



CogniTExtes

Revue de l'Association française de linguistique cognitive

Volume 19 | 2019

Corpora and Representativeness

Conversational corpora : when “big is beautiful”

Frederick J. Newmeyer



Electronic version

URL: <http://journals.openedition.org/cognitextes/1584>

DOI: 10.4000/cognitextes.1584

ISSN: 1958-5322

Publisher

Association française de linguistique cognitive

Electronic reference

Frederick J. Newmeyer, « Conversational corpora : when “big is beautiful” », *CogniTExtes* [Online], Volume 19 | 2019, Online since 17 June 2019, connection on 19 June 2019. URL : <http://journals.openedition.org/cognitextes/1584> ; DOI : 10.4000/cognitextes.1584

This text was automatically generated on 19 June 2019.



CogniTExtes est mis à disposition selon les termes de la licence Creative Commons Attribution - Pas d'Utilisation Commerciale - Pas de Modification 4.0 International.

Conversational corpora : when “big is beautiful”

Frederick J. Newmeyer

EDITOR'S NOTE

Please also read Maarten Lemmens' response to this paper further on in this issue (<http://journals.openedition.org/cognitextes/1616>).

1. Introduction

- 1 The goal of this paper is to examine the relationship between corpus size and conclusions drawn from corpora regarding questions of grammatical theory. Usage-based grammarians typically assert that if one focuses on naturally occurring discourse drawn from corpora, then grammar will reveal itself to be primarily a matter of memorized formulas and very simple constructions. As Sandra Thompson has put it, ‘what we think of as grammar may be best understood as combinations of reusable fragments’ (Thompson 2002 : 141). This paper challenges that view. Appealing to a 170MB corpus of conversational English, it argues that introspective data and corpus-derived data do not lead to different conclusions about the nature of linguistic theory. Usage-based grammarians have been led to their incorrect conclusions primarily by appeal to corpora that are too small to reveal the full complexity and abstractness of English grammar.
- 2 Overwhelmingly, generative grammarians have used introspective data in formulating their theories. This is the case for the reasons outlined in Schütze (1996) : They provide data not obtainable from spontaneous speech or recorded corpora ; they provide negative information, that is, information about what is not possible for the speaker ; they allow for the easy removal of irrelevant data, such as slips of the tongue, false starts, etc. ; and, more controversially, they allow the researcher to abstract away from the communicative function of language and thereby to focus on the mental grammar as a structural system.

Furthermore, it is sometimes claimed that the factors affecting judgments tend to be less mysterious than those affecting use, and that in providing an alternative path to the grammar from language use, 'we have a basis on which to search for the common core that underlies both kinds of behaviour' (Schütze 1996 : 180).

- 3 There have been many different critiques of introspective data and it is not my intention to address them all. The most pervasive is that an introspection is an experiment, though a totally uncontrolled one, and therefore subject to the (normally unconscious) theoretical biases of the investigator. Therefore, the argument goes, introspective judgments are too unreliable to serve as a data base for linguistic theorizing. I do not wish to deny that exclusive reliance on introspective judgments has had negative effects, though how often theoretical proposals have been led astray by the use of such judgments is an open question (see Cowart 1997 and the papers in Schindler to appear for discussion).

- 4 The two major (albeit not mutually exclusive) alternatives to introspection are experimental evidence, which is favoured mainly by psychologically-oriented linguists, and corpus-based evidence, which is favoured mainly by functionally-oriented, variation-oriented, and NLP-oriented linguists. The latter alternative, in a nutshell, holds that we speak in order to communicate and so it follows that only communication-based corpora are valid as evidence when probing the nature of grammar. As Michael Tomasello put it :

The data in generative grammar analyses are almost always disembodied sentences that analysts have made up ad hoc, ... rather than utterances produced by real people in real discourse situations ... [Only the focus on] naturally occurring discourse [has the potential to lead to] descriptions and explanations of linguistic phenomena that are psychologically plausible. (Tomasello 1998 : xiii).

- 5 In other words, nothing good theoretically can come out of a data base that is essentially rotten. What critics such as Tomasello advocate is a near-exclusive focus on natural conversation. Since, at least as they see things, conversation is the principal function of language, it stands to reason, they would say, that the properties of grammars should reflect the properties of conversation. In other words, the complex abstract structures of generative grammar are an artifact of the appeal to introspectively-derived example sentences that, they say, nobody would ever actually use. The conclusion that they draw is that if one focusses on naturally occurring conversation, grammar will turn out to be a collection of stock memorized phrases, 'formulas' or 'fragments' as they are often called, and very simple constructions.

- 6 At the very best, comments like those of Tomasello are non-sequiturs. A much older parallel criticism by Clark and Haviland questions the use of introspective judgments as data because '[w]e do not speak in order to be grammatical ; we speak in order to convey meaning' (Clark and Haviland 1974 : 116). But Wexler and Culicover (1980) replied that even if this is true :

... no logical relation can be established between the purpose of our speech and the kind of data on which to base linguistic theories. Function can also be established in biology. For example, the function of certain molecules is associated with genetic transmission. But this function once again does not dictate choice of data. We do not say that the biologist's enterprise is odd because he uses X-ray photographs although it is not the purpose or function of the relevant molecules to provide photographs. (Wexler and Culicover 1980 : 395)

- 7 Turning to experimental data, I have little to say about it, except to point out that corpus-oriented linguists tend to reject it as well :

The constructed sentences used in many controlled psycholinguistic experiments are themselves highly artificial, lacking discourse cohesion and subject to assumptions about default referents [...] (Bresnan 2007 : 297)

