Anatol Stefanowitsch and Susanne Flach

5 The Corpus-Based Perspective on Entrenchment

5.1 Introduction

Entrenchment is a fundamentally cognitive notion, referring to the degree to which a linguistic structure of any degree of complexity or schematicity forms an established unit of the mental grammar of a speaker. It may therefore seem that linguistic corpora, which are essentially large samples of usage data, have no particular role to play in estimating degrees of entrenchment of linguistic structures. Instead, psycholinguistic methods may be perceived as more suitable because they appear to measure the phenomenon more directly.

However, this perception is mistaken for two reasons. First, entrenchment is a theoretical construct, a hypothesized aspect of mental representations. As such, we cannot directly measure degrees of entrenchment, or even verify the existence of entrenchment, by any currently available means. Regardless of whether we attempt to measure entrenchment experimentally or on the basis of corpora, we must rely on operational definitions that approximate the theoretical construct but are not identical with it. In Section 5.2.1, we sketch ways of conceptualizing the relationship between mental representations and usage data that allow the corpus-based operationalization of the former on the basis of the latter.

Second, entrenchment as a theoretical construct has always been defined with respect to what we might loosely refer to as frequency of occurrence. Thus, it seems not only possible but obvious to turn to corpus-based operationalizations to investigate (degrees of) entrenchment. In Section 5.2.2, we discuss this issue in more detail and touch on the relationship between corpus-based measures of entrenchment on the one hand and experimental ones on the other.

Although at first glance raw frequencies of occurrence may seem to be a straightforward way of measuring entrenchment, matters are slightly more complicated. Linguistic units may differ along at least two dimensions, complexity and schematicity, and simple frequency counts are useful approximations of entrenchment only in the case of minimally complex and minimally schematic expressions. In Section 5.3.1, we discuss complexity and schematicity in more detail. In Section 5.3.2, we discuss established corpus-based measures of entrenchment for different types of units. Specifically, we deal with simple, nonschematic units in Section 5.3.2.1, with different perspectives on complex, nonschematic units in Section 5.3.2.2, and with schematic units in Section 5.3.2.3.

DOI 10.1515/9783110341423-006

Although the relationship between corpus frequency and entrenchment is overwhelmingly taken for granted in the cognitive linguistic research literature, it has recently been questioned. Section 5.4 addresses the major points of criticism.

5.2 Corpora, Cognition, and Entrenchment

5.2.1 Corpora and Cognition

It is fairly uncontroversial that linguistic corpora (collections of authentic spoken or written language use) can be a useful tool for linguistic research. There is wide agreement in applied linguistics, for example, that dictionaries, reference grammars, and to some extent language-teaching materials should be based on the analysis of corpora and use citations from corpora to illustrate the phenomena under discussion. Similarly, it is regarded as a matter of course in sociolinguistics and discourse analysis that language variation and the structure of linguistic interaction are investigated largely on the basis of samples of authentic language use. Finally, there are areas of linguistic study, such as historical linguistics, in which no other source of data exists in the first place.

There is considerably less agreement on whether corpora have a place in the investigation of the language system in the context of syntactic theorizing, let alone in contexts where it is explicitly investigated as a cognitive phenomenon. In these areas of research, psycholinguistic experiments (ranging from simple grammaticality judgments to complex stimulus-response designs) are generally seen as the most obvious, or even the only, source of empirical data. Although corpora and corpus-linguistic methods are being adopted by a growing number of researchers in syntax-theoretic and cognitive-linguistic frameworks (see, e.g., Glynn & Fischer, 2010; Gries, 2003; Gries & Stefanowitsch, 2006; Perek, 2015; Schmid, 2000; Schneider, 2014; Wulff, 2008), they are sometimes still explicitly advised against (most notably by Chomsky, 1957, pp. 15–17; see also Andor, 2004, for Chomsky's recent view on the value of corpus data) or ignored entirely.

This hesitant, sometimes downright hostile attitude toward corpora in cognitive or theoretical approaches to language is due at least in part to the assumption that their object of study is incompatible with the type of data collected in corpora. While the object of study is the language system or its representation in speakers' minds (e.g., langue, competence, i-language, linguistic cognition), corpora are widely understood to contain linguistic usage (e.g., parole, performance, e-language, linguistic interaction). However, the incompatibility is more apparent than real once the relationship between corpora and cognition is made explicit. There are two main ways in which this relationship can be conceptualized (see also Stefanowitsch, 2011).

