forces of influence. Their theories are built within a vacuum called syntax. Once syntactic autonomy is in question, any existing research results will be undermined (since it is no longer justifiable to study syntax in a highly isolated manner). Psycholinguistics is one area where Chomskyans hope to obtain empirical evidence for their autonomous view of syntax.

Grodzinsky (2000) suggested that Broca's area is where the Chomskyan syntactic processor resides. However, many experiments generated from this and similar claims, for example, Müller et al. (2003), found that Broca's area was activated not by syntactic processing alone, but also by lexical-semantic processing. (Conversely, as previously mentioned, the processing of sentences involves not only Broca's area, but also other areas in the brain, even areas which were largely ignored by the classic aphasia-syndrome model, i.e. the cerebellum and various nerve tissues in the brain stem – see Moro et al. 2001)

Newmeyer (1997) suggested that genetic dysphasia is good evidence for pointing "incontrovertibly to a genetic basis for autonomous grammar" on the ground that genetic dysphasia is a specific language impairment (SLI) which is "by definition, a language disorder unaccompanied by non-linguistic deficits." However, equating dedicated language malfunction to the existence of an autonomous syntactic faculty may not be entirely appropriate. There may be other reasons causing the language deficiency. Bates (2003), for example, citing Bishop (1997) and Leonard (1997), refers to a theory of SLI being caused by a general deficit in auditory processing rather than the damage to a mysterious "mental organ" of language. Other researchers also pointed out that SLI patients are in fact "mentally subnormal" (see Cowley 2001). Arguing for syntactical autonomy on the basis of SLI is unsustainable.

It would seem then, that psycholinguistics does not support an autonomous view of syntactic processing as strong as that claimed by Chomsky and his followers. What about the science of biology which Chomskyans often refer to in their discussions of innateness and UG theories?

4. Biological basis

The success of generative grammar was such that, in the 1960's and perhaps 1970's, researchers from other fields (mathematics, computer science, biology, engineering…) noticed and started trying out the idea of "generative" in their separate fields. To date, some interdisciplinary researchers still seize Chomsky's main theses as the starting point in their language-related research, presumably because Chomskyan theories are "neat" (as opposed to the more "untidy" and "unprincipled" theories of grammar offered by other approaches) and easy to integrate into their research methodologies. If Chomsky's theories are undermined, then many theories built upon them in other fields may also be challenged.

Both Nowak & Komarova (2001) and Page (in press) used mathematical models to simulate children's first language learning under the framework of UG. Both studies developed sophisticated mathematical models from either the language evolution or language acquisition point of view. Although the researchers found nothing wrong with UG suppositions mathematically, they came to a similar conclusion regarding the variation of UG. While Nowak & Komarova suspected that UG is "only very roughly defined by our genes" and can vary to some extent among individuals, Page mentioned the possibility of "much less structured UGs" simulated by neural network algorithms, which can enable children to successfully learn a language if the input is simple, like mothers' language.

According to Jerne (1985: 1059), quoted by Lorenzo & Longa (2003), the "hypothesis of an inheritable capability to learn any language means that it must somehow be encoded in the DNA of our chromosomes." Different attempts have been made to examine the relationship between human genes and language with some interesting results. Tsonis et al. (1997), for example, used Zipf's law, referring to Zipf (1949), to calculate whether there are structures in DNA similar to human languages and concluded that "DNA sequences show no linguistic properties." On the other hand, Ji (1997) introduced a very interesting model, where "cell language" was identified and proposed to be similar to human languages. For example, cell language uses molecules as signs and the messages to be conveyed are "gene-directed cell processes" like replication and translation (the process of forming a protein molecule). Ji wisely observed that "human language can be said to consist of two components – cultural and natural" and went on to propose a biological model of human language to account for the inheritance of language behaviours through the encoding of DNA. In this model, Ji proposed that the natural component of human language is encoded in DNA, which is in turn translated into texts of the cell language, which are used to develop what Ji called the "language-enabling brain structure"

(LEBS). Upon interacting with culture, the LEBS then gives rise to Chomsky's UG, which further interacts with culture to yield linguistic performance. Figure 2 provides an illustrative version of Ji's model.