- 8 By and large, generative grammarians have simply ignored the critique of introspective judgments coming from usage-oriented linguists. And when we have talked about data, it has been more a matter of defending the reliability and consistency of introspective judgments than challenging the idea that a focus on conversation as opposed to introspection leads to a simplified view of grammar. In this paper, however, I argue that introspective data do not lead to grammars that are markedly different from those whose data base is naturally occurring conversation. It follows, then, that introspective data are no less relevant than conversational data to the construction of an adequate grammatical theory.
- 9 For all the hostility that exists to introspective data, not that many linguists have actually taken the trouble to argue that conversational data lead to wildly different grammars than those based on introspective data. In this paper I present and critique two of them. The first is a paper jointly written by Sandra Thompson and Paul Hopper (Thompson and Hopper 2001) that argues that conversational data refute a key hypothesis of generative grammar that was arrived at using introspective data. The second is a full-length volume written by Jim Miller and Regina Weinert entitled *Spontaneous Spoken Language* (Miller and Weinert 1998). The remainder of the paper takes on some general issues about the value of introspection vis-à-vis conversational corpora.
- 10 My data base is the Fisher English Training Transcripts, which I will refer to as
- 11 the ‘Fisher corpora’. The transcripts comprise a 170MB corpus of over 11,000 complete telephone conversations, each lasting up to 10 minutes and containing over 6,700,000 words. All of the examples in this paper are drawn from the Fisher corpora unless noted otherwise.
- 12 The paper is organized as follows. Sections 2 and 3 critique the Thompson & Hopper article and the Miller & Weinert book respectively. I argue that a too-small database led them to incorrect conclusions. Section 4 illustrates the (perhaps surprising) complexity of conversational speech, while §5 argues that very few English constructions are restricted to particular genres. The following §6 addresses the claim that grammars are essentially collections of reusable fragments. Section 7 discusses the pros and cons of conversational corpora and §8 is a brief conclusion.

2. A critique of ‘Transitivity, clause structure, and argument structure : Evidence from conversation’ (Thompson and Hopper 2001)

- 13 One construct that is central to virtually every formal approach is ‘argument structure’, roughly the sorts of dependents that can occur with particular heads. In fact, in most formal theories, syntactic structure, in one way or another, is a projection of argument structure. Therefore, it is not surprising that Thompson and Hopper would want to zero in on argument structure as part of their attack on formal linguistics. They begin by criticizing the use of introspective judgments about verbs like *spray*, *load*, *cover*, *pour*, and so on in the argument structure literature (see Levin 1993 and the references cited there). They do not deny that *spray* and so on have pretty well-defined complement types. But

their feeling is that since these verbs are so rare in conversational speech, they are simply irrelevant to deep grammatical analysis :

[T]he apparent importance of [argument structure] may be an artifact of working with idealized data. Discussions for argument structure have to date been based on fabricated examples rather than on corpora of ordinary everyday talk [...]
(Thompson and Hopper 2001 : 40)

- 14 Their position is that to fully understand how grammar works, one needs to focus on grammatical behaviour of the most common verbs, which, they write, leads to the conclusion that argument structure is a theoretically useless concept. Why ? Because according to Thompson and Hopper :

... the more frequent a verb is, the less likely it is to have any fixed number of 'argument structures'. (Thompson and Hopper 2001 : 49)

- 15 If that were true, then we might have good reason to rethink the importance of argument structure, as well as the introspective data that led us to regard argument structure as important.

- 16 Space limitations do not permit a discussion of Thompson and Hopper's entire paper. I limit my focus to one key part of it, namely their discussion and analysis of the verb *get*. *Get* is the most frequently-used verb in English conversation, after *have* and *be*.

- 17 They write :

Get is a prime example of a verb with no easily imagined argument structures, precisely because it is used in so many lexicalized 'dispersed' predicates and specific constructions'. (Thompson and Hopper 2001 : 49)

- 18 An examination of the Fisher corpora shows nothing like that to be true. I looked at over 1000 instances of *get*, which I divided up proportionally to the frequency of their morphological variants (*get*, *gets*, *getting*, *got*, and *gotten*). My conclusion was that, Thompson and Hopper to the contrary, *get* does indeed have 'easily imagined argument structures'. *Get* certainly has *more* argument structure possibilities than the average verb, but there is nothing difficult to imagine about them. The breakdown is in (1) :

(1) Subcategorization frames for the verb *get* :

___NP 511 49.1 %
___AP 329 31.6
___PP 88 08.5
___Past Part 44 04.2
___NP XP 42 04.0
___to VP 19 01.8
___other 7 00.8
TOTAL 1040

- 19 As can be seen, over 95 % of the instances of *get* occurred in only seven subcategorization frames and over 80 % of its occurrences were before a bare NP or a bare AP. The seven leftover cases were hard for me to classify. For example, *get* occurs before *home* and *here* :

(2) a ___*home* 4
b ___*here* 1

- 20 The categorial status of *home* or *here* is not obvious. There was also an example of *get* with a Quantifier Phrase and one before a *wh*-complement :

(3) a B : how old how old um **how much did you get** when you start
b B : i had **gotten where i was taking** a i think it was uh some some brand

- 21 In (3a), the QP *how much* is probably best analyzed as an AP, though that is open to question. Example (3b) was my only example of *get* before a subordinate clause.