First, and most obviously, there is what we will refer to as the *corpus-as-output* view. From this perspective, a corpus is a sample of the language use of a particu-

lar group of speakers representative for a particular speech community—a snapshot, as it were, of the linguistic performance they collectively generate on the basis of the individual linguistic representations in their minds. The corpus-as-output view is not incompatible with a cognitive approach: As in other methodological frameworks relying on the observation of naturally occurring behavior, we can draw inferences about the mental representations underlying this behavior. Specifically, we can attempt to model mental linguistic representations based on observed patterns of language use in combination with general assumptions about cognitive mechanisms employed in turning linguistic representations into linguistic action.

However, the corpus-as-output view is more straightforwardly compatible with research questions that allow the researcher to remain agnostic or even apathetic with respect to cognition, such as the research areas mentioned at the beginning of this section and much of the research that explicitly describes itself as "corpus-linguistic," regardless of whether this research is descriptive, theoretical, diachronic, or applied (e.g., Biber, Johansson, Leech, Conrad, & Finegan, 1999; Hilpert, 2013; Hunston, 2002; Hunston & Francis, 2000; Leech, Hundt, Mair, & Smith, 2009; Mair, 2004, 2006; McEnery & Hardie, 2012).

Second, there is a somewhat less obvious perspective that we will refer to as the *corpus-as-input* view. From this perspective, the corpus is a (more or less representative) sample of the language use that members of a particular speech community are exposed to during language acquisition and on the basis of which they construct their mental representations in the first place. Although this model is unlikely to appeal to generative linguists, who axiomatically minimize the relevance of the input to the process of constructing linguistic representations, it is widely accepted in usage-based models of first- and second-language acquisition (see, e.g., Ambridge, Kidd, Rowland, & Theakston, 2015; Ellis & Wulff, 2015; Lieven & Tomasello, 2008; MacWhinney, 2008; Tomasello, 2003; see also MacWhinney, Chapter 15, this volume, and Theakston, Chapter 14, this volume).

If we use corpora as a model of linguistic input, we must, of course, take care to construct them in such a way that they approximate the actual input of a given (average member of a) speech community. This is especially important in the context of first-language acquisition research because children are initially exposed to a rather restricted input of exclusively spoken language of a familiar register, to some extent adapted (consciously or subconsciously) to their limited linguistic skills (e.g., baby talk, motherese, caregiver language). To model this input, large samples of naturally occurring adult–child interactions must be collected (see, e.g., the Manchester Corpus [Theakston, Lieven, Pine, & Rowland, 2001]; and CHILDES [MacWhinney, 2000]).

Usage-based models of language are not limited to the initial period of language acquisition, however: Exposure to and experience with performance data are seen as central in the shaping and reshaping of the linguistic competence of speakers throughout their lifetimes (see, e.g., Hoey, 2005; Langacker, 1987, 1991). Although the acquisition of general grammatical schemas ("rules") is complete at some point,

Exhibit 5.1 Data Sources

BNC. British National Corpus. Available at http://www.natcorp.ox.ac.uk (Aston & Burnard, 1998).
BROWN. A Standard Corpus of Present-Day Edited American English. Available at http://www.nltk.org/nltk_data (Francis & Kucera, 1979).
COCA. Corpus of Contemporary American English. Available at http://corpus.byu.edu/coca (Davies, 2009).
ENCOW14. Corpora from the Web, English Version, Release 2014, Slice AX03. Available at https://webcorpora.org (Schäfer & Bildhauer, 2012).

the acquisition of vocabulary and (semi-)idiomatic expressions continues well into adulthood, and—crucial in a discussion of entrenchment—the quantitative distribution of phenomena in the input will continue to influence the representation of these phenomena. In this wider context, large, register-mixed corpora such as the *British National Corpus* (*BNC*; Aston & Burnard, 1998; see Exhibit 5.1 for the data sources used in this chapter) may not be perfect models of the linguistic experience of adult speakers, but they are reasonably close to the input of an idealized average member of the relevant speech community. In our discussion of entrenchment, we will adopt the corpus-as-input view and mention practical limits where applicable.

5.2.2 Corpora and Entrenchment

Although entrenchment is a (hypothesized) cognitive phenomenon, it does not seem to play an important role in the cognitive sciences. It is not mentioned at all, for example, in recent handbooks of psychology and psycholinguistics such as Reisberg (2013) or Traxler and Gernsbacher (2006). Instead, it originates in Langacker's (1987) *Foundations of Cognitive Grammar*, where, after a detailed discussion of the extent to which combinations of simple linguistic units may themselves be units of the linguistic system, he introduced the notion as follows:

> Linguistic units are more realistically conceived of as falling along a continuous scale of entrenchment in cognitive organization. Every use of a structure has a positive impact on its degree of entrenchment, whereas extended periods of disuse have a negative impact. With repeated use, a novel structure becomes progressively entrenched, to the point of becoming a unit; moreover, units are variably entrenched depending on the frequency of their occurrence. (p. 59)

Clearly, Langacker's (1987) idea of entrenchment is related to the psycholinguistic notion of repetition priming and its long-term effects, and in experimental studies in cognitive linguistics, the notions seem to be largely equated with each other. Entrenchment is typically operationalized in terms of reaction times and accuracy of responses to certain stimuli, that is, the same aspects of behavior that are used in psycholinguistics to operationalize long-term effects of priming (e.g., facilitation, stimulus-response binding; see, e.g., Blumenthal-Dramé, Chapter 6, this volume).