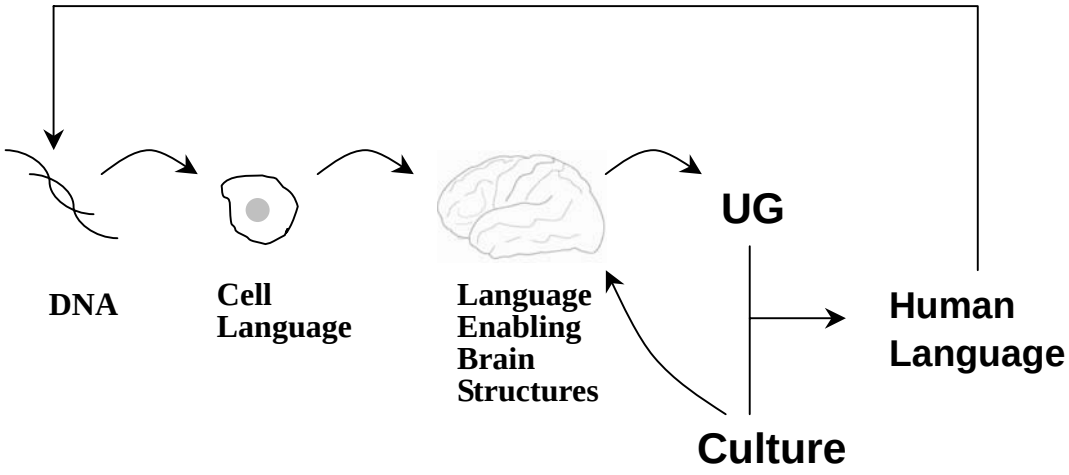

Figure 2: Biological model of human language (adapted from Ji 1997)

Disregarding the form of the UG Ji had in mind in his model, Ji had the wisdom to include the influence of culture in the model.

The importance of culture (or environmental factors) to the development of language has been repeatedly emphasised by non-Chomskyan researchers. Snow (1996), for example, pointed out: "For language as for other similarly complex domains of psychological functioning, both a well-designed brain and a well-designed environment are prerequisite to normal development" (p. 377). She cautioned against the unjustifiable attempt to identify biology with heredity (i.e. genes) and went on to cite researches which demonstrated the "postnatal, experiential influences on brain structure" (p. 293). In other words, even biological structures, not logical structures alone, can be changed by our environment (this is easy to see – consider the development of cancer). To insist on a biological core lurking in the genes, controlling the vital part of language processing and unaffected by any environmental factors while doing so, is like insisting on a cube sugar being undissolvable in water.

Bates (2003) was not against the idea of language being innate. What ought to be questioned is the form of innateness. She illustrated her point with the evolution of the giraffe's long neck. The elongation of the neck itself could not have happened without bringing along quantitative changes in other parts of the giraffe's body, like the legs or the lungs, which must have changed in proportion in order to support this quantitatively changed part of the anatomy. Note "quantitative changes" are

emphasised here to show that the evolution was only a new adaptation of the old anatomy, not a "qualitative" change whereby any new part was generated by the biological system. Likewise, there is not a "syntactic organ" in the human body which grew out of nowhere and, once having appeared, encapsulated itself and became "autonomous". Syntax is no different from other language components like phonetics or semantics, which are so distinguished only to facilitate scientific investigation in the first place. Language itself, in turn, is not so different from other cognitive abilities of human beings. As human language evolved to become different from monkey languages for example, various language participating structures of the body must have changed and contributed jointly to the new function. A distributional view is more likely than a modular one. Language exists because our environment so demands. In other words, language evolves because social organisations and cultural activities put pressure on human beings (somewhat like leaves on high trees to giraffes) and this should be considered in the study of language, no matter which perspective language is viewed from.

5. Corpus linguistics

Corpus linguistics is the opposite of generative grammar. The former starts from *performance* data (i.e. collecting language actually used by the population); while the latter starts from the retrospection of *competence* (i.e. contriving sentences in an "undisturbed" manner). Whereas corpus linguistics investigates language produced under various social circumstances; generative grammarians retrieve sentences from the "autonomous" syntactic module which was individually prepared and readily present at birth. An important difference was observed by Meyer (2002) that corpus linguists mainly do the *describing* of language, while generative grammarians mainly do the *explaining*. Unfortunately, the two approaches do not converge and compensate for each other. As Meyer said, "corpus linguists are very sceptical of the highly abstract and decontextualized discussions of language promoted by generative grammarians…" (p. 3). Generative grammarians, such as Borsley and Newmeyer, on the other hand, criticized functional linguistics (which subsumes the corpus linguistics), saying things like: "At every stage in the development of the field, there has been a significant minority attempting to reassert the 'common sense' view of language" (Newmeyer 1997: 49)[1].