- 22 What is interesting is where one never gets *get*. There is nothing like the subcategorizations in (4) :
- (4) a ___# (*I got)
 b ___ADV (*I got easily ; *tickets for the concert get easily)
 c ___(that) S (*I got (that) he finally believed me)
 d ___for NP to VP (*I got for him to believe me)
 e ___NP's Ving (*I got Mary's helping)
- 23 In other words, *get* occurs in *more* argument structures than *spray* or *load*, but that is hardly an interesting fact in my opinion.
- 24 Finally, it is not the case that the uses of *get* are more construction-specific than those of other verbs. *Get* occurred before a past participle 44 times in my sample. 24 different participles were employed : *accepted, affected, arrested, asked, called, called for, carded, cashed, exposed, hit, ignored, interrupted, married, paid, past, plagued, raided, reassigned, rejected, set up, stationed, stored, stuck, treated*.
- 25 Given the space, I could provide similar arguments for the next most common verbs in conversation : *say, go, know, think, see, come, want, and mean*. There is nothing about their behaviour that challenges traditional views of argument structure based on introspective judgments.
- 26 The question is what caused Thompson and Hopper to get things so wrong. I would say that the problem is the small size of their corpus. The Thompson and Hopper paper is based on only 446 clauses from three face-to-face multi-party conversations. How could one possibly derive valid generalizations about the properties of English from such a small corpus ? So here we see that big really is beautiful !

3. A critique of *Spontaneous Spoken Language* (Miller and Weinert 1998)

- 27 Miller and Weinert (1998) is a major study of the grammatical properties of English conversation. As it turns out, all of the constructions in (5) are missing from their corpus :
- (5) Constructions missing from Miller and Weinert (1998)
 a. adverbial clauses of concession introduced by *although*
 b. adverbial clauses of reason introduced by *since*
 c. gapping
 d. conditional clauses signaled by subject-auxiliary inversion
 e. accusative-infinitive sequences ('exceptional case marking')
 f. gerunds with possessive subjects
 g. gerunds with an auxiliary
 h. initial participial clauses preceding a main clause
 i. infinitives in subject position
 j. infinitives with auxiliaries
- 28 Yet all of these occur in the Fisher corpora :
- (6) a [adverbial clauses of concession introduced by *although*]
 B : **although they may not agree with war** then they are going to support the u._s. government and they're going to support the u._s. soldiers
 b [adverbial clauses of reason introduced by *since*]
 A : **since i've never been much of a power grabber myself** i don't really understand people that that are
 c [gapping]

A : but at the same time you might not have not being in that situation might have had gave you a different outlook on the world on the world and life and such and **and me the same**

d [conditional clauses signaled by subject-auxiliary inversion]

A : **had i known then what i know now**

e [accusative-infinitive sequences ('exceptional case marking')]

A : um **i consider myself to be a pretty open minded person**

f [gerunds with possessive subjects]

A : you know **going back to his firing of his economic advisors**

g [gerunds with an auxiliary]

B : i was kinda surprised they'd i could i could fit in **because of my having been born in england** i i i thought it would just be americans

h [initial participial clauses preceding a main clause]

A : **hoping i never get that far** i just wanna make sure that i don't end up on every committee

i [infinitives in subject position]

A : to yeah to **to get to where they need to do so sunday** would kinda be like the first day of the festivities

j [infinitives with auxiliaries]

A : yeah you know **i wouldn't have wanted to to have brought -em up in a in a Christian controlled**

- 29 The absence of all of these ordinary English constructions from Miller and Weinert's database would be inconsequential if it had not led them to the inevitable conclusions about the bankruptcy of formal linguistic theory. They write that :

[t]he properties and constraints established over the past thirty years by Chomskyans [are based on sentences that] occur neither in speech nor in writing [or only] occur in writing. (Miller and Weinert 1998 : 379)

- 30 And on the basis of that mistake, they go on to question whether English grammar could be anything like what formal linguists propose. But Miller and Weinert's corpus contained only 50,000 words ! How could one possibly reach a reasonable hypothesis about English grammar from a corpus that small ? Indeed, when one considers that the average speaker utters about 16,000 words *per day* (Mehl, Vazire, Ramirez-Esparza, Slatcher, and Pennebaker 2007), it is clear that one cannot conclude anything about grammatical knowledge from a corpus of 50,000 words.

4. The grammatical complexity of everyday speech

- 31 Conversational speech reveals a depth of syntactic knowledge that is absolutely stunning, and which supports the standard introspective judgments that we find in the literature. Consider, for example, a few cases of long-distance *wh*-movement :

(7) a B : so **what do you think that um we should do** um as far as w- we're standing right now with our position

b B : getting back to this subject **where do you want to go with it**

c A : **when do you expect to get together**

- 32 Along the same lines, speakers and hearers can link embedded gaps in relative clause constructions to their antecedents :

(8) a B : you know **when i move away and get the things that i want to have** and retire early and enjoy you know what i mean

b A : actually **following the rules that they need to be following they are doing things that they shouldn't be doing**

c B : that right **if i had time to cook the things that i like to cook** then it would be in home

- 33 To produce and comprehend utterances like these, the language user has to hold in mental storage a position for an unexpressed direct object in a different clause and to link a fronted *wh*-element or lexical antecedent to that position. It is not a matter here of
- 34 ‘fragments’ or ‘formulas’, but rather of a sophisticated engine representing grammatical knowledge.
- 35 Anaphoric relations are among the most difficult grammatical phenomena to extract from corpora, given the difficulty of formulating the right search criteria. But persistence provides some interesting (and I would say surprising) results. For example, one often hears that cataphors (i.e., backwards anaphors) only occur in linguists’ introspective judgments or possibly in educated speech or writing. But in fact they are found in conversation, and both in pronominal and elliptical form :
- (9) a A : when their sons die with with money he rewards the parents and and the parents are quite happy about it
b A : um overseas we i don’t know why we don’t but everybody has flags here we have huge flags on the street
- 36 There are also examples of both forward and backward sluicing in conversation :
- (10) a A : i know i know i’m going to get married some time **but i don’t know when**
b B : we’re supposed to just give one another’s opinion about uh if you like eating at home or if you like eating out more and i **guess why**
(11) a A : **i just i don’t know why but** i don’t usually get sick in the winter time
b B : **i don’t know why but** for whatever reason every night the cat comes and like meow outside our door
- 37 After decades of research, we still do not know what the conditions are for appropriate cataphors and sluices. Nevertheless, speakers handle the relevant structures without effort.