However, note that Langacker (1987) does not define entrenchment in psycholinguistic terms but in terms of language "use" (presumably both in the sense of the input that the speaker is confronted with and the language output that they themselves produce). Moreover, he explicitly refers to frequency of occurrence as the driving force of entrenchment, a point he repeats in a later work (Langacker, 2008, p. 238), where he suggests "observed frequency" as the basis for estimating degree of entrenchment empirically.

Because corpora are essentially large samples of linguistic usage data that may serve as models of the linguistic input and output of (adult) speakers, they are the most obvious place to turn to to operationalize Langacker's (1987) definition. This methodological insight is captured in Schmid's (2000) "from-corpus-to-cognition principle," which simply states that "frequency in text instantiates entrenchment in the cognitive system" (p. 39). This principle will serve as our departure point for discussing corpus-based measurements of entrenchment.

5.3 Measuring Entrenchment in Corpora

Both Langacker's definition of entrenchment and Schmid's (2000) operationalization suggest that estimating entrenchment on the basis of a linguistic corpus is straightforward: If frequency drives entrenchment, the number of times that a particular phenomenon occurs in our corpus should be a direct measure of its entrenchment in the cognitive system.

However, as we will show, this interpretation of *frequency* as "(raw) token frequency" is too narrow for all but the simplest units of language. The entrenchment of a monomorphemic word may be measured in terms of token frequency, but linguistic units may differ from such simple units along two dimensions: complexity and schematicity. The dimension of complexity concerns the internal structure of linguistic units, that is, the question of whether, and to what degree, a unit consists of identifiable subunits at the same level of articulation.[1] The dimension of schematicity concerns the question of whether, and to what degree, a linguistic unit is phonologically specified, that is, associated with a particular sound shape. Let us illustrate these dimensions in more detail before we return to the question how they relate to the notion of frequency.

1 Note that morphemes typically consist of more than one phoneme (or, in the case of written language, grapheme), that is, while they are simple (unanalyzable) units at the level of linguistic signs (the *first articulation*, cf. Martinet, 1960, pp. 13–14), they are complex at the level of phonology (the *second articulation*, cf. Martinet, 1960, p. 15), the units of which could themselves be investigated in terms of their entrenchment. We do not discuss the entrenchment of phonemes or graphemes further in this chapter, but at least some of the measures discussed with respect to complex units are presumably also relevant to this issue.

5.3.1 Types of Linguistic Units: Complexity and Schematicity

Figure 5.1 shows the two dimensions as a system of axes and provides examples of (potential) units at different points.

Let us begin with the dimension of complexity. Monomorphemic words are maximally simple: They cannot be analyzed into smaller meaningful units. Multimorphemic words are slightly more complex, at least from the perspective of the analyst: They consist of at least one root and one or more affixes. Still further up the dimension of complexity, we find multiword expressions, such as adjective–noun compounds (e.g., *working class*, *higher education*) and fixed phrases (e.g., *a great time*, *a great deal*, *for old time's sake*). These phrases include *idioms of decoding*, the meaning of which cannot be derived from their component parts and must therefore necessarily be stored and processed as units (e.g., *great deal*). They also include *idioms of encoding*, which are semantically relatively transparent but must be stored and processed as units nevertheless because their *existence* is not predictable from its component parts. For example, someone who does not know the expression *great time* will be able to derive its meaning to some extent from the meaning of *great* and *time* when they encounter it in an appropriate context; however, they would not be able to predict a priori that this phrase can be used to describe an enjoyable situation—in French, the direct translation *grand temps* means 'high time' in the sense of urgency, and in German *große Zeit* is not a fixed phrase at all, and if used, it would most likely be interpreted as '(the)