Most existing research on corpus linguistics concentrates on lexicography and phraseology. Sophisticated statistical models have also been developed based on

[1] By using the term 'common sense', Newmeyer supposedly meant to degrade performance related studies to the level of folk thinking, not serious, scholarly research.

language corpora to account for the constructs of sentences or entire texts. The results of corpus-based research not only added to our knowledge about language, but also contributed significantly to academic and social activities, for example in the areas of literature, history, legal and business activities, language learning and language engineering.

Roughly speaking, there are two kinds of approaches to corpus-based research. One is to use concordance lines and collocational data to analyse a body of language (e.g. Stubbs 2002). The other is to use mathematical formulae to build statistical models from corpora (e.g. Charniak 1993). The former often focuses on establishing subtheories of (the use of) language; the latter often contemplates the application of mathematical models to practical use. Concordance-based investigation into corpora has produced a significant amount of literature on the behaviours of words, phraseology, semantics and aspects of surface grammar. The results have been applied fruitfully to compilation of dictionaries, language teaching, translation, literature and socio-cultural studies, and so on. Statistical modelling of corpora, on the other hand, contributed to natural language processing including speech recognition, language generation, machine translation, and so on.

So far, most contributions made by corpus linguistics to the understanding of language are in the areas of lexis and phrases, although grammar is also covered in much corpus-based research. The grammars covered in corpus-based investigations are often concrete and specific and by no means intended to be all-encompassing, or "universal." Little attempt has been made to explore the "core" of grammar of a highly abstract nature. Unlike generative grammarians who use more of a deductive approach (they, for example, contrived UG first and then went on to examine whether sentences follow UG principles), corpus linguists use an inductive approach and start from observations of language in use. Thus the observed rules in corpus linguistics tend to be locally relevant and less pompous looking. Meyer (2002: 12) thus described, for example,

> a very common use of corpora: to provide a detailed study of a particular grammatical construction that yields linguistic information on the construction, such as the various forms it has, its overall frequency, the particular contexts in which it occurs… and its communicative potential.

Research into grammar based on corpora is also possible with a view to yielding more generalisable result. One recent study by Barnbrook & Sinclair (2001), for example, looked at "definition sentences" (e.g. *A cat is an animal that chases rats*) in

a dictionary corpus and inferred what they called a "local grammar."[2] According to them, a parser based on this grammar was constructed, which could not only parse every definition sentence in the corpus, but "will also interpret any such sentences occurring in the language at large" (p. 249). An important observation made by the authors was that, because local grammars like this have a clear communicative function in mind, their parsing performance will be higher than a generally constructed parser. Presumably, if there is a battery of these grammars and parsers, then ordinary texts can be satisfactorily analysed based on these functional terms.

One recurrent theme in current linguistics research is that grammar and lexis are very difficult to separate. Hunston (2002: 150), for example, claimed: "The point about lexical and grammatical facts being separable has been countered by the notion of 'pattern grammar'." For Hunston, grammar can be interpreted as a kind of "pattern flow" where one pattern leads to or embeds in another. Thus, to some extent, phraseology can be "used as a description of what happens in English, in lieu of, say, a grammatical rule…" (p. 152). This is not to say that there is only phraseology and no grammar in human language, but that the role of grammar may not be so important as some think. The other way of saying this is to put the classical distinction between grammar and words at risk and say something like "there is grammar in words," or "words subsumes more grammar than we thought."

Finally, a more synthesised view of grammar and diction can be found in Francis (1993), who explained a "corpus-driven approach to grammar." As a result of detailed analyses of large corpora, a kind of cross-referencing machinery is generated where all lexical items are marked with the grammatical structures they frequently appear in, and all grammatical structures identified are marked with the lexis and phrases they prefer. Altogether, this constitutes a comprehensive and descriptive grammar of English.

6. Case study

Stubbs (2002) is another interesting work which explained the concept of "extended lexical unit," which may well generate thoughts about treating grammar of human language in an alternative fashion.