5. The relative independence of structure type and genre

- 38 It is simply not the case that introspection leads to sentences that are confined to a large degree to literary genres. In fact, the differences between spontaneous conversation and what is found in literature are almost always quantitative, rather than qualitative. Consider Douglas Biber’s book *Variation across speech and writing* (Biber 1988). Biber takes 67 grammatical features of English, some of them pretty exotic, and calculates their frequency in 23 different genres, some spoken and some written. Only three of these features occurred in face-to-face conversations at a frequency of less than 0.1 times per thousand words :
- (12) Rare features in the Biber (1988) corpus :
a present participial clauses (e.g. *stuffing his mouth with cookies, Joe ran out the door*)
b past participial clauses (e.g. *built in a single week, the house would stand for fifty years*)
c split infinitives (e.g. *he wants to convincingly prove that*)
- 39 It is hard for me to imagine that any of these sentence types would be rejected by English speakers’ introspective judgments. And all three features are rare in academic prose as well : 1.3, 0.4, and 0.0 times per thousand words respectively in that genre. In fact, it was not difficult to find examples of all three in the Fisher Corpora :

- (13) a B : **having angst** i don't have any like firsthand experience with separations
or anything cause i mean
b A : but **compared to the comedies** now it it's tame
c B : right and they tried **they tried to really make it so people wouldn't get a
long**

40 Table 1 gives the ten most frequent grammatical features in the two genres :

Table 1 : the most frequent grammatical features in two English genres

RANK	FACE-TO-FACE CONVERSATIONS	ACADEMIC PROSE
1	nouns	nouns
2	present tense	prepositions
3	adverbs	attributive adjectives
4	prepositions	present tense
5	first person pronouns	adverbs
6	contractions	type-token ratio (the number of different lexical items in a text, as a percentage)
7	type-token ratio	nominalizations
8	attributive adjectives	BE as main verb
9	BE as main verb	past tense
10	past tense	agentless passive

41 The only features that made the top ten in face-to-face conversations, but not in academic prose were (unsurprisingly) first person pronouns and contractions. Facts like these suggest that the gulf between introspective data and data drawn from corpora, whether conversational or scholarly, is insignificant.

42 The *Longman Grammar of Spoken and Written English* (LSGWE ; Biber et al. 1999) stresses that our preconceived notions about what is common in conversation and what is common in formal academic writing tend to be quite unreliable. For example, it notes that ‘when using a relative clause with the head noun way, academic writers might be expected to use a combination of preposition + relative pronoun — *in which* — since this form explicitly marks how way integrates with the relative clause’ (LSGWE : 7) as in :

- (14) **The way in which this happens** gives important information on the inner organization (Biber et al. 1999 : 7)

43 However, the LGSWE reports that writers of formal prose commonly leave out both the relative pronoun and the preposition, as in :

- (15) Silicates are classified and named according to **the way the tetrahedra are linked**. (Biber et al. 1999 : 7)

- 44 Interestingly, the full combination of preposition and relative pronoun is not rare in conversation :

(16) a. A : the result of september eleventh one of the things that i kind of consciously did was change **the way in which i approach language teaching**
 b. B : **the way in which i was able to get tickets** uh is pretty much no longer don't have my uh in anymore

- 45 It is also not difficult to find syntactic phenomena in formal speech or writing that are characteristic of (or said to be characteristic of) informal conversation. For example, Thompson (2002) argues that there is little or no true syntactic subordination in conversational speech. That is, in a sentence like *I think that we should leave*, the main clause is *that we should leave*, with *I think* having essentially fragmentary status. One of her arguments is that *I think* frequently occurs as a parenthetical, as in *We'll be better off, I think, if we take the Lincoln Highway*. But, in fact, it is not hard to find the same phenomenon occurring in writing :

(17) a. The last natural blondes will die out within 200 years, **scientists believe**.
 [BBC News World Edition, 27 September 2002]
 b. Election will be a turning point, **commentators say** [Taipei Times, 10 January 2006]
 c. '09 Afghan pullout too soon, **experts say** [National Post, 10 January 2008]
 d. Facts prove no match for gossip, **it seems** [New York Times, 16 October 2007]

- 46 What we learn from the work of Biber and others is that there are very few construction types that are confined exclusively to conversational speech or exclusively to writing. That of course does not imply that the same structures appear with equal frequency from genre to genre. Of course they don't. In both informal conversation and in formal writing, we find complement-taking verbs like *think*, *know*, and *believe* both with and without a following *that*-complementizer. However, omitting the *that* is far more common in conversation than in writing. What is important is that our mental grammars provide the possibility of the *that*-complementizer and we can choose to employ it or omit it as we wish (even though the choice we make depends in part of the level of speech).