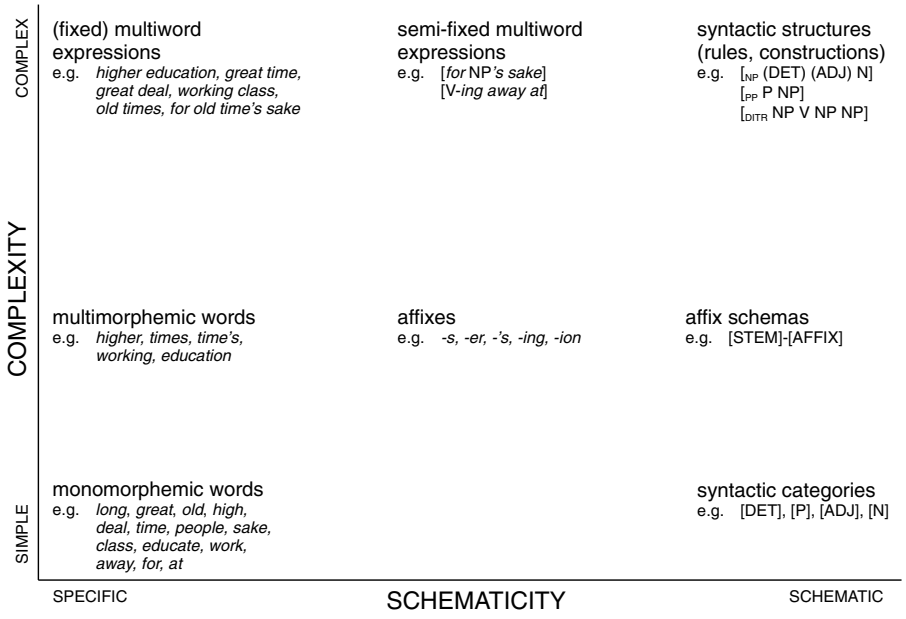

Fig. 5.1: The syntax-lexicon continuum (see Croft, 2001, p. 17; Langacker, 1987, pp. 25–27).

great age (of)' (on idioms of encoding/decoding, see Makkai, 1972, pp. 25–26; Wulff, 2013, p. 276).

Next, let us look at the dimension of schematicity. Again, we find monomorphemic words as one end point of this dimension: They are associated with a specific phonological sound shape (which is frequently invariant or at least predictable by general rules). At the other end of this dimension we find simple but maximally schematic units—syntactic categories such as determiner, adjective, and noun, for example. These are simple in that they cannot be decomposed into smaller units but schematic in that they do not have a specific sound shape associated with them. For example, although the class of adjectives shares syntactic properties (all adjectives can modify a noun, be used as a predicate with a copula, or both), there is nothing that they have in common in terms of phonology (*long* and *high* do not share any phonemes, differ in syllable structure, and the former shares its form with the verb *to long* while the latter shares its form with the noun *high*, as in *an all-time/record high*).

Diametrically opposed to the minimally schematic and minimally complex monomorphemic word are maximally schematic and complex units of grammatical structure. Depending on one's theoretical outlook, this category contains purely formal units, such as phrase structures (noun phrase, prepositional phrase, or even more abstract units such as the X-bar schema of government and binding theory or the head feature principle of Head-Driven Phrase Structure Grammar), or fully schematic meaningful constructions, such as the ditransitive construction with the meaning 'X cause Y to receive Z' (Goldberg, 1995, pp. 141–151) or the resultative construction with the meaning 'X cause Y to become Z.'

Between these extreme points, we find constructions falling along intermediate points at one or both dimensions. For example, there are the maximally complex but only partially schematic semifixed expressions variously referred to as *constructional idioms* (Jackendoff, 2002), *grammar patterns* (Hunston & Francis, 2000), and *variable idioms* (e.g., Stefanowitsch & Gries, 2003). The pattern *for NP's [noun phrase] sake*, for instance, has highly conventionalized instances such as *for God/Christ/heaven's sake*, but also productive instances such as *for her sanity's sake*, *for money's sake*, *for your job's sake* (all of which occur only once in the *BNC*). Another example is *V away at* (as in *eat/chip/hammer/work/scrub away at sth.*). Affixes (which contain a slot for the base to which they attach) may be seen as units of medium complexity and medium schematicity, and general morphological schemas such as STEM-AFFIX ("suffix") and AFFIX-STEM ("prefix") as units of medium complexity and high specificity. Depending on how a given model handles lemmas (in the sense of a construct covering all word forms derived from a particular stem), these may also be thought of as falling somewhere in the middle on both dimensions: Their form is partially specific (phonological content shared by all forms) and partially schematic (to be specified by affixes when deriving the word forms). The treatment of affixes and lemmas depends on the particular theoretical framework adopted, however, and the following discussion does not hinge crucially on this issue.