According to Stubbs (2002: 102), extended lexical units "refers, not to a list of fixed phrases, but to abstract semantic units, which have typical but variable lexical realizations." Note the words *typical* and *variable*. One may well remember the prototype theory in semantics where there is family resemblance between members of

[2] The term is somewhat misleading since it refers to grammar for a special kind of language (or, sublanguage) rather than for some part of a general language.

a category, but there is no necessary and sufficient condition for them. For example, penguins and ostriches are all BIRDS, although sparrows may be a more prototypical bird for people in many countries. In Stubbs' scheme, extended lexical units are a kind of linguistic device important to the encoding of meaning in text. They constitute a level of text organisation between lexis and syntax, which is highly complex and relatively abstract. Extended lexical units involve the selection of words based on collocational restrictions and local context following the decision of a central word. For example, according to Stubbs (p. 86), the frequent collocates of *backdrop* include *provide*, *take*, and *perfect*; while typical phrases where *backdrop* will appear are: *provide the perfect backdrop for*, *take place against a backdrop of*, and so on.

As an illustration of the idea of extended lexical units, data from a Chinese "fire report" corpus I collected is presented in the form of concordance lines in (1).

(1)

00001:	點都是灌木叢和草原，	火勢一發不可收拾	，從空中俯瞰，數公里
00002:	左右鄰居全是家具行，	火勢一發不可收拾	，迅速蔓延到後方的鐵
00003:	日晚間突然起火燃燒，	火勢一發不可收拾	，由於主要火勢集中在
00004:	間突然傳出火警，由於	火勢一發不可收拾	，消防人員緊急出動多
00005:	木造和鐵皮加蓋，因此	火勢一發不可收拾	，短短半個小時內，就
00006:	不足，房屋又是木造，	火勢一發不可收拾	，雖然沒有造成任何人
00007:	，延燒 42 小時，不但	火勢一發不可收拾	，損失慘重，更燒出高
00008:	平房，突然起火燃燒，	火勢一發不可收拾	，一連波及附近 6 棟木
00009:	造平房突然冒出火花，	火勢一發不可收拾	，還波及到旁邊另外一
00010:	房，全都是易燃物質，	火勢一發不可收拾	，目睹整個經過的民眾
00011:	全部都是木材易燃品，	火勢一發不可收拾	，嘉義市消防局出動了
00012:	棄物等易燃物品，使得	火勢一發不可收拾	，猛烈的火舌不斷往外
00013:	場堆放許多易燃物品，	火勢一發不可收拾	，消防局動用 11 輛消
00014:	於四周都是木造樓房，	火勢一發不可收拾	，嘉義市警局出動 10
00015:	，由於山上滿佈雜草，	火勢一發不可收拾	，所幸老天爺幫忙，降
00016:	存放了大批化學物品，	火勢一發不可收拾	，足足燃燒一個多小時
00017:	爲燒垃圾導致火燒山，	火勢一發不可收拾	，延燒將近 10 個小時
00018:	由於現場放有瓦斯桶，	火勢一發不可收拾	。所幸火勢在一個小時
00019:	26 日上午發生大火，	火勢一發不可收拾	，台北縣警消共出動了
00020:	學工廠剛剛發生大火，	火勢一發不可收拾	，台北縣消防局已經派
00021:	有人員傷亡。 猛烈的	火勢一發不可收拾	，甚至還傳出陣陣的爆
00022:	一些油墨易燃物，因此	火勢一發不可收拾	，最後竟然還延燒到旁
00023:	，現場搶救困難，以致	火勢一發不可收拾	。 附近居民表示，電
00024:	易燃的乾草和赤竹林，	火勢一發不可收拾	，熊熊大火不斷沿著山

00025: 內堆放大量易燃物品， 火勢一發不可收拾 ，近兩千坪的廠房，在
00026: 木材等易燃物品，因此 火勢一發不可收拾 。 位在宜蘭市佔地將
00027: 物燥加上風速助長，致 火勢一發不可收拾 ，從嶺東路往上竄燒到
00028: 物品，悶燒速度很快， 火勢一發不可收拾 ，雖然台北縣消防局調
00029: 附近都是乾燥的稻草， 火勢一發不可收拾 ，從水庫的南面不斷延
00030: 是一些易燃物質，因此 火勢一發不可收拾 ，黑夜中就看到一團火
00031: 堆放的都是易燃物品， 火勢一發不可收拾 ，現場還不時傳出爆炸
00032: 裡頭堆放大量木材造成 火勢一發不可收拾 ，幸好沒有人員傷亡，
00033: 廠內堆放的大量木材， 火勢一發不可收拾 ，火大得連幾公里外都
00034: 材行堆放大量木材造成 火勢一發不可收拾 ，嘉義市消防隊出動了
00035: 部都是木造隔間，使得 火勢一發不可收拾 。這起火災真正的原因
00036: 內堆放大量易燃物品， 火勢一發不可收拾 ，甚至波及到一旁的民
00037: 的鐵皮建築物，以致於 火勢一發不可收拾 ，300 多名警義消全
00038: 屋內都是木製的裝璜， 火勢一發不可收拾 ，經消防人員全力灌救

