In defense of frequency generalizations and usage-based linguistics. An answer to Frederick Newmeyer's "Conversational corpora : when big is beautiful"

Maarten Lemmens

CogniTextes
Revue de l'Association française de linguistique cognitive

Volume 19 | 2019
Corpora and Representativeness

In defense of frequency generalizations and usage-based linguistics. An answer to Frederick Newmeyer's "Conversational corpora : when big is beautiful"

Maarten Lemmens

Electronic version
URL: http://journals.openedition.org/cognitextes/1616
DOI: 10.4000/cognitextes.1616
ISSN: 1958-5322

Publisher
Association française de linguistique cognitive

Brought to you by Université de Lille

Electronic reference
Maarten Lemmens, « In defense of frequency generalizations and usage-based linguistics. An answer to Frederick Newmeyer's "Conversational corpora : when big is beautiful" », *CogniTextes* [Online], Volume 19 | 2019, Online since 17 June 2019, connection on 05 September 2019. URL : http://journals.openedition.org/cognitextes/1616 ; DOI : 10.4000/cognitextes.1616

In defense of frequency generalizations and usage-based linguistics. An answer to Frederick Newmeyer's "Conversational corpora : when big is beautiful"

Maarten Lemmens

EDITOR'S NOTE

Please also read Frederick Newmeyer's position paper (http://journals.openedition.org/cognitextes/1584) and rejoinder to Maarten Lemmens' response paper in this issue (http://journals.openedition.org/cognitextes/1657).

1. Introduction

1 In his paper "Conversational corpora: when big is beautiful", Newmeyer sets himself the goal of evaluating the relationship between corpus size and conclusions drawn from corpora regarding questions of grammatical theory. He formulates a strong critique against corpus research based on too small (conversational) corpora (such as Thompson & Hopper 2001) and in doing so, explicitly rejects the usage-based approach to language in which they are embedded. The gist of his critique is that the findings of these corpus studies do not lead to a different model of language than what has been suggested by introspection-based analyses. Ironically perhaps, he proves them wrong by using a (conversational) corpus himself, even if he says that what he is doing is "simply trying to meet Thompson and Hopper and others on their own terms". "What we need", Newmeyer says, "are decent-sized corpora of the linguistic behaviour of particular individuals, or at

least of individuals in a particular speech community, narrowly-defined". However, as such corpora do not exist, we have no choice, Newmeyer says, but to use introspective judgments which, if only "for this reason alone", are irreplaceable. The crucial message Newmeyer wants to send home is that "[m]any 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."

2 As a usage-based linguist who makes ample use of (extensive) corpus data, I could not be more thrilled to hear a highly respected generative linguist plead for the creation and use of better and larger corpora. I furthermore strongly agree with him that corpus or frequency studies "should never be used as a substitute for careful grammatical analysis". Corpus studies should be done with great care, where one pays attention to (i) the scope and/or representativity of the data, (ii) methodological rigidity, (iii) analytical correctness (i.e., correct analysis of the data), and (iv) correct interpretation of the results. Nothing is more damaging to linguistics than bad corpus work. I also endorse Newmeyer's view 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." This is plainly true, even if I would perhaps be more cautious as to the use of introspective judgment to build a (universal) theory of grammar (see below for some discussion).

3 Despite these positive elements, I wish to take issue with some of Newmeyer's claims. First of all, I provide evidence against his critique of claims he attributes to usage-based linguistics regarding the complex and all-pervasive role of frequency with respect to the structure of grammar. Still today, many people mistake the usage-based linguists for being a simplified behavioristic learning model which it clearly is not. For example, a crucial distinction that must not be overlooked in how people learn grammar from input is that between token and type frequency; it is particularly the latter which contributes to the creation of grammatical structures. As a result, his criticism needs reevaluation (see below). Even if Newmeyer's overt criticism targets the (corpus-based) frequency analyses (which usage-based linguists suggest influences the structure of the grammar), an additional underlying goal of his paper is a defense of using introspective judgments as a basis for a grammatical theory. Ironically, he disqualifies usage-based corpus analyses by drawing on a corpus analysis himself, with the ultimate aim to reject this kind of data in favour of introspection-based data yet simultaneously calling for bigger (conversational) corpora.

4 Again, I applaud Newmeyer's call for using larger corpora, but I cannot help wondering whether studies that would use such corpora would succeed in convincing him, given his conception of the linguistic model that usage-based linguists propose and of how corpus analyses come to support such a view. As we will show below, if interpreted correctly, the large body of corpus-based work ***does*** suggest a different model of grammar. While I can still understand Newmeyer's wish to counter corpus-based analyses "on their own terms", his own use of corpus data emanates from a particular theoretical *a priori* as to what linguistic data mean (see section 3 below).

5 Finally, usage-based linguistics is still a fairly young model; it started in the late 80s of the previous century (cf. the publication of Langacker's *Foundations of Cognitive Grammar* in 1987 (Vol. 1) and 1991 (Vol. 2)) and has gained more momentum in the last two decades, especially within the framework of Construction Grammar (see Goldberg 1995, 2006; Hilpert 2014). While understandably Newmeyer may not find it desirable to get too

heavily involved in a linguistic model that he fundamentally disagrees with, some of the points he raises have been refined, elaborated and confirmed by more recent work. See, for example, Schmid & Handl (2010), Herbst *et al.* (2014), Perek (2015), or Divjak *et al.* (2017), to name but a few.

6 This response will take up some of Newmeyer's critical comments in more detail. While I disagree with many of Newmeyer's observations, especially on the theoretical level, the main objective of this response is to present a more complete description of the usage-based alternative ***in answer to*** Newmeyer's objections to frequency-based analyses. This is presented in the next section. In section 3 we take up and argue against aspects of his critique of the specific case studies he cites.

2. Toward a better understanding of a usage-based theory of language

7 First of all, a terminological clarification is in order: "usage-based" and "corpus-based" are two different things. The confusion is understandable, but the distinction nevertheless remains fundamental. A usage-based model of language does not mean one looks at, or is constrained to look at, (lots of) corpus data; as explained below, it holds that grammar is ***learned*** from, and subsequently shaped by, usage (hence usage-*based*). Logically then, one cannot omit usage from grammar, but that does not mean that grammar can *only* be studied by looking at usage events as given by corpus data.

8 Newmeyer's critique raises a number of issues which he finds problematic in usage-based linguistics: the structure of grammar, the nature of usage events, negative evidence, the emergence of grammar, the role of frequency, and the place of introspection in linguistic analysis. In what follows, we will take up these different points in turn.

2.1 A structured inventory of conventionalized symbolic units

9 Newmeyer attack on usage-based grammarians and in particular their "frequency generalizations" tends to oversimplify various features of usage-based approaches.