5.3.2 Usage Intensity and Entrenchment

We are now in a position to take a closer look at the relationship between frequency and entrenchment for units of different degrees of schematicity and complexity. Because we use the term *frequency* in a variety of technical ways, we adopt the ad hoc term *usage intensity* as a cover term. In our terminology, the entrenchment of a unit depends on its usage intensity in language use (as sampled in linguistic corpora); usage intensity must be conceptualized in different ways for different types of units.

Before we discuss measures of different types of units, there is an important caveat to acknowledge. In the preceding section, we used the term *linguistic unit* in the sense of form–meaning pairs—that is, Saussurean *signs* of various degrees of complexity (de Saussure, 1916, pp. 97–100) or Langackerian *symbolic units* (Langacker, 1987, pp. 58–60). Corpora, of course, do not contain such units, but only their forms. Because there is no one-to-one relationship between forms and signs/symbols, we cannot straightforwardly equate tokens in a corpus with linguistic units. For example, the string *bank* belongs to at least three linguistic units: one with the meaning 'financial institution,' one with the meaning 'land along a river,' and one with the meaning 'set of similar things' (*a bank of lights*). If we want to compare, for example, the entrenchment of the words *bank* ('land along a river') and *shore* ('land along the edge of a body of water'), we cannot simply use the frequencies of the respective strings in a corpus—24,546 for *bank(s)* and 2,281 for *shore(s)* in the *BNC*—because the former belongs to the unit 'financial institution' in approximately 95% of the hits. This must, of course, be kept in mind particularly in the case of simple units, where the strings must be manually coded for which linguistic unit they belong to before they are counted. With complex units, it is less of a problem; it is common knowledge in machine translation research that multiword units are less ambiguous than single word units (e.g., in the complex units *along the bank*, *investment bank*, and *bank of lights*, the string *bank* unambiguously belongs to the units 'land along a river,' 'financial institution,' and 'set of similar things,' respectively). This means that we can afford to be less concerned about ambiguity and manual coding the more complex the units investigated are.

5.3.2.1 Simple Units

In the case of maximally specific, maximally simple units such as monomorphemic words, usage intensity can indeed be equated with raw token frequency, that is, the number of times that the unit occurs in a given corpus. There is no reason to measure usage intensity in any other way, and indeed, there does not seem to be any alternative. Consider Table 5.1, which lists the 10 most frequent adjectives and nouns in the *BNC* (morphologically complex words are shown in parentheses).

Tab. 5.1: The 10 Most Frequent Adjectives and Nouns in the *British National Corpus*

Adjectives				Nouns			
Rank	**Word**	***n***	**%**	**Rank**	**Word**	***n***	**%**
1	*other*	129,885	0.0011586	1	*time*	151,754	0.0013537
2	*new*	113,561	0.0010130	2	*people*	121,584	0.0010846
3	*good*	76,551	0.0006829	3	*way*	95,351	0.0008506
4	*old*	52,436	0.0004678	4	*(years)*	88,571	0.0007901
5	*(different)*	47,521	0.0004239	5	*year*	73,009	0.0006513
6	*great*	43,924	0.0003918	6	*(government)*	61,789	0.0005512
7	*(local)*	43,783	0.0003906	7	*day*	58,802	0.0005245
8	*small*	41,812	0.0003730	8	*man*	57,589	0.0005137
9	*(social)*	41,629	0.0003713	9	*world*	57,397	0.0005120
10	*(important)*	38,679	0.0003450	10	*life*	54,903	0.0004898

Note. Unambiguously tagged words only; $N = 112{,}102{,}325$.

For the morphologically simple words, it is obvious that, with the caveat discussed earlier, their raw token frequency should be indicative of their relative entrenchment; for example, *new* should be more entrenched than *old* and *good*, and *time* should be more entrenched than *people* and *way*. As long as we base our entrenchment estimates on a single corpus (or on different corpora of exactly the same size), we can express token frequency as an absolute frequency, but we can also express it independently of corpus size as an unconditional probability p(*w*), that is, the likelihood that the word will occur in any given amount of text.

It is less obvious how morphologically complex words fit into the picture. In at least some cases, different forms derived from the same stem presumably contribute jointly to the entrenchment of the stem—the forms *year* and *years* taken together are more frequent than the form *people*, so we would predict the stem *year* to be more entrenched than the stem *people*. However, it is still to some extent controversial, under what circumstances morphologically complex words are actually recognized (i.e., stored and processed) as complex entities (see, e.g., Blumenthal-Dramé, 2012; Ford, Davis, & Marslen-Wilson, 2010; Hay & Baayen, 2005; Marslen-Wilson, Komisarjevsky, Waksler, & Older, 1994).