6. On grammars as ‘combinations of reusable fragments’

- 47 I simply do not understand what it might mean to describe grammars as ‘combinations of reusable fragments’. I have no problem with the idea that many of the more commonly-used phrases are stored in memory, but the idea that a *grammar* might be a stock of fragments strikes me as utterly implausible. How many such ‘fragments’ would it take to characterize the syntactic competence of a speaker of English ? Hundreds of thousands ? Millions ? More likely, I would say, tens of millions, if fragments are lexically specified. Consider the following 20 lines from a small part of one of the conversations in the Fisher corpora :

4.58 5.46 A : hi
 5.81 9.13 B : hi so did you hear what the topic is
 8.61 10.89 A : yes it's about terrorism right
 10.18 11.59 B : yeah
 11.91 12.95 B : um
 13.52 16.71 A : so what are your feelings on that [laughter]

15.44 20.00 B : i have [laughter] i personally can't imagine anyone staying calm
[laughter]
19.20 21.21 A : yeah nor can i yeah
20.87 26.07 B : um you would even i- though if you're panicked i would assume you
would try and
26.38 31.45 B : keep your head clear enough to to act to protect yourself but
29.29 30.42 A : right
31.31 39.34 A : yeah i don't know if there was an explosion or something i don't it
it's a shock so i don't know that anybody can really think about it and control
themselves
31.65 32.40 B : um
39.02 41.91 B : right even with all the um
42.74 43.80 B : (([sigh] the))
43.93 50.33 B : the publicity and media coverage you know that's been on that topic
in the last
47.24 48.58 A : (([mn] right))
50.51 53.01 B : twenty months it's still um
53.16 55.95 B : is something that you wouldn't be
56.15 59.81 B : prepared for and be able to take in stride i don't think

- 48 There are certainly formulaic expressions here : *hi, right, take in stride, I don't think*, and possibly a few others. But in other respects, the transcript reveals a sophisticated knowledge of syntax that defies any meaningful analysis in terms of ‘fragments’. The speakers know how to handle purpose clauses, *wh*-inversion, relative clause attachment, participial complements, and much more. If these are somehow to be subsumed under the rubric of ‘fragments’, then I would say that this infinitesimally small sample of natural speech would have to contain at least two dozen fragments. How many more would be needed to describe a typical speaker's daily output ?
- 49 There are, to be sure, some superficially startling statistics in the literature about the formulaicity of spoken language. For example, Altenberg (1998) found no less than 80 % of the words in the London-Lund corpus to form part of a recurrent word combination. But he counted ‘any continuous string of words occurring more than once in identical form’ (Altenberg 1998 : 101). After limiting himself to word combinations consisting of at least three words occurring at least ten times in the corpus and eliminating unintentional repetitions (*the the the, I was I was*, etc.), the resulting material consisted of only 6,692 tokens representing 470 different types of word combinations. Those 6,692 tokens represent only 1.3 % of the entire corpus.
- 50 In a later study, Erman and Warren (2000) estimated that 58.6 % of spoken texts are filled with what they call ‘prefabs’, where a prefab is ‘a [memorized — FJN] combination of at least two words favoured by native speakers in preference to an alternative combination which could have been equivalent had there been no conventionalization’ (Erman and Warren 2000 : 31). But consider the criterion for identifying prefabs that they appeal to the most, namely ‘restricted exchangeability’ :

By restricted exchangeability is meant that at least one member of the prefab cannot be replaced with a synonymous item without causing a change of meaning or function and/or idiomaticity. For instance, *good friends* in *they are good friends* cannot be changed into *nice friends* without losing the implication of reciprocity ; *not bad* (meaning ‘good’) cannot be changed into **not lousy* without a change of meaning and loss of idiomaticity. *I can't see a thing* cannot be **I can't see an object* without loss of the non-literal hyperbolic meaning ; *I'm afraid* — a pragmatic prefab used to soften a piece of bad news cannot be **I'm scared* or *frightened*. (Erman and Warren 2000 : 32)

- 51 If we take their strategy for identifying prefabs literally, then none of their examples are prefabs, since none of the contrasting words are truly synonymous. *Good* and *nice* almost always have different meanings, as do *bad* and *lousy*, *thing* and *object*, and *afraid* and *scared* / *frightened*. Are any two words true synonyms? I doubt it. In fact, it was Dwight Bolinger, whom they cite as a precursor, who wrote: ‘The natural condition of language is to preserve one form for one meaning’ (Bolinger 1977: x). As far as I can see, the only workable criterion that they have for prefab status is the intuitive idea that some combinations of words (e.g. *not bad* vs. *not lousy*) are produced more frequently than others. (I write ‘intuitive idea’ since they provide no text counts for individual prefabs.) The fact that *not bad* might well be a memorized fragment does not entail that language users cannot and do not compute its meaning and structure by means of principles of grammar.
- 52 In putting all of their eggs in the basket of ‘fragmentation’, many usage-based linguists fall prey to the converse of what Ronald Langacker has aptly termed the ‘rule/list fallacy’: ‘the assumption, on grounds of simplicity, that particular statements (i.e. lists) must be excised from the grammar of a language if general statements (i.e. rules) that subsume them can be established’ (Langacker 1987: 29). For example, the fact that one has learned to multiply does not entail that one might not have committed to memory the fact that twelve times twelve equals 144. But the more extreme usage-based linguists seem to adopt the position that rules should be excised from the grammar if one can establish the need for listing the items in question. That cannot be correct.
- 53 Any open-ended system where users have the ability to interpret novel strings has no alternative but to posit rule-like mechanisms alongside lists. And those who place formulaic language on center-stage tend to focus almost exclusively on language production, all but ignoring comprehension, and show no interest at all in language users’ ability to make judgments of the well-formedness of sentences that they have never heard. Interpreting novel strings and making judgments of well-formedness require computational ability — that is, they require a grammar.

7. Some general issues regarding conversational corpora

- 54 The remainder of this paper is devoted some general issues regarding conversational corpora. Section 7.1 stresses the ways in which they are of great value, while the following 7.2 points to some of their limitations.