10 First of all, in his introduction, he states the following (my italics):

> 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.*

11 At various places in the paper, Newmeyer reiterates (with some variation in the wording) the phrase in italics in the quote above (e.g., "a memorized stock of [formulas/fragments] and simple constructions"). He explicitly refers to Thompson (2000) and Thompson & Hopper (2001) who are said to adopt such an approach. However, nowhere in these papers do these authors (or other usage-based linguists, for that matter) talk about "(very) simple constructions" or of grammar ***merely*** being a stock of memorized formulas. Sure, corpus-based studies (like these two papers in question) have revealed that we do use more stock phrases than has been assumed in the (generative) literature for the last 50 years. In fact, Thompson & Hopper, unlike what Newmeyer's phrasing suggests, are quite more nuanced in their description of grammar:

> "what we think of as grammar is a complex of memories we have of how our speech community has resolved communicative problems. 'Grammar' is a name for the

> *adaptive, complex, highly interrelated, and multiply categorized sets of recurrent regularities* that arise from doing the communicative work humans do." (2001 : 48, emph. ML)

12 This description is far away from characterizing grammar as "a stock of memorized formulas and very simple constructions". In fact, the common (and oft-cited) usage-based definition of grammar is that of *a structured inventory of conventionalized form-meaning pairs* . The paper by Thompson & Hopper repeatedly clarifies that while the grammar may contain "reusable fragments", it also contains *schemas* (or *schematic constructions*); this is what the phrase "multiply categorized sets of recurrent regularities" in the quote above refers to. Even if fundamentally different in kind, these schemas do the work of "grammar rules" in a generative approach. Newmeyer (quite correctly) observes (§6) that grammatical constructions as used in conversation reveal "a sophisticated knowledge of syntax that defies any meaningful analysis in terms of 'fragments' ". Again, usage-based linguists do not claim that grammar is 'just' an inventory of stock phrases; quite the contrary, the usage-based definition of grammar is precisely that it contains both specific memorized stock phrases and schematic units, a point to which we return in section 2.4 below. Again, one of the reasons usage-based linguists emphasize the (complex) role of frequency is that it has been more than ignored in the last half century (see, e.g., Ellis (2006) for a good discussion).

2.2 Usage events

13 Newmeyer also takes issue with the focus on conversation that he attributes to proponents of usage-based theories:

> 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.

14 Again, this oversimplifies what Tomasello and other usage-based linguists are saying. The focus on conversation as the natural habitat of language (and per extension, of grammar) pertains primarily to the acquisition of grammar. As Chouinard & Clark (2003:638) state quite clearly:

> Children don't learn language in a void ; they learn it in conversation. They learn how to express their intentions and interpret the intentions of others as they use conversation to accomplish such goals as deciding what cereal to have for breakfast, getting help with a game, or finding a toy.

15 In other words, the idea is that grammar is learned in highly contextualized usage events which, at that age, typically is conversational interaction.[1] Obviously, the grammatical knowledge of adult speakers exceeds such conversational interaction. As Tomasello phrases it: "The linguistic skills that a person possesses at any given moment in time [...] result from her accumulated experience with language across the totality of usage events in her life" (2000:61). Nowhere is it said (or should it be said) that this experience be *restricted* to conversation. Surely, conversational interaction, especially for young children, is the most common habitat for language (also for adults), but it clearly is not the only kind of linguistic usage events speakers are confronted with during their life, quite the contrary. Our linguistic experience, and by extension, our grammar is shaped by a wide and varied range of usage events, such as listening to a lecture, debating, reading a novel or an academic paper, writing letters or a diary, watching a theatre play, talking to

oneself, etc. Again, the reason why usage-based linguists insist on the importance of authentic data (of which conversational data is but one type) is that such usage data has been ignored in the (generative) linguistic tradition for far too long, and even deemed irrelevant since pertaining to linguistic performance and not to linguistic competence. This has not only led to a too narrow focus on grammar and communication but also –and this is much more damaging— to ignoring the capacity of children to ***learn*** grammar (in conversational interaction).[2] Before we turn to how grammar can be learned, and what the grammar looks like within a usage-based approach, it is warranted that we briefly consider another point raised by Newmeyer regarding the advantage of introspective data over the use of corpora.

2.3 Negative evidence

16 Newmeyer voices a commonly heard plea for the use of introspective data, against the use of corpus data: "[introspective data] 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". His point here is that it is impossible to model the grammatical knowledge of the speaker on the basis of corpora, since they only provide ***positive*** evidence, i.e. what ***is*** possible for the speaker. In the last section of his paper (§7.2.1), Newmeyer makes this position even more explicit:

> No corpus can provide sentences that *do not occur*. [...] 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. (his italics)

17 This view is a nice illustration of what Stefanowitsch (2006) has called the "raw frequency fallacy", i.e. the erroneous assumption that only observed frequencies count as evidence and that, logically, corpora cannot tell us anything about ungrammaticality. However, corpora ***do*** provide negative evidence, provided one makes the distinction between accidental and significant absences. Accidental absences concern constructions that do not occur in the data sample but could have occurred, while significant absences are those that do not occur for a reason, i.e. that they are not allowed by the system. This difference can be established by (fairly straightforward) statistical comparison of what can be expected and what is observed (see Stefanowitsch's paper for a telling example on the ditransitive construction).

18 What is more, the conversational input that children are exposed to also provides them with strong negative evidence, be it of a less disruptive kind than many people have in mind. Chouinard & Clark's (2003) study nicely shows how pragmatic reasoning (the Gricean maxim of Manner) provides children with strong implicit negative evidence when adults retake and correct their "erroneous" (or rather, unconventional) utterances. This pragmatic reasoning will push children to conclude that certain (hypothetically possible) patterns are disallowed by the system. More generally, children get constant feed-back on the communicative effectiveness of their linguistic productions.

2.4 The frequency-biased learning grammar

19 While Newmeyer's position is exclusively concerned with "the construction of an adequate grammatical theory", there is a larger underlying assumption regarding the

role of negative evidence (or rather, the absence thereof) in the acquisition of (un)grammaticality in the course of language acquisition. More precisely, the absence of negative information in attested data (corpora) is mobilized as evidence that grammar and grammaticality cannot be learned on the basis of the input. This view is usually part of a larger argument about the *poverty of the stimulus*, i.e. the view that the input is too impoverished to account for our capacity to create novel sentences.[3] The poverty of the stimulus is taken to lend support to the claim that if grammaticality cannot be learned, it must be derived from some innate (i.e., pre-existing) structure, as is explicitly acknowledged by Chomsky when he says:

> Language acquisition seems much like the growth of organs generally ; it is something that happens to a child, not that the child does. And while the environment plainly matters, ***the general course of the development and the basic features of what emerges are predetermined by the initial state. But the initial state is a common human possession.*** It must be, then, that in their essential properties and even down to fine detail, languages are cast to the same mold. (2000 : 7 ; emph. ML)

20 Newmeyer is not concerned with the ***acquisition*** of grammar (or grammaticality), but as his views emanate from the same conceptions about positive and negative evidence, it is warranted that we clarify this further, to reassess the role of the input.