There is a wide agreement in the literature, however, that at least some morphologically complex words are stored and processed as units, especially if they are highly frequent or if they are phonologically or semantically nontransparent. For example, the adjective *important* is almost nine times more frequent than the verb *import*, the adjective *imported*, and the noun *import* combined, and the stem has a different meaning in the latter three. It is plausible to assume that *important* is treated like a simple unit by speakers, rather than being derived from the stem *import* and the suffix *-ant* when needed. This is even clearer in the case for the adjective *social*, which is theoretically derivable from the bound stem *soc(i)-* (that also occurs in *society*, *sociology*,

sociopath, etc.) and the suffix *-al* (also found in *local*, *governmental*, etc.): again, *social* is more frequent than all other forms potentially derived from *soc-* taken together, and the stem is pronounced differently in the adjective ([səʊʃ]) than in the other words ([səs] in *society*, [səʊs] in *sociology*, *sociopath*, etc.). Again, it is plausible to assume that speakers treat the adjective *social* like a simple unit.

Whether we treat a given word as a simple unit or as a combination of a stem and one or more affixes is to some extent an empirical issue (see the literature cited earlier in the chapter) and to a large extent a theoretical decision in the context of a particular model of language. Crucially, however, whatever decision one makes, corpora will provide the token frequencies needed to determine the entrenchment of the units one has postulated. If we treat a unit as simple and specific, its entrenchment is measured by its raw frequency; if we treat a unit as complex and/or schematic, this introduces complications, which we turn to next, beginning with the issue of complexity.

5.3.2.2 Complex Units

Instead of stem–affix combinations (the status of which as complex units is at least theoretically debatable in all cases, e.g., in word-and-paradigm models of morphology, and widely debated in some cases, as in the literature cited earlier), we demonstrate the problem of measuring the entrenchment of complex units with the less debatable case of multiword units. The usage intensity of complex units can be measured in three related, but distinct, ways.

5.3.2.2.1 Frequency-Based Measurement

The first way of estimating the entrenchment of complex units is to treat them analogously to simple units and measure their usage intensity in terms of the token frequency of the expression as a whole (as, e.g., Hunston & Francis, 2000, and Sinclair, 1991, do, albeit without explicit reference to *entrenchment*). Table 5.2 shows the 10 most frequent adjective–noun pairs in the *BNC*.

At first glance, it is certainly plausible that these adjective–noun combinations are highly entrenched (at least in the British English of the 1990s). However, it is less plausible that the order of frequency reflects degrees of entrenchment—for example, it is surprising that the combination *other people* should be more entrenched than the proper name *Soviet Union*, or even the combination *young people*.

One problem with simply counting the frequency of occurrence of complex units is that this ignores the individual frequencies of their components: *other* and *people* are both much more frequent overall than *Soviet* and *Union*, so the co-occurrence is less surprising in the case of the former than in the case of the latter. In other words, raw frequency counts can be misleading in the case of complex units because they ignore the a priori likelihood of co-occurrence.

Tab. 5.2: The 10 Most Frequent Adjective–Noun Combinations in the *British National Corpus*

Rank	Bigram	*n*	%	*n* per million words
1	*Prime Minister*	9,461	0.00008440	84.40
2	*other hand*	5,566	0.00004965	49.65
3	*Labour Party*	4,257	0.00003797	37.97
4	*long time*	4,229	0.00003772	37.72
5	*other people*	4,126	0.00003681	36.81
6	*hon. friend*	4,099	0.00003656	36.56
7	*local authorities*	4,028	0.00003593	35.93
8	*great deal*	4,021	0.00003587	35.87
9	*Soviet Union*	3,895	0.00003475	34.75
10	*young people*	3,609	0.00003219	32.19

Note. Unambiguously tagged words only; $N = 112{,}102{,}325$.

5.3.2.2.2 Probability-Based Measurements

The second way of estimating the entrenchment of complex units is to take their complexity into account and to measure their usage intensity by relating the token frequency of the expression as a whole to the individual token frequencies of its component parts. The most straightforward way of doing this is to calculate the conditional probability $p(w_{n+1}|w_n)$, that is, the likelihood that we will encounter a word w_{n+1} given that we have just encountered a word w_n (calculated, obviously, by dividing the frequency of the bigram $w_n w_{n+1}$ by the frequency of w_{n+1}). This is referred to as *transitional probability* in the computational and corpus-linguistic literature (see, e.g., Bush, 2001; Bybee & Scheibman, 1999; and Krug, 2003, for work relating transitional probability directly to the notion of entrenchment; see also Saffran, Newport, & Aslin, 1996, on first-language acquisition). Note that transitional probability is equivalent to the psycholinguistic notion of *cue reliability*, that is, the reliability with which a given linguistic Phenomenon A predicts the occurrence of another Phenomenon B (see, e.g., MacWhinney, 2008; note that cue reliability and related measures are routinely derived from corpora).