7.1. Some positive features of conversational corpora

- 55 There is clearly no substitute for conversational corpora if one’s interest is the study of
- 56 the structure of conversations and broader discourses. That point should be uncontroversial. But they are also quite useful for grammatical theorists. We have just seen how they can be used to rebut extravagant claims about what is supposedly not found in ordinary usage. And that works both ways. Syntacticians using only introspective data tend to be much too quick to label a sentence type ‘ungrammatical’ when it is easy to find evidence that that sentence type is commonly used. Consider an example from Bresnan, Cueni, Nikitina, and Baayen (2007). Most treatments of the dative

alternation say that the verb *give* appears with a prepositional object only if there is movement to a goal. So supposedly the (a) sentences of (18-19) are grammatical and the (b) sentences are ungrammatical :

- (18) a The movie gave me the creeps.
- b (*)The movie gave the creeps to me.
- (19) a The lighting here gives me a headache.
- b (*)The lighting here gives a headache to me.

57 But as Bresnan et al. demonstrate on the basis of conversational evidence, sentences like those in (b) are not infrequent :

- (20) a This life-sized prop will give the creeps to just about anyone !
- b That smell would give a headache to the most athletic constitution.

7.2. Some negative features of conversational corpora

58 The following subsections outline briefly the limitation of conversational corpora.

7.2.1. The first limitation of conversational corpora : their not providing ungrammatical sentences

59 No corpus can provide sentences that *do not occur*. Yet ungrammatical sentences have played a key role in the development of grammatical theory. Even the absence of a construction type from a conversational corpus of millions of words is no guarantee that it does not form part of the linguistic competence of a native speaker. Ironically, even the Thompson and Hopper paper appeals to ungrammatical sentences in several places to help underscore its points. In other words, if for no other reason there will always be a place for introspective data.

60 Consider an important related point. Corpus-oriented linguists focus almost exclusively on language *production* and tend to ignore comprehension completely. After all, how might one reliably extract comprehension data from a corpus ? No doubt carefully-designed experiments are possible, but introspective judgments are still our best hope for deciding if a sentence is comprehensible not. Along the same lines, linguists like Hopper and Thompson never discuss our ability to make judgments about sentences that we have never heard. As stressed above, interpreting novel strings and making judgments of well-formedness require computational ability, that is, they require a grammar, not just a memorized stock of formulas and simple constructions.

7.2.2 The second limitation of conversational corpora : their conflating the grammars of speakers of different varieties of the same language

61 The second limitation is based on the fact that nothing can necessarily be concluded about the linguistic competence of an individual speaker on the basis of corpora including utterances from various speakers who are not all members of the same speech community. The Fisher transcripts go out of their way to include American English speakers from different walks of life, different regions, and different income levels. That is a very good thing if one wants a feel for the diversity of the language spoken across the country. But it is a disaster if one is probing the grammatical competence of an individual speaker. And that after all is what generative grammatical theory is all about. Psychologically speaking, there is the I-language of an individual and there are the universals common to all grammars, but really nothing in between. In grammatical

theory there is no concept like ‘pan-American English’, which can be motivated by pooling the output of a large number of speakers.

- 62 Consider a concrete example. Many speakers of American English produce what are called ‘positive *anymore*’ sentences. In fact, they occur in the Fisher corpora :

(21) a B : most of my time is leisure **anymore** so and

b A : it’s a fact of life **anymore**

c A : well you know the ones that are on t.v. **anymore** are getting pretty racy

- 63 I had never heard a ‘positive *anymore*’ sentence until I was away at university and the first time I heard one I could not parse it. I am still not sure how to use it appropriately.

- 64 How could my grammar conceivably be the same as that of an English speaker who uses the construction natively ? In fact, Labov (1972) has used positive *anymore* sentences as an argument against pan-dialectal grammars, given that most people who do not use the construction do not know what it means. By the way, despite what is often said, it is not a simple synonym of ‘*nowadays*’. So (22a) (a constructed example) is possible, but not (22b) :

(22) a I was dealt good hands when we started playing bridge an hour ago, but anymore they’re really crappy.

b * I was dealt good hands when we started playing bridge and hour ago, but nowadays they’re really crappy.

- 65 Consider as well the nonstandard usages by *one speaker* in the Fisher corpora. Speaker A in (23) uses the constructions outlined in (23a-f) :

(23) a The invariant *be* construction :

A : aw i don’t have a best friend my best friend is god so **he be the one to give it to me**

b Coordinated object pronouns in subject position :

A : yeah you know **me and you must be in the same situation**

c *Been* as a simple past :

A : my best friend but uh my best friend is acting up and **he been acting up** and i’m tired of him acting up so i think i’m going to go about my business so

d *Go to* as a synonym for *start* :

A : because **people go to acting funny** when they get money

e *Ain’t got to* as a synonym for *don’t have to* and negative concord :

A : yeah well i pay i pay for it where **i ain’t got to worry about no mortgage**

f Uninflected third person singulars :

A : and you should see my my matter of fact i was just in my bathroom and **my bathroom look like a million dollars**

- 66 I understand all of these sentences, but there is no reason to think that they are generated by my grammar. And without question, there are construction types licenced by my grammar that would be totally foreign to the grammar of Speaker A.

- 67 In brief, there is no way that one can draw conclusions about the grammar of an individual from usage facts about communities, particularly communities from which the individual receives no speech input. There are many non-sequiturs in the literature that arise from ignoring this simple fact. So, Manning (2002) observes that Pollard and Sag (1994) consider sentence (24a) grammatical, but that they put an asterisk in front of sentence (24b) :

(24) a We consider Kim to be an acceptable candidate.

b *We consider Kim as an acceptable candidate.