21 As usage-based linguists working on acquisition have shown, grammaticality ***can*** be learned on the basis of the input. We will not reproduce the argument in full here, but the gist of it is that

1. the input is not as degenerate as generative linguists assume,
2. children are doing something and language acquisition doesn't simply happen to them,
3. there is (strong) negative evidence in the input which leads children to acquire grammaticality, and
4. there are frequency biases in the input, where token frequency leads to entrenchment and type frequency, to abstraction.

22 It is precisely the interaction between type and token frequency mentioned in (iv) which determines the shape of grammar and explains where Newmeyer goes wrong in his interpretation of what usage-based linguistics means. He finds that "the idea that *a grammar* might be a stock of fragments" (his emphasis) utterly implausible. If it can be of any consolation to him, so do usage-based linguists. He reserves a strong criticism for "the more extreme usage-based linguists who put prefabs or memorized fragments at center stage and who "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". Like him, I shy away from such extremism, which would be plainly wrong and in contradiction with what usage-based linguists actually say. Langacker's rule/list fallacy, mentioned by Newmeyer, precisely points at the ***co-existence*** of grammatical rules and their instantiations, highlights that a 'stock of fragments' is not incompatible with a 'grammar'. A usage-based model of language accepts the existence of stock phrases in the grammar ***as well as*** grammatical patterns (defined as schematic form-meaning pairs). Again, one needs to distinguish type and token frequency to understand how acquiring such grammatical patterns work.

23 The token frequency of a linguistic expression captures the repeated occurrence of that expression; high token frequency leads to ***entrenchment***. In simple terms, this means that the expression is memorized as is, much like a prefab or an idiom. For example, repeated

occurrence of *work my head off* will lead to the memorisation of precisely this phrase. Type frequency concerns the occurrence of minimal variations in the input which lead to ***abstraction***, i.e. it leads to the creation of more schematic structures. (The presence of such schemas in the grammar and their role in productivity and creativity in language is explicitly referred to in the paper by Thompson & Hopper that Newmeyer criticizes.)

24 For instance, for the above example *work my head off*, variations in the body part that is referred to (e.g., *work my **ass** off*) will trigger an abstraction process leading to a more schematic structure that captures the analogy, e.g., *work my <BODY PART> off*, where one of the slots (here the one pertaining to the body part) is (somewhat more) open.[4] This is essentially a categorization process. The higher the type variation, the more open the particular slot; that is, its criterial features will become more general, which paves the road for novel items to occur in it. Similar type variations will also affect the verb slot in the expression, e.g., ***laugh** one's head off*, ***dance** one's head off*, ***bawl** one's head off*, etc.). This leads to further schematization that is partially schematic and partially filled: *V one's BODY PART off* (note that *one's* also represents a schematization here). Even the particle slot can become amenable to schematization, e.g., *cry one's eyes **out***). This will eventually lead to a fairly schematic pattern, e.g., *V POSS BODY-PART PRT* which sanctions instances like *He worked his head off* or as *They cried their eyes out*. Jackendoff calls such patterns "constructional schemas" where he says that "Pieces of syntactic structure can be listed in the lexicon with associated meanings, just as individual words are" (2008:15).

25 As shown here, these schematizations are the outcome of analogical reasoning and they do come with their semantic and syntactic restrictions derived from the observed analogies. As far as semantics is concerned, the overall meaning of both the schema and the different instantiations remains fairly constant: 'V to a high degree'. As to the syntactic constraints, the subject is, as a rule, coreferential with the possessive determiner (**He$_i$ worked their$_j$ head off*), even if this is not a categorical restriction: if someone's 'V-ing to a high degree' can be seen as affecting someone else, this can be easily overruled, e.g. *Eight, nine months ago, he traps me in the Marquee, **talks my head off**. He's going on about nothing and everything* (COCA, Fiction corpus, 2014). These grammatical structures thus come with constraints derived from their contexts of use. In other words, the grammatical patterns are derived from usage, but once they are in place, they will constrain what is possible in the language. As Bybee puts it, "Usage feeds into the creation of grammar just as much as grammar determines the shape of usage." (2006:730).

26 At the same time, such schemas lead to productivity (or creativity), as new instances can be built when needed. For example, the schema above allows for novel instances, such as *I typed my fingers off taking notes during the lecture*. In such novel uses, there should be an interpretable link between the action denoted by the verb and the body part (out of context, *I worked my fingers off* is hard to figure out). Strikingly, this requirement is less strong for the more conventional patterns *V one's head off* and *V one's ass off* for which any verb seems to be possible. Nevertheless, a quick corpus search on COCA[5] shows that there is less variation than one might assume. For *V one's head off* (145 instances), the lion share of the verbs (142 instances or 98%) are communication verbs (*scream*, *laugh*, *bark*, *talk*, *yell*, *shout*, *bawl*, *lie*, *cry*, *giggle*, *sing*, *moan*, *howl*, *chat*, *shriek*, *holler*) or verbs that can be perceived as such (*snore*, *query*); the only other verb attested is *work* (3 occurrences). For *V one's ass off* (178 attestations), the distribution is different: only 28% of them (50 verbs) are communication verbs, the rest (128 or 97%) are action verbs of various sorts (e.g., *work*, *run*, *dance*, *play*, *sweat*, *flirt*, *write*, etc.).

27 In sum, the 'mini-grammar' for the semi-open construction *V one's BODY PART off* presents a schematic network of semantically and formally related structures, and the network contains both schematic structures as well as specific instances. This means that there is no more strict distinction between lexical items (lexicon) and 'rules' (syntax). Viewing the grammar as a structured schematic network in fact gives you the best of both worlds: it combines grammatical productivity with item-specific preferences. These preferences can be caused by differences in frequency, but not necessarily, as frequency is not the only causal factor giving rise to such preferences.