Table 5.3 shows the 10 adjective–noun combinations with the highest transitional probability/cue reliability in the *BNC*, that is, the highest probability that the adjective in question will be followed by the noun in question. For reasons we discuss subsequently, we have discarded here and in all following tables all cases that occur fewer than three times and all cases in which the adjective and the noun never occur outside of the combination. We have also removed manually all cases in which the first and/or second word is erroneously tagged as an adjective or noun, respectively.[2]

2 These were mostly proper names, such as *Ronny Johnsen*, or foreign language items such as *ambre solaire* (French) and *faerie queene* (Middle English), as well as a few misspelt or mistokenized items. The *BNC* is tagged using a stochastic tagger that will guess (often, but by no means always correctly) the part of speech of an unknown word based on the part of speech of the preceding word.

Tab. 5.3: Transitional Probability (or Cue Reliability) of Adjective–Noun Combinations in the *British National Corpus*

Rank	Bigram	*n* (Bigram)	*n* (Adjective)	*n* (Noun)	*p*(N|Adj)
1	*ulcerative colitis*	728	754	1,004	0.9655
2	*corned beef*	78	82	1,484	0.9512
3	*arachidonic acid*	72	76	4,898	0.9474
4	*scrolled area*	216	229	34,786	0.9432
5	*stainless steel*	285	307	3,647	0.9283
6	*sclerosing cholangitis*	68	74	200	0.9189
7	*foregone conclusion*	78	85	5,008	0.9176
8	*varicose veins*	58	64	760	0.9062
9	*adoral shields*	114	127	559	0.8976
10	*helping hand*	120	134	32,513	0.8955

At first glance, it is difficult to assess the degree to which these combinations are entrenched. On the one hand, some of them feel like strongly entrenched units (e.g., *corned beef*, *stainless steel*, *foregone conclusion*). On the other hand, combinations such as *arachidonic acid* or *sclerosing cholangitis* are likely to be unfamiliar to most members of the speech community. Thus, it is not the units themselves that are necessarily strongly entrenched, but the (directional) relationship between the second and the first element. In other words, if speakers know the adjective (which is vastly more likely in the case of *stainless* than in the case of *arachidonic*), they will be able to predict the following noun with a high degree of accuracy.

Of course, we can also measure the entrenchment of the relationship in the opposite direction, that is, the conditional probability $p(w_n|w_{n+1})$. This corresponds to the psycholinguistic notion of *cue availability* (see MacWhinney, 2008), that is, the degree to which a linguistic Phenomenon B is available as a cue to a different Phenomenon A. In the case of adjective–noun combinations, the cue availability is p(Adj|N), that is, the probability that a particular adjective will precede a given noun. Table 5.4 shows the adjective–noun pairs with the highest cue availability.

Clearly, cue reliability and cue availability measure different things: There is no overlap between the 10 adjective–noun pairs with the highest cue reliability and those with the highest cue availability, and in fact, pairs with a high cue reliability generally have a low cue availability and vice versa. Take the phrases *stainless steel* (from Table 5.3) and *global warming* (from Table 5.4). Although *stainless steel* has a high cue reliability of 285/307 = 0.9283, it has a rather low cue availability of 285/3647 = 0.0781. Conversely, *global warming* has a high cue availability of 599/683 = 0.877 but a rather low cue reliability of 599/3521 = 0.1701.

In the context of language processing, this difference is presumably relevant: If we hear *stainless*, we have a high expectation that *steel* will follow, but if we hear *global*, we do not necessarily expect *warming* (at least not in the 1990s; the cue reli-

Tab. 5.4: Cue Availability of Adjective–Noun Combinations in the *British National Corpus*

Rank	Bigram	*n* (Bigram)	*n* (Adjective)	*n* (Noun)	Cue reliability
1	*muscular dystrophy*	77	607	83	0.927711
2	*false pretences*	86	3,530	96	0.895833
3	*cerebral palsy*	102	478	115	0.886957
4	*global warming*	599	3,521	683	0.877013
5	*still lifes*	54	2,763	63	0.857143
6	*intestinal pseudo-obstruction*[a]	24	838	28	0.857143
7	*grand theogonist*	28	4,352	33	0.848485
8	*multiple sclerosis*	142	2,204	171	0.830409
9	*major histocompatibility*	36	23,581	44	0.818182
10	*intestinal pseudoobstruction*[a]	18	838	22	0.818182

Note. All combinations where one of the two parts occurred fewer than three times were removed beforehand; mistagged combinations were removed manually (e.g., *faerie queene*, *paba-udca disulphate*).
[a]All calculations are based on orthographic strings so that different spellings of the same word are counted as separate types.

ability may have changed in the meantime). In contrast, if we are unsure whether we have heard *warming* or *warning*, the absence of the adjective *global* would lead us to tend toward *warning*; but if we are unsure whether we have heard *steel* or *seal*, the absence of the adjective *stainless* would not provide much of a cue.