In (1), 火勢 ("fire flames") is the centre of the extended lexical phrase, and 火勢一發不可收拾 ("Once fire broke out it could not be contained") is a typical realisation of this unit. Other similar patterns are:

(2) 00208: 但風勢助長大火延燒， 火勢 延燒猛烈迅速，從第一
00209: 左右，突然發生火警， 火勢 延燒猛烈，消防人員擔
00210: 型機車早上突然起火， 火勢 延燒的很迅速，造成 4
00211: 積大量塑膠製品，使得 火勢 延燒相當迅速，爆炸聲
00212: 房屋，又沒有防火巷， 火勢 延燒迅速，而鄰近還有
00197: 的家具行被大火燒燬， 火勢 延燒 1 個小時才被控制
00198: ，經開闢防火線滅火， 火勢 延燒 6 小時才控制撲滅
00199: 鐵皮廠房更是得見骨， 火勢 延燒一個小時才控制，
00202: 於山路崎嶇救援困難， 火勢 延燒了好幾個小時才被
00206: 是持續灌救，直到晚間 火勢 延燒將近一公頃才獲得

Arguably, the fuzzy pattern of 火勢延燒_____才被控制/撲滅 ("the fire's momentum continued to spread _____ before it was contained"), where the empty slot can be filled by a quantity phrase referring to either time or space, can be subsumed under a larger pattern including the instances in (1), in which the central word 火勢 heads the unit followed by a structure denoting the consequence of the fire. In this manner, many parts of a text can be explained by the extension and cohesion between these lexical units. Grammar, in the sense of "strict rules", would then play a less important role, as the reality may well be – for some languages other than

English at least.

7. Conclusion

Referring to Sampson (1997), Cowley (2001) put it somewhat humorously: "What Sampson, to my judgement, successfully demolishes is Pinker's thesis that we possess innate knowledge of the X bars, traces, cases and other representations posited in generativist theory," and then somewhat and alarmingly to current and future linguistic workers: "Forty years of intensive research have failed to provide any empirical base for their theorizing."

A recent book on machine translation (Nirenburg et al. 2003) consists of 36 researchers' articles, only three of which refer to Chomsky's early work published in 1955, 1957, and 1965. What is the significance of this? Machine translation is one of the forms of language engineering which plays an increasingly important role in information processing and communication. A personal experience may help exemplify this point. The other day I was submitting an English article to a translation journal which demands a French abstract. I used a Web-based MT system to translate my English abstract into French and showed it to a French professor prior to submitting the paper. The professor was amazed to find that the abstract was quite good and usable. Machine translation has its own linguistic models of a theoretical nature to which generative grammar appears to contribute little. The same is true in fields like language teaching and research in the arts in general. Although generative grammar appears to generate research or ideas in some science disciplines, the degree and nature of influence remains to be confirmed.

It may be argued that generative grammar is like theories about the universe, such as the Big Bang theory of the origin of the universe, or the putative existence of Black Holes. Such theories don't have obvious practical use but they add to human knowledge or imagination about the universe. The Big Bang theory however, is supported by the discovery of traces of the explosion and astronomical data showing the continuing expansion of the universe. If generative grammar is to become a convincing theory of human language, more concrete evidence must be supplied, such as the finding of a genome which unambiguously moves a wh- word, assigns a θ-role, or guards the principle of c-command. But until this is possible, generative grammar remains a "toy theory" of little practical use and it may even divert researchers from the pursuit of knowledge. For example, Andrew Radford, a writer of syntax textbooks for linguistics students commented in one of his recent publications:

It seems reasonable to suppose that competence will play an important part

> in the study of performance, since you have to understand what native speakers tacitly know about their language before you can study the effects of tiredness, drunkenness, etc. on this knowledge (Radford 1997:2).