28 Generative linguists might object that this is not really grammar as they see it, a set of procedural or algebraic rules that generate all possible sentences of a language and only those, but such objection would miss the point entirely. For clarity of exposition of how type frequency works, we have taken the example of the semi-open construction *V one's body part off*; however, similar operations of abstraction are at work for all other possible constructions in the grammar. At the highest level of abstraction, one finds highly schematic constructions such as the past tense construction (*V-ed*), the passive construction, the ditransitive construction (NP V NP NP, e.g., *I gave John the book*), the caused motion construction (*He pushed the cart into the barn)*, or the intransitive motion construction (*He walked into the classroom*). These highly schematic constructions allow for much creativity within the confines of their syntactic and semantic constraints.

29 The effects of type and token frequency on the structure of grammar can thus be summarized as follows. High type frequency leads to abstraction which increases productivity, since it reduces the association of the pattern with a particular lexical item and consequently also loosens up its criterial features that constrain items occurring in this slot. High type frequency thus facilitates novel uses. In more technical parlance, what children actually learn is "pattern extraction mapped to communicative intent, with generalisation to more schematic representations" (Lieven 2014). Conversely, high token frequencies of particular instances of a given schema will lead to entrenchment, the (structured) storage of these instances within the schematic network. Unless they occur with extremely high frequency, it is very unlikely that these stored instantiations are truly frozen stock phrases; such may be the case for example for *I love you* or *I dunno* which not only occur extremely frequently but often also have a particular communicative function. Mostly, the stored phrases remain somewhat open and variable, as in our example *work one's head off* where the verb and possessive can vary in form depending on the grammatical context, etc.

30 In short, when Newmeyer says: "It is not a matter [...] of 'fragments' or 'formulas', but rather of a sophisticated engine representing grammatical knowledge", he ignores the sophisticated grammatical knowledge that a usage-based grammar presents, which is actually even more powerful than the generative 'engine' that he has in mind, as it can also deal with sentences that extend the schematic constraints (in generative terms: that violate the grammar rules), as illustrated by, e.g., *Australians drink pubs dry after World Cup defeat to France*[6] (violation of selection restrictions), *The audience laughed him off the stage* (transitive use of intransitive *laugh*), or *There is too much apple in the cake* (transformation of count noun to mass noun).[7] Again, generative linguists might argue that some (if not all) of these extensions are not due to grammar, but to lexical operations, e.g., a metonymical extension from object (apple) to the object's substance. That may be so, but the fact remains that it stretches *grammatical* constraints for which a lexical account often does not work. Generative linguists might also object that these are spurious or

idiosyncratic extensions, often with a deliberate humorous effect. However, when one starts looking at lots of data (as Newmeyer suggests himself) such 'idiosyncracies' turn out not to be all that spurious as one may have assumed (see also Hilpert 2014:Chapter 1 for an interesting discussion).

31 Despite the (highly) simplified nature of the above summary, it should be clear that children, operating on the input via analogical reasoning and categorization, build a grammar that eventually allows them to be linguistically creative. As already indicated above, via low disruptive negative evidence provided by adult retakes of their unconventional utterances, children acquire grammaticality. Frequency biases in the input further strengthen their notion of grammaticality via statistical pre-emption, the repeated experience of an item in a competing construction or pattern. This principle of probabilistic learning is what Tomasello refers to when he says: "children not only say what they hear, but the more they hear it, the more it seems to them that this is the only way it can be said" (2000: 72). The idea is that (high) token frequency (and positive evidence in general) not only leads to the entrenchment of a given structure, but also to the discouragement of potential alternatives.

32 At the end of his article, Newmeyer rejects the input as a possible basis for grammatical competence, showing (via a transcript of an exchange from his corpus) how conversation is unbelievably messy, with incomplete sentences, sidetracks, insertions, interruptions, restarts, etc. Usage-based linguists do not deny that conversations are messy, quite the contrary. However, the example given by Newmeyer is a conversation by expert speakers (adults) with years of training behind them. They have mastered the communicative skills that underlie such conversations (including managing multiple tracks within one and the same conversation). Children do not yet possess these advanced skills. As studies in language acquisition have shown (see Chouinard & Clark 2003 and the references therein), the input that children are exposed to when they are in the full process of learning language is quite error-free, well-structured, and –importantly–adapted to their level of understanding.

2.5 Frequency effects on the structure of grammar

33 In a usage-based approach, "grammar is the cognitive organization of one's experience with language" (Bybee 2006:711). There is ample research that shows that frequency of experience with particular constructions (or instances of constructions) does have an impact on this cognitive organization. Given the frequency biases in the input, speakers are not only very much aware of what is conventional and/or frequent in their language, their linguistic behaviour also shows to be influenced by these biases (see, e.g. Kuyper 1996 Goldberg *et al.* 2006, or work by Elissa Newport and colleagues[8]). Newmeyer himself admits that this is so: "frequency is an important factor leading to the shaping and reshaping of grammar." He uses this to say that "appeals to frequency should never be used as a substitute for careful grammatical analysis". I couldn't agree more! He then goes on to link this up with introspective judgments: "Frequency generalizations derived from conversational corpora do not challenge theories constructed on the basis of introspective judgments." This a conclusion I disagree with, as there are clear arguments to say that introspective-based theories have been challenged by such frequency-based studies.

34 First of all, Newmeyer contradicts himself on this point. He cites corpus-work by Bresnan *et al.* that corrected "extravagant claims about what is supposedly not found in ordinary usage" that had been made by syntacticians using only introspective data (his observation). If introspective judgments of well-formedness are at the heart of the model that Newmeyer advocates, how can such correction by corpus data not be seen as challenging existing theories or claims? I would expect Newmeyer to agree that, independent of the theoretical framework, appeals to introspection should never be used as a substitute for careful grammatical analysis.

35 A second reason why I disagree with Newmeyer that frequency analyses of corpus data have not challenged theories constructed on the basis of introspective judgements is that they have confirmed the existence of semi-open constructions (such as *V one's <body part> off*, *the X-er the Y-er*, *What's X doing Y*, etc.) with their idiosyncratic constraints and preferences. As indicated above, these findings suggest that a strict dichotomy between lexicon and syntax is untenable (which is, however, a point that Newmeyer's theory continues to uphold). How could one not consider this as theory-challenging or calling for a (drastic) change of model?

36 Even more fundamentally, if it ***is*** possible to learn grammar and acquire a deep sense of grammaticality on the basis of the input, a process in which frequency plays a complex but essential role, why would one want to continue postulating an innate, universal grammar wired into the human genome? This question is all the more pressing in view of the fact that assuming grammar to be a common human possession creates a dualism in the model that Newmeyer seems to defend, since one part of the language (the grammatical core, containing abstract regularities) is innate, while another part (the lexicon, containing the idiosyncratic) is learnt.