However, with respect to the entrenchment of the unit as a whole, the directionality of the priming relationship is irrelevant; a high probability in either direction should favor entrenchment, while a low probability should disfavor it. Thus, a combined measure may be the best indicator of entrenchment, and there are several fairly common association measures that combine the two probabilities. The most obvious of these is *cue validity* (defined as the product of the two probabilities; see Bates & MacWhinney, 1987, p. 164); others are the *Dice coefficient* (defined as the harmonic mean of the two probabilities) and *minimum sensitivity* (defined as the smaller of the two probabilities). These measures have in common that they give more weight to the smaller of the two probabilities and therefore yield similar results. Table 5.5 shows the 10 adjective–noun pairs with the highest cue validity in the *BNC*.

Like transitional probability/cue reliability and cue availability, the combined measures yield mixed results; although it is plausible in most (perhaps all) cases that there is a strong association between the adjective and the noun, the adjectives and nouns themselves are in many cases infrequent, and thus the combinations are unlikely to be entrenched for an average member of the speech community.

The reason many extremely rare phrases rank highly when using probability-based entrenchment measures is simply that probability-based measures are insensitive to raw frequency. A combination such as *ulcerative colitis*, which despite its terminological status is likely to be familiar to a relatively large proportion of the speech community, has almost the same probability-based entrenchment values as

Tab. 5.5: Cue Validity of Adjective–Noun Pairs in the *British National Corpus*

Rank	Bigram	*n* (Bigram)	*n* (Adjective)	*n* (Noun)	Cue validity
1	*myalgic encephalomyelitis*	7	8	7	0.8750
2	*x-linked agammaglobulinaemia*	5	5	6	0.8333
3	*polychlorinated biphenyls*	26	31	27	0.8076
4	*ulcerative colitis*	728	754	1,004	0.7001
5	*endoplasmic reticulum*	29	30	41	0.6837
6	*ornithischian pisanosaurus*	2	3	2	0.6667
	popular-democratic interpellations	2	2	3	0.6667
	thievin' 'aybag	2	3	2	0.6667
	triple-combed burgonet	2	2	3	0.6667
	twin-elliptic harmonograph	2	3	2	0.6667

Note. $N = 112{,}092{,}864$.

the combination *endoplasmic reticulum*, which is unlikely to be known to anyone who is not a molecular biologist, simply because the proportional frequencies of the noun, the adjective, and the combination are the same—the fact that *ulcerative colitis* is 25 times more frequent has no influence on the measures.

Although this in itself may seem to be a desirable property of probability-based measures in some contexts, it has the unfortunate consequence that the importance of rare combinations is often overestimated: Probability-based measures react more sensitively to small differences for low frequencies than for high frequencies. If, for example, the combination *ulcerative colitis* occurred one time less than it actually does (727 instead of 728 times), the cue validity would change from 0.7001 to 0.6982, a barely noticeable difference of 0.0019. If, however, the combination *twin-elliptic harmonograph* occurred one time less (once instead of twice), the cue validity would change from 0.6667 to 0.167—a drastic change of 0.5. This problem is most dramatic with rare combinations of rare words. There are a number of adjective–noun combinations in the *BNC* that only occur one to three times, but where the adjective and the noun never occur by themselves, such as *honey-throated harangueress* and *flinted knife-sharpener* (once), *ultra-religious chasidim* and *histidine-containing phosphocarrier* (twice), or the nonce-word combinations *slithy toves* (from Lewis Carroll's *Jabberwocky*) and *Vermicious Knids* (from Roald Dahl's *Charlie and the Great Glass Elevator*; three times each). All these combinations will have probability-based measures of 1, that is, they will be estimated as maximally entrenched.

One might assume that this is a problem of corpus construction, that such overestimates could be avoided if the corpus did not contain samples from scientific discourses or literary works by authors known for coining words. However, as cases such as *honey-throated harangueress* and *flinted knife-sharpener* show, even nonspecialized discourse will necessarily contain rare combinations of rare words. We can try to avoid the problems of probability-based measures to some extent by discarding rare combina-

tions from our analysis (which is what we did earlier), but in doing so, we are combining probabilities with frequencies in an arbitrary and unsystematic way that