To equate performance to abnormal conditions like tiredness and drunkenness in human communication is a serious misconception. "Performance" is a perfectly normal and essential part of language which should be distinguished from "performance errors." Radford seemed to imply that once a sentence is created from their syntactic module in its pure form, it can be delivered straight away to the listener or reader, without having to meet any functional requirements demanded by various discourse constraints. This is certainly not the case. Knowing that there is a functional component in language and refrain from investigating it is one thing; while being entirely ignorant of it is quite another.

The reason I chose to introduce corpus and grammar in this way was actually generated by Borsley's (2002) review of Hunston & Francis' (2000) work (i.e. *Pattern Grammar*), where Borsley apparently thought generative grammar so important that for a book with *grammar* in its title not to refer to it was ignorant and was "notable for the narrowness of its authors' intellectual horizons." I think the opposite is true. Studying corpora provides us with the widest scope possible in linguistics. It is the generative grammarians who drove Bloomfield and the structuralists away (who should in fact be credited for their comprehensive fieldwork) and confined themselves in the ivory tower of syntax to whom the label "narrow scope" seems better suited. Thus instead of dwelling on the content of grammar-related research based on corpora (which is still coming together anyway), I focused on pointing out the fundamental implausibility of generative grammar. By doing so, I hope to divert the attention of as many future linguistic researchers as possible to the more useful and realistic approaches to language, including sociolinguistics, psycholinguistics, applied linguistics, discourse analysis, lexical semantics, functional-systematic grammar, computational linguistics, and so on, and of course, corpus linguistics.

References

Aitchison, Jean. 1998. *The Articulate Mammal: An Introduction to Psycholinguistics (4th ed)*. London: Routledge.

Barnbrook, Geoff & Sinclair, John. 2001. Specialised Corpus, Local and Functional Grammars. In *Small Corpus Studies and ELT: Theory and Practice*, ed. by Ghadassey, M, Henry, A. & Roseberry, R., 237-276. Amsterdam: John Benjamins.

Bates, Elizabeth. 2003. Natura e cultura nel linguaggio [On the nature and nurture of language]. In R. Levi-Montalcini, D. Baltimore, R. Dulbecco, & F. Jacob (Series Eds.) & E. Bizzi, P. Calissano, & V. Volterra (Vol. Eds.), Frontiere della biologia [Frontiers of biology]. Il cervello di Homo sapiens [The brain of homo sapiens]. Rome: Istituto della Enciclopedia Italiana fondata da Giovanni Trecanni S.p.A., 241-265. (Accessible online: http://crl.ucsd.edu/~bates/papers.html)

Bierwisch, Manfred. 2001. "Generative Grammar". In J. Smelser, Neil J. & Paul B. Baltes (Hg.), *International Encyclopedia of the Social and Behavioral Sciences*, 6061 – 6067.

Bishop, D.V.M. 1997. *Uncommon understanding: Development and disorders of comprehension in children*. Hove, UK: Psychology Press/Erlbaum.

Borsley, Robert D. 2002. Review of S. Hunston and G. Francis, Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English, Amsterdam: John Benjamins, 2000. *Lingua* 112: 235–241.

Charniak, E. 1993. *Statistical Language Learning*. MIT Press: Cambridge, Mass.

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

Chomsky, Noam. 1986. *Knowledge of Language*. Praeger: New York.

Cook, Vivian. 1988. *Chomsky's Universal Grammar: An Introduction*. Oxford: Blackwell.

Cowley, S.J. 2001. The baby, the bathwater and the "language instinct" debate. *Language Sciences*. 23: 69-91.

Dingwall, William O. 1998. The biological bases of human communicative behaviour. *Psycholinguistics (2nd ed)*, ed. by Gleason, J.B. & Ratner, N.B. Orlando, FL: Harcourt College Publishers.

Francis, G.ill. 1993. A corpus-driven approach to grammar: Principles, methods and examples. In Baker, M., Francis, G.. & Tognelli-Bonelli E. (eds). *Text and technology: In honour of John Sinclair*, 137-156. Amsterdam: John Benjamins.

Grodzinsky, Yosef. 2000. The neurology of syntax: language use without Broca's area. *Behavioral and Brain Sciences*. 23(1): 1-71.