- 68 Manning then produces examples from the *New York Times* of sentences like (24b). Perhaps (24b) is generated by Pollard’s grammar and perhaps it is not.

- 69 Perhaps (24b) is generated by Sag’s grammar and perhaps it is not. But we will never find out by reading the *New York Times*. The point is that we do not have ‘group minds’. No input data that an individual did not experience can be relevant to the nature of his or her grammar.
- 70 Now one might object that in this paper I have been as guilty as Manning.
- 71 After all, I have been appealing to the Fisher Corpus to make claims about grammatical competence in English in general. But I am simply trying to meet Thompson and Hopper and others on their own terms. What we need are decent-sized corpora of the linguistic behaviour of particular individuals, or at least of individuals in a particular speech community, narrowly-defined. Do they exist ? I do not think so. Until they do, for this reason alone, introspective judgments are irreplaceable.

7.2.3. The third limitation of conversational corpora : their leading linguists to exaggerate the importance of text frequency to the shaping of grammar.

- 72 Frequency of use, as calculated on the basis of evidence from conversational (and other) corpora is uncontroversially an important factor in directing grammatical change. Frequency drives the grammaticalization of locative nouns to adpositions, pronouns to person markers, auxiliaries to tense and aspect particles, and much much more. But a word of caution is necessary here. Joan Bybee, for one, has often pointed to the effect of frequent use on constituent structure. For example, Bybee and Scheibman (1999) give some pretty good evidence that in frequent phrases like *I don’t know*, the subject and the auxiliary form a surface constituent, not the auxiliary and the verb. So consider the two bracketings in (25a-b) :
- (25) a [I] [don’t know] [‘Classical’ analysis]
 b [I don’t] [know] [Bybee & Scheibman analysis]
- 73 They appeal to the evidence supporting (25b) to dismiss traditional generative constituent analysis. But other tests — binding relations for example — support the traditional analysis. What is going on then ? As far as I can see, what we have here is another example of a ‘bracketing paradox’, that is, a situation where a single string requires different analyses at different levels of grammar. (26) gives two well-known examples of bracketing paradoxes :
- (26) a transformational grammarian (lexically [*transformational*] [*grammarian*], but semantically [*transformational grammar*] [*ian*])
 b this is the cat that ate the rat (syntactically [*this is*] [*the cat that ate the rat*], but phonologically [*this is the cat*] [*that ate the rat*])
- 74 I assume that *I don’t know* should be handled in more or less the same way. I certainly do not see anything there that would challenge standard models of grammar.
- 75 Another important point is that the frequent use of a construction type in one language is not necessarily a reliable guide to what occurs crosslinguistically. For example, most English speakers control both ‘preposition stranding’ (27a) and ‘pied-piped’ PPs (27b) :
- (27) a B : this is joe pinatouski **who am i speaking to**
 b A : **to whom am i speaking**
- 76 But stranding is used vastly more often than pied-piping. In the Fisher corpora, the PP *to whom* occurs only 8 times, while the full sentences *Who am I speaking to ?* and *Who am I talking to ?* occur 24 times and 26 times respectively. One might predict on this basis that stranding would be more common than pied-piping crosslinguistically. Such is not the

case, however. Stranding is attested only in Germanic (but not in German and only marginally in Dutch) and marginally in French.

- 77 Here is one more example of how frequency fails to predict typological distribution. Keenan and Comrie (1977) showed that if a language can form relative clauses at all, then it can form them on subjects. One might predict then that subject relatives would be *used more often* than object or oblique relatives. Apparently, this is not consistently the case. Fox and Thompson (1990) found that with nonhuman referents and when the head NP is a matrix subject, 77 % of English relative clauses are object relatives.
- 78 In sum, frequency is an important factor leading to the shaping and reshaping of grammar. But appeals to frequency should never be used as a substitute for careful grammatical analysis. Frequency generalizations derived from conversational corpora do not challenge theories constructed on the basis of introspective judgments.

7.2.4. The fourth limitation of conversational corpora : their chaotic nature

- 79 The fourth limitation is probably the most serious. The fact is that conversation is unbelievably messy. What we say is constrained in large part by our grammars, of course, but also by so much more. When we talk, we get distracted, interrupted, and we often change our minds about what we want to say. And all of this happens in mid-stream. Consider a typical exchange from the Fisher corpora :

B : do you have that problem
 A : well i i you know i think
 A : i'm not sure
 B : yeah
 B : flu in your lungs you're saying not just your uh
 A : yeah right just kind of the up you know your chest and your throat and your uh nasal cavities or whatever the heck it is and all that stuff
 B : i get it all in my head and my throat but very very seldom ever any chest problems or any anything that makes me you know nauseous that's not very
 A : right
 A : right
 A : mm
 B : common for me

- 80 In this discourse, we superficially have predicates without subjects, three complement-taking verbs stacked up one after the other with no complements, *say* used seemingly intransitively, and *up* used with a definite article. Only the most dyed-in-the-wool empiricist would argue that English grammar licences sentences like these. There is such a gulf between the ‘syntax of conversation’ (if one would want to call it that) and our mentally-stored grammatical competence that it is plainly dangerous to make too many conclusions about the latter from the nature of the former. Perhaps there is some way not involving introspection to filter out the dysfluencies, but I am not sure how that might be done.

8. Concluding remarks

- 81 The bottom line is that no one form of data is theoretically privileged with respect to any other. Introspective data, conversational data, experimental data, data from ancient manuscripts, and so on all have their place and their pitfalls. Generative grammarians have undeniably tended to appeal to introspective data. But that has been more a

function of convenience than of anything else. It is interesting that Chomsky, at least in the early years, was very critical of introspective data. In *Syntactic Structures* he wrote :

It is also quite clear that the major goal of grammatical theory is to replace this obscure reliance on intuition by some rigorous and objective approach. (Chomsky 1957 : 94)

82 And eight years later in *Aspects of the Theory of Syntax* he wrote :

Perhaps the day will come when the kinds of data that we now can obtain in abundance will be insufficient to resolve deeper questions concerning the structure of language. (Chomsky 1965 : 21)

83 Has that day come ? Quite possibly. Very few generativists would argue that introspective data are 'sufficient' to resolve the deeper questions.

84 To conclude briefly, when introspective data goes head-to-head with data drawn from corpora of conversation, there is little reason to think that the theory of grammar derived from one would differ greatly from a theory derived from the other. Many scholars have thought the corpus-derived data and introspective data do lead to different theories, but they have arrived at this conclusion only because of the small size of their corpora. In short, big is beautiful.

BIBLIOGRAPHY

Altenberg, Bengt. 1998. On the phraseology of spoken English : The evidence of recurrent word combinations. In A. P. Cowie (ed.), *Phraseology : Theory, Analysis and Applications*, 101-122. Oxford : Clarendon Press.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge : Cambridge University Press.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, & Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London : Longman.

Bolinger, Dwight. 1977. *Meaning and Form*. London : Longman.

Bresnan, Joan W. 2007. 'Is syntactic knowledge probabilistic ? Experiments with the English dative alternation'. In Sam Featherston & Wolfgang Sternefeld (eds.), *Roots : Linguistics in search of its evidential base*. Berlin : Mouton de Gruyter, 75-96.

Bresnan, Joan W., Anna Cueni, Tatiana Nikitina, & Harald Baayen. 2007. 'Predicting the dative alternation'. In G. Boume, I. Kraemer, & J. Zwarts (eds.), *Cognitive foundations of interpretation*. Amsterdam : Royal Netherlands Academy of Science, 69-94.

Bybee, Joan L. & Joanne Scheibman. 1999. 'The effect of usage on degrees of constituency : The reduction of *don't* in English'. *Linguistics* 37: 575-596. Reprinted in *Frequency and the organization of language* by Joan Bybee. 2007, Oxford: Oxford University Press, 2294-2312.

Chomsky, Noam. 1957. *Syntactic structures*. Janua Linguarum Series Minor, 4. The Hague : Mouton.

--- 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

- Clark, Herbert H. & Susan E. Haviland. 1974. 'Psychological processes as linguistic explanation'. In David Cohen (ed.), *Explaining linguistic phenomena*. Washington: Hemisphere Publishing Corporation, 91-124.
- Cowart, Wayne. 1997. *Experimental syntax : Applying objective methods to sentence judgments*. Newbury Park, CA: SAGE Publications.
- Erman, Britt & Beatrice Warren. 2000. 'The idiom principle and the open choice principle'. *Text* 20: 29-62.
- Fox, Barbara A. & Sandra A. Thompson. 1990. 'A discourse explanation of the grammar of relative clauses in English conversation'. *Language* 66: 297-316.
- Keenan, Edward L. & Bernard Comrie. 1977. 'Noun phrase accessibility and universal grammar'. *Linguistic Inquiry* 8: 63-99.
- Labov, William. 1972. 'Where do grammars stop ?' In Roger W. Shuy (ed.), *Monograph Series on Languages and Linguistics, 23rd Annual Round Table. Sociolinguistics: Current trends and prospects*. Washington: Georgetown University School of Languages and Linguistics, 43-88.
- Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Volume 1: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
- Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press.
- Manning, Christopher, D. 2002. 'Probabilistic syntax'. In Rens Bod, Jennifer Hay, & Stefanie Jannedy (eds.), *Probabilistic linguistics*. Cambridge, MA: MIT Press, 289-341.
- Mehl, Matthias R., Simine Vazire, Nairán Ramirez-Esparza, Richard B. Slatcher, & James W. Pennebaker. 2007. 'Are women really more talkative than men ?' *Science* 317 : 82.
- Miller, Jim & Regina Weinert. 1998. *Spontaneous spoken language: Syntax and discourse*. Oxford : Clarendon.
- Pollard, Carl & Ivan A. Sag. 1994. *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.
- Schindler, Samuel (ed.). To appear. *Linguistic Intuitions, Evidence, and Expertise*. Oxford: Oxford University Press.
- Schütze, Carson. 1996. *The empirical basis of linguistics: Grammaticality judgments and linguistic methodology*. Chicago: University of Chicago Press.
- Thompson, Sandra A. 2002. 'Object complements' and conversation: Towards a realistic account'. *Studies in Language* 26: 125-164.
- Thompson, Sandra A. & Paul J. Hopper. 2001. 'Transitivity, clause structure, and argument structure: Evidence from conversation'. In Joan L. Bybee and Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins, 27-60.
- Tomasello, Michael (ed.). 1998. *The new psychology of language: Cognitive and functional approaches to language structure*. Mahwah, NJ: Lawrence Erlbaum.
- Wexler, Kenneth & Peter Culicover. 1980. *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

ABSTRACTS

The goal of this paper is to examine the relationship between corpus size and conclusions drawn from corpora regarding questions of grammatical theory. Many linguists, typically those with a ‘usage-based’ orientation, assert that introspective data, unlike corpus-derived data, is of little relevance to the construction of the correct theory of language. This paper challenges that view, arguing that introspective data and corpus-derived data do not lead to different conclusions about the nature of linguistic theory.

INDEX

Keywords: argument structure, data, corpus-derived data, introspective generative syntax, usage-based linguistics

AUTHOR

FREDERICK J. NEWMAYER

University of Washington, University of British Columbia, and Simon Fraser University