37 Finally, there is an important point to make about diachronic change and its relation to the grammar model (language phylogeny). Newmeyer accepts that "[f]requency of use [...] is uncontroversially an important factor in directing grammatical change". He himself cites examples, such as the grammaticalization of locative nouns to adpositions, pronouns to person markers, auxiliaries to tense and aspect particles, and so on. Even if these changes often affect grammatical structures quite drastically, Newmeyer rejects the idea that this should lead to a fundamental change of how we should see grammar: "I certainly do not see anything there that would challenge standard models of grammar." Elsewhere, Newmeyer (2003:698) has made this point even more strongly:

> One also has to take issue with the view, expressed so often by advocates of usage-based grammar, that grammars are fragile, fluid, temporary objects. [...] In my view, one of the basic things to explain about grammars is their stability, at least where there is no significant language contact to complicate things. [...] I suspect that we could carry on a conversation with Shakespeare, who lived four hundred years ago. And the problems we would have with him would more likely be lexical and low-level phonological rather than syntactic. Preposition stranding survives, despite its being functionally odd, typologically rare, and the object of prescriptivist attack for centuries.

38 In this earlier publication, Newmeyer reacts against Bybee and Hopper's view that "mental representations are seen as provisional and temporary states of affairs that are sensitive, and constantly adapting themselves, to usage" (2001:2). The mental representations are those that result from the fixing of linguistic groups as structural units. Newmeyer interprets this to mean that "normal human languages are not any different from trade pidgins [...] where there are hardly any rules and communication is

largely based on world-knowledge and context" (2003:698). The above account of usage-based grammar should suffice to understand that the comparison with pidgin does not hold; grammatical patterns (schemas) do get fixed in grammar, also through usage. Usage-based linguists see grammar as an adaptive, co-evolving system of mappings of patterns with communicative intent.[9] In fact, it is again (extreme high) frequency of use of such meaningful grammatical patterns that can be an explanatory factor for the stability of grammar over time, generation after generation.[10] Preposition stranding may be such a pattern that 'survives'; frequency may play a role in this, but its salience may also be due to the fact that it 'stands out' structurally (preposition without its object) or intonationally (stress on preposition which is typically unstressed). The last point is quite essential: frequency is an undeniable structuring factor for grammar, but it is not the only one. Again, a usage-based model gives you the best of two worlds, as it combines structural stability with adaptability.

2.6 Introspection and (un)grammaticality

39 A final point concerns Newmeyer's defense of introspective judgement of well-formedness as a reliable source for modelling the mental grammar; this is part of his critique of grammar being characterized as a 'stock of fragments' (which we have argued to be an incorrect interpretation of what usage-based linguists actually say). He observes that "[i]nterpreting novel strings and making judgments of well-formedness require computational ability — that is, they require a grammar." As explained above, usage-based theories do not deny the existence of grammar, quite the contrary. A usage-based model of language ***does*** contain grammatical patterns, defined as schematic form-meaning pairs, which speakers mobilize to interpret and produce novel strings. However, unlike Newmeyer, I truly doubt that making of judgments of well-formedness is what speakers do when they hear novel utterances. Rather, what they do is to ***interpret*** these novel sentences (as indicated in the first part of Newmeyer's observation) to understand their meaning and the communicative intent of the interlocutor. In other words, speakers naturally figure out the meaning of sentences (be they novel or not), but as a rule they do not go around judging whether the sentences they hear are grammatically correct or not. It seems to me that Newmeyer comes very close here to conflating grammaticality and acceptability here, a distinction which for Chomsky (1965) was essential and to be related to, respectively, competence and performance.

40 Newmeyer is absolutely right in saying that ungrammatical sentences will continue to play an important part in linguistic analyses and thus introspection will always have its place in linguistics. However, its role is fundamentally different in the generative and the usage-based perspective. In Newmeyer's approach, introspection is used to build the model; this means that a single counterexample is thus in principle enough to invalidate the model. In a usage-based model, introspection (or ungrammaticality) is used not to prove the model but to *explain why* a given construction may not occur. For example, a *to*-dative construction with *know* (e.g., *He knew it to me*) may not occur because its meaning (the merger of the semantics of *know* and that of the *to*-dative construction) would simply be uninterpretable. This relates to Goldberg's (2006:39-40) *semantic coherence principle* which holds that lexical items and constructions can only merge if their semantics are compatible. Another factor may be conventionalized choices; for example, causative events involving bodily processes cannot be expressed in English with a lexical causative

(**I {laughed/cried/bled/sweat/sneezed} him*) but require a periphrastic causative (*I made him {laugh/cry/bleed/sweat/sneeze}*). While one can come up with a semantic motivation for this preference (e.g., the semi-autonomous nature of the bodily process), it still remains, to a large extent, a conventionalized choice.[11]

2.7 Interim summary

41 In sum, in many ways, Newmeyer's position is quite close to a usage-based account since he insists, as do many usage-based linguists, on the proper linguistic analysis of the data. The issue is whether Newmeyer would still object to the usage-based approach to language as described above, which holds that grammatical structures exist and that they can be learned from the input (via regular cognitive operations that humans also mobilize in other areas). A model which consequently does away with the idea of the poverty of the stimulus but preserves the idea of (acquired) grammaticality. A model which does away with the dualism inherent to the nativist perspective that holds that part of language is acquired and part is innate. A model in which frequency, interpreted correctly, does play an important role in structuring the inventory of conventional linguistic units that grammar is said to be.

42 A large part of Newmeyer's critique has to do with the use of corpora, in particular those that are too small to lead to valid conclusions about grammar. He thus calls for larger corpora. As said, we couldn't agree more! At the same time, the analysis of corpora, in particular if they get very large, is more complicated that Newmeyer seems to be aware of. In his paper, he critically comments on a number of specific corpus studies which he counters with his own corpus analysis. However, his criticism is not always without problems; the next section takes up a few of those problems.

3. Newmeyer's critique of specific corpus-studies

43 As already indicated, Newmeyer's paper is somewhat ambiguous, since, first of all, he criticizes the claims made by (usage-based) corpus studies by drawing on a corpus himself which then is taken to show the importance of introspection-based data. Newmeyer argues in favour of introspection data which also "allow for the easy removal of irrelevant data, such as slips of the tongue, false starts, etc."; at the same time, he calls for using larger corpora. However, using larger corpora means that manual scrutiny of the data becomes less feasible, and thus there is no more "easy removal" of unwanted or irrelevant data.[12] His call for large corpora is thus quite surprising, in view of his insistence on introspection data but also in view of his criticism of frequency-based analyses. Besides this rhetorical contradiction in Newmeyer's reasoning, some further comments are nevertheless in order with respect to issues related to corpus analysis.

44 The first two comments are methodological. Newmeyer counters findings from other studies by drawing on a corpus himself, the Fisher English Training Transcripts or the Fisher corpus for short. However, we are not provided with comprehensive meta-information on this corpus: What time period is covered by the data? Who are the speakers (region, age, social class, etc.)? What is the topic of conversation? etc. Without this information, Newmeyer's claims are as unqualifiable as the ones he criticizes.

45 A second methodological comment is that we have no information about how the data was extracted from this corpus. Moreover, Newmeyer merely uses this corpus to find some counterexamples, an approach that is typical of the introspection-based modelling. That is, his method is more 'corpus-illustrated' than 'corpus-based'. Corpus analysis, at least if meant to show for language modelling, analyses the corpus data exhaustively, which Newmeyer does not do. When evaluated in this way, corpus evidence has the merit of allowing robust generalisations which are not invalidated by a single (i.e., marginal) counterexample.

46 How this can lead to inaccurate analyses is illustrated in Newmeyer's criticism of Biber's work on linguistic variation in genre; Biber has suggested a list of 67 linguistic features that would allow one to identify the genre of a text. Newmeyer singles out three of these features (present participle constructions, past participle constructions, and split infinitives) that, according to Biber, occur least in face-to-face conversations (less than 0.1 times per 1,000 words). Newmeyer counters: "it was not difficult to find examples of all three in the Fisher corpus" and he lists three such counterexamples, one for each construction.[13] Newmeyer follows a line of reasoning typical of categorical introspection-based modelling, in which a single occasional counterexample suffices to invalidate the model. This is not at all what Biber's model says, which is concerned with frequency of occurrence of features (or their combined occurrence) as indicative of certain genres; saying that particular constructions occur less than 0.1 times per 1,000 words is not making a categorical statement. Newmeyer is not entirely fair in his counterargument, as he does not quantify his findings nor do we know whether the Fisher corpus is at all reliable to counter Biber's claims. Finally, and more dramatically, upon qualitative inspection of Biber's examples of participial constructions (examples (1a,b) below) and the counterexamples from the Fisher corpus given by Newmeyer (examples (2a,b) below), one cannot be but struck by an intuitive evaluation of style difference, Newmeyer's counterexamples being more typical of spoken language.

(1)	a	stuffing his mouth with cookies, Joe ran out the door (Biber 1988)
	b	built in a single week, the house would stand for fifty years (Biber 1988)
(2)	a	having angst i don't have any like firsthand experience with separations or anything cause i mean
	b	but compared to the comedies now it it's tame

47 The examples from Newmeyer's spoken corpus are built with highly frequent forms (*having X* and *compared to X*) which seems less so for Biber's examples. For instance, in the spoken part of the COCA (118 million words), there are only ten attestations of *stuffing* used in a present particle construction as the one above. In addition, eight of these are of the form *stuffing X into Y* and only two cases of *stuffing X with Y*. In other words, there are most likely still striking differences across genres in both form and frequency of these constructions which, however, Newmeyer chooses to ignore, since he has pitched the discussion at the highest possible level, that of the type of construction.[14]

48 As said, Newmeyer uses his corpus data mainly to find counterexamples but often does not really present a full-fledged analysis. This is quite clear in his discussion of anaphoric

relations (§4), where he says that "one often hears that cataphors (i.e., backwards anaphors) only occur in linguists' introspective judgments or possibly in educated speech or writing". This section merely presents a number of counterexamples, without any further analysis regarding their internal complexity or conventionality. The only thing he says is that "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." It is not clear how not yet knowing much about these constructions has anything to do with usage-based or corpus-based linguistics.

49 Another critique by Newmeyer is targeted at Miller and Weinert's (1998) study of the grammatical properties of English conversation. Newmeyer's critique is that their work does not mention 10 types of constructions (listed under (5)) which, however, all occur in the Fisher corpus. Ignoring the 'single counterexample proves you wrong' fallacy, Newmeyer's observations are wrong on yet another ground, i.e. comparing two different kinds of corpora. Miller & Weinert have actually used four different corpora for English, which have specific features: (i) they are relatively small (taken together about 200,000 words in total)[15], (ii) they mostly contain conversations by young speakers (primary school children and adolescents), and –importantly—(iii) they all pertain to Scottish English. It is thus only to be expected that the Fisher corpus (which we understand to contain American English adult conversations) contains a different range of constructions than a corpus largely made up of Scottish English adolescent speech.

50 Newmeyer's substantially edited version of their conclusions, described as being about "the bankruptcy of formal linguistic theory", distorts their words into a strong anti-Chomskyan statement:

> [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.

51 They actually said the following (with apologies for the longer quote which, however, is essential to our point; the italics highlight the passage corresponding to the quote above):

> The properties and constraints established over the past thirty years by Chomskyans are boundary conditions in the sense that no sentential construction will break through the boundaries and remain acceptable ; in fact, sentences breaking through the boundaries occur neither in speech nor in writing. But the examples that come closes to the boundaries all occur in writing, and many of these occur as examples in work of syntax. As their first language children do no acquire the written variety of their native tongue but the structures and vocabulary that they hear in the spontaneous speech around them. Anticipating the discussion below of possible mechanisms of acquiring first language, we ask at this point whether it makes sense to think of children as acquiring a grammar that will generate all and only the correct sentences of their native language, including its written variety, when their task is to develop rules for the structures of spontaneous spoken language.

52 In these statements (which figure in the conclusion to their book), they carefully situate their work in response to Chomskyan linguistics. For one, they observe (in the part preceding the above quote) that the intuitions on which the Chomskyan linguistic model is based are largely those of a small groups of (highly) educated individuals working with language professionally (linguists, graduate students, etc.) and cannot be held representative of the grammar of (all) speakers of English.[16] Secondly, their point in the above quote is not at all a statement about "the bankruptcy of formal linguistic theory" but is a careful critical reflection on the status of the properties and constraints of

written language (on which most formal theories are based) in view of the fact that "children do not acquire the written variety of their native tongue but the structures and vocabulary that they hear in spontaneous speech around them". In other words, if one reads their work carefully, one sees that their point is not so much to present a comprehensive account of all English constructions possible in conversation, but rather to show that the constructions that they find in their spoken data differ quite strongly from those discussed in the generative literature which quite legitimately leads one to question the validity of the latter, in particular in view of language acquisition. This resonates quite well with the usage-based views on emergent grammar. This shows how Newmeyer's interpretation needs to be qualified.

53 In sum, Newmeyer's criticism of the different studies is unconvincing and his own corpus analysis is methodologically unclear and selective. Its main purpose seems to be to present counterexamples, to salvage the introspection-based method as a basis for the construction of a (theory of) grammar.

4. Concluding remarks

54 Newmeyer is quite sharp in his rejection of the studies by usage-based linguists who argue for the importance of frequency-based generalizations to the structure of grammar. His critique misses the point, however, as there are quite some inaccuracies in his understanding and rendering of the assumptions of usage-based linguistics. First of all, he incorrectly characterizes the usage-based model of grammar as a "stock of fragments" and "very simple constructions". This not at all how usage-based linguists define grammar. They characterize it as a structured inventory of form-meaning pairs, which contains both fixed or semi-fixed expressions and grammatical constructions. For both, frequency plays an important role: token frequency for the former, type frequency for the latter. Newmeyer entertains a too simple view on the role of frequency in the construction of grammar, which is much more than simply counting occurrences and its effects cannot be countered by a single counterexample. Grammatical structure can be learned from the input which does not, contrary to what many people may still think, involve simplistic behavioristic learning, but results from abstracting frequency-biased regularities of the form-meaning mappings in the input. Implicit negative evidence and statistical pre-emption further conspire to deeply engrained knowledge of grammaticality. Obviously, frequency does not explain everything in grammar. Factors such as contextual salience, semantic simplicity, relevance or pragmatic practice also play an important role in the acquisition of form-function mappings.

55 Newmeyer holds a different view of how or on what basis grammar should be modelled. He argues that corpora do not provide a better basis for grammatical theory than does introspection-based reasoning, unless these corpora are sufficiently large. While I wholeheartedly second his appeal to use such larger conversational corpora, I take issue with his critique of the usage-based approach as it emerges from this paper, since his recount is partial and in the service of his own views. In addition, his own use of corpus data is merely to provide counterexamples (the extent of which we do not know) and he does not provide the careful grammatical analysis that he himself calls for.

56 This response is an invitation to him to reconsider what usage-based linguistics is and to acknowledge that it can do the 'grammar work' that he unjustifiedly claims it cannot.

Surely, many intriguing questions remain about how speakers handle (complex) grammatical structures and much more empirical work (of various sorts) is needed to answer those. Analysing bigger conversational corpora is certainly one laudable way to go, as long as it is done carefully, with proper grammatical analyses. Independent of which theoretical framework we adopt, let us, above all, continue to do just that.

BIBLIOGRAPHY

Bybee, Joan L. 2006. From usage to grammar: The mind's response to repetition. *Language* 82: 711–733.

Bybee, Joan L., & Paul J. Hopper. 2001. Introduction to frequency and the emergence of linguistic structure. In Joan L. Bybee & Paul Hopper (eds.), *Frequency and the emergence of linguistic structure*, 1-24. Amsterdam: John Benjamins.

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

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax.* Massachusetts: M.I.T. Press.

Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*. Cambridge, Ma: Cambridge University Press.

Chouinard, Michèle & Eve Clark. 2003. Adult reformulations of child errors as negative evidence. *Journal of Child Language* 30: 637-669.

DĄbrowska, Ewa. 2008. Questions with long-distance dependencies: A usage-based perspective. *Cognitive Linguistics* 19: 391-425.

DĄbrowska, Ewa. 2010. The mean lean grammar machine meets the human mind. In Schmid, Hans-Jörg & Susanne Handl. *Cognitive Foundations of Linguistic Usage Patterns*: 151-170. Berlin : Mouton de Gruyter.

Divjak, Dagmar, Ewa DĄbrowska & Antti Arppe. 2016. Man Meets Machine. Evaluating the psychological reality of corpus-based probabilistic models. *Cognitive Linguistics* 27 (1): 1-34.

Ellis, Nick. 2006. Frequency effects in language processing. *Studies in Second Language Acquisition* 24: 143-188.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. Chicago: Chicago University Press.

Goldberg, Adele E. 2006. *Constructions at work. The nature of generalization in language*. Oxford: Oxford University Press.

Herbst, Thomas, Hans-Jörg Schmid, & Susen Faulhaber (eds.) 2014. *Constructions, Collocations, Patterns.* Berlin: Mouton de Gruyter.

Hilpert, Martin. 2014. *Construction grammar and its application to English*. Edinburgh: Edinburgh University Press.

Hilpert, Martin. 2015. From *hand-carved* to *computer-based*: Noun-participle compounding and the upward strengthening hypothesis. *Cognitive Linguistics* 26(1): 113-147.

Jackendoff, Raymond. 2008. Construction after construction and its theoretical challenges. *Language* 84: 8-28.

Langacker, Ronald W. 1987. *Foundations of cognitive grammar*. vol. *I: Theoretical Prerequisites.* Stanford: Stanford University Press.

Langacker, Ronald W. 1991. *Foundations of Cognitive Grammar*. vol. *II: Descriptive application*. Stanford: Stanford University Press.

Lauwers, Peter & Dominique Willems. 2011. Coercion: Definition and challenges, current approaches, and new trends. *Linguistics* 49(6): 1219-1235.

Lemmens, Maarten. 2017. A cognitive, usage-based view on lexical pragmatics. A reply to Hall. In I. Depraetere & R. Salkie (eds.), *Semantics and Pragmatics: Drawing a Line*, 101-114. Switzerland: Springer.

Lemmens, Maarten. 2015. Cognitive semantics. *Routledge Handbook of Semantics.* Editor: Nick Riemer, London & New York: Routledge, 90-105.

Lemmens, Maarten. Forthc. *Usage-based perspectives on lexical and constructional semantics*. China: Shanghai Foreign language University Press.

Levinson, C. Stephen. 2003. Language and Mind: Let's Get the Issues Straight! In Gentner, Dedre & Susan Goldin-Meadow (eds.), *Language in Mind. Advances in the Study of Language and Thought*, 25-46. Massachusetts, CA: MIT Press.

Lieven, Elena. 2014. First language learning from a usage-based approach. In Herbst, Thomas, Hans-Jörg Schmid, & Susen Faulhaber (eds.), *Constructions, Collocations, Patterns*, 9-32. Berlin: Mouton de Gruyter.

Miller, Jim & Regina Weinert. 1998. *Spontaneous spoken language: Syntax and discourse*. Oxford: Clarendon.

Newmeyer, Frederick J. 2003. Grammar is grammar and usage is usage. *Linguistic Society of America* , 79(4): 682-707.

Perek, Florent & Maarten Lemmens. 2010. Getting at the meaning of the English *at*-construction: the case of a constructional split. *CogniTextes* 5, http://cognitextes.revues.org/331

Perek, Florent. 2015. *Argument structure in usage-based construction grammar : Experimental and corpus-based perspectives*. Amsterdam: John Benjamins.

Romain, Laurence. 2017. Measuring the alternation strength of causative verbs: a quantitative and qualitative analysis of the interaction between verb, theme and construction. *Belgian Journal of Linguistics* 31, 213-235.

Schmid, Hans-Jörg & Susanne Handl. 2010. *Cognitive Foundations of Linguistic Usage Patterns*. Berlin: Mouton de Gruyter.

Stefanowitsch, Anatol. 2006. Negative evidence and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 2, 61-77.

Street, James & Ewa DĄbrowska. 2010. More individual differences in Language Attainment: How much do adult native speakers of English know about passives and quantifiers? *Lingua* 120, 2080-2094.

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*, 27-60. Amsterdam: John Benjamins.

Tomasello, Michael. 2000. First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics* 11: 61-82.

NOTES

1. I prefer using the term *conversational interaction* as it covers a wider range of linguistic usage events than those typically associated with the term *conversation*, such as (interactive) storytelling.

2. While in the last decades, more attention has been paid to spontaneous spoken interaction, there still remains a strong bias towards written language in most linguistic work. This bias is understandable, as written data (most of which is now readily available in electronic form at a simple mouse click) is so much easier (and cheaper!) than oral data, which requires much more effort and funding, both in collecting and subsequently transcribing the data. In recent years, also the need for a multimodal analysis which includes, e.g., gesture or facial expressions in linguistic analyses, has come to complicate data collection and analysis even more.

3. Chomsky describes the poverty of the input as "the degenerate quality and narrowly limited extent of the available data" (1965:58).

4. See also Lemmens (2015, 2017, forthc.) for related discussions on such schematization processes applied this and other examples.

5. *Corpus of Contemporary American English*, available at https://corpus.byu.edu/coca/ (consulted via institutional login from Université de Lille, France). The search pattern used was *VERB _app* head off* (*_app** stands for possessive pronoun or determiner) which also returned constructions like *chop his head off* which have been manually deleted from the extractions. I hasten to add that the example is merely used for illustrative purposes, and not a full-fledged corpus analysis.

6. Adapted from https://www.newshub.co.nz/home/world/2018/06/australians-still-in-good-spirits-drinking-pubs-dry-after-world-cup-defeat-to-france.html, last accessed Nov. 25, 2018

7. This is sometimes referred to as "coercion", see Lauwers & Willems (2011).

8. For some relevant publications by E. Newport and her team on the role of input frequency on learning, see http://cbpr.georgetown.edu/faculty/elissa_newport.

9. See Levinson (2003) for a more semantically oriented discussion of language as part of a gene-culture coevolution.

10. For an innovative discussion of how abstract grammatical patterns may get entrenched, see Hilpert (2015) on the "upward strengthening hypothesis".

11. We may add that these conventionalized choices are language-specific; other languages may choose to encode these events differently. One cannot help being surprised by the fact that Newmeyer feels the need to point out that "the frequent use of

a construction type in one language is not necessarily a reliable guide to what occurs crosslinguistically". This is quite obviously true in a usage-based approach which defines grammar as a *conventionalized* inventory of symbolic units. In Newmeyer's model, which embraces a universal view on grammar, such crosslinguistic differences are only "superficial appearances" (Chomsky 2000:7).

12. The presence of unwanted or irrelevant data is in fact a false counterargument against the use of (larger) corpora, since unless this "noise" is structural (e.g., systematic and numerous mismatches due to an incorrect query), the proper statistical tests will correct for this. And contrary to what Newmeyer believes, bigger is not always beautiful, as at some point the corpus will reach a saturation point and statistical tests risk becoming unreliable giving false positives, since the discrepancies between the distributional frequencies being compared become too large.

13. The criticism of Biber's work aims at countering the claim that introspection "leads to sentences that are confined to a large degree to literary genres". Strikingly, this claim is posited as a given, without any further evidence or source. However, it is not at all clear why that would be so.

14. In all fairness, much early work in usage-based linguistics (notably, in Construction Grammar) is pitched at a too schematic level, being concerned with, e.g., the ditransitive construction, the caused motion construction, etc. More recent work (see Perek & Lemmens 2009, Perek 2017, Romain 2017) makes a case for more low-level schematisations, especially for argument structure constructions. See also DĄbrowska (2008, 2010) for psycholinguistic evidence on how low-level schemas for long dependency WH-question are being favoured by speakers at the expense of more general schemas.

15. Newmeyer incorrectly reports that their corpus only contains 50,000 words, mentioning only one of the four corpora they have used.

16. See recent psycholinguistic work by Street & DĄbrowska (2010) on differences in ***grammatical*** proficiency as correlating with degree of education.

ABSTRACTS

In his paper "Conversational corpora : when big is beautiful", Newmeyer sets himself the goal of evaluating the relationship between corpus size and conclusions drawn from corpora regarding questions of grammatical theory. He formulates a strong critique against corpus research based on too small (conversational) corpora and in doing so, explicitly rejects the usage-based approach to language in which they are embedded. He argues that, unless they are based on large (conversational) corpora, frequency analyses do not give sufficiently reliable analyses compared to introspection-based analyses. In this response, I will counter some of the critique that Newmeyer levels against usage-based (or frequency-based) models, showing that, first of all, his criticism needs to be reevaluated and secondly, frequency-based analyses (and a usage-based approach more generally) do imply a radically different view on grammar which surpasses some of the shortcomings of introspection-based models.

Dans son article « Conversational corpora : when big is beautiful », Newmeyer se donne comme objectif d'évaluer la relation entre la taille des corpus et les conclusions que l'on peut en tirer en termes de théorie grammaticale. Sa critique appuyée de recherches menées sur des corpus (conversationnels) de taille limitée l'amène à rejeter plus largement le cadre dans lequel ces recherches s'inscrivent, en l'occurrence, les approches de la langue fondées sur l'usage. Il défend notamment l'idée que les résultats d'analyses de fréquence fondés sur des corpus trop petits sont moins fiables que les analyses fondées sur l'introspection. Dans la présente réponse, j'oppose des contre-arguments à la critique que Newmeyer fait des modèles fondés sur l'usage (ou la fréquence). Je montre d'abord que ses critiques doivent être réévaluées, puis que les analyses fondées sur la fréquence (et, d'une manière plus générale, sur l'usage) impliquent une conception radicalement différente de la grammaire, qui dépasse certaines limites des modèles fondés sur l'introspection.

INDEX

Keywords: usage-based linguistics, frequency-based analysis, corpora, introspection, negative evidence
Mots-clés: linguistique de l'usage, analyse fondée sur la fréquence, corpus, introspection, preuve par la négative

AUTHOR

MAARTEN LEMMENS

Université de Lille & UMR 8163 STL « Savoirs, Textes, Langage », CNRS, France