Hunston, Susan. 2002. *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Hunston, Susan. & Francis, G.. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English*. Amsterdam: John Benjamins.

Jerne, Niels. 1985. The generative grammar of the immune system. *Science*. 229: 1057-1059.

Ji Sungchul. 1997. Isomorphism between cell and human languages: molecular biological, bioinformatic and linguistic implications. *BioSystems.* 44: 17-39.

Kaan, E. & Swaab, T.Y. 2002. The brain circuitry of syntactic comprehension. *Trends in Cognitive Science*. 6: 350-356.

Leonard, L.B. 1997. *Specific language impairment.* Cambridge, MA: MIT Press.

Lorenzo, G. & Longa, V.M. 2003. The minimalist program as a biological framework for the study of language. *Lingua*. 113: 643-657.

Meyer, C.F. 2002. *English Corpus Linguistics : An introduction*. Cambridge: Cambridge University Press.

Moro A., Tettamanti M., Perani D., Donati C., Cappa S.F. & Fazio F. 2001. Syntax and the brain: disentangling grammar by selective anomalies, *NeuroImage*, 13: 110-118.

Müller, R.-A., Kleinhans, N. & Courchesne, E. 2003. Linguistic theory and neuroimaging evidence: An fMRI study of Broca's area in lexical semantics. *Neuropsychologia*. 41: 1199-1207.

Newmeyer, F.J. 1997. Genetic dysphasia and linguistic theory. *Journal of Neurolinguistics.* 10(2-3): 47–73.

Nirenburg, S.H., Somers, H. & Wilks, Y. (eds). 2003. *Readings in machine translation.* Cambridge, Mass.: MIT Press.

Nowak, M.A & Komarova, N.L. 2001. Toward an evolutionary theory of language [Opinion]. *Trends in Cognitive Sciences*. 5(7): 288-295.

Page, Karen M. (in press). How much evidence does a child need in order to learn to speak grammatically. *Bulletin of Mathematical Biology*.

Radford, Andrew. 1997. *Syntax: A Minimalist Introduction*. Cambridge: Cambridge University Press.

Snow, Catherine E. 1996. Toward a rational empiricism: why interactionism is not behaviourism any more than biology is genetics. *Toward a Genetics of Language*, ed. by Rice, Mabel L. Mahwah, NJ: Lawrence Erlbaum Associates.

Stubbs, Michael. 2002. *Words and Phrases: Corpus Studies of Lexical Semantics.* Oxford: Blackwell Publishing.

Taylor, John. 2002. *Cognitive Grammar*. Oxford: Oxford University Press.

Tsonis, Anastasios A., Elsner James B., Tsonis Panagiotis A. 1997. Is DNA a language? *Journal of Theoretical Biology*. 184(1): 25-29.

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

[Received 30 January 2004; revised 10 May 2004; accepted 12 May 2004]

Centre for Applied Language Studies
University of Wales Swansea
Singleton Park, UK
c-c.shei@swansea.ac.uk

語料庫與文法從反面談起

解志強

應用語言學研究中心
威爾斯大學斯灣西分校

文法這觀念對每個人有不同意義。對於衍生文法學者來講，文法是與生俱來的，是自給自足的，也是全人類通用的。對功能語法學家來講，文法只是人類所使用的諸多表達手段的其中一種。到底文法是不是一種天生的語言核心，如同衍生文法家所言，可以和語用完全脫節，可能終究還是需要心理語言學、神經語言學和生物學方面的研究來定案。在此同時，語料庫語言學仍舊是較爲可靠的描述語言實際使用狀況的工具。經由語料庫的研究，許多字典得以編纂完成，許多語言結構和規則得以問世。本文將進行一項選擇性與評估性的回顧與探討，主要是以心理語言學的角度來看衍生文法學派的主張是否正確。作者主張所謂的自給自足的句法部門並不存在，將文法和語境的研究完全分離是錯誤的。雖然與生俱來的說法的確有可能，然而人腦中的文法部門仍難免受到環境和互動的持續影響和塑造。或許，文法在人類語言當中的地位，並不像衍生語言學派所以爲的那麼重要。從語料庫的角度來研究文法和語言的關係，或許更貼近於事實。

關鍵詞：衍生文法學、心理語言學、失語症、腦神經攝影、語料庫語言學、延伸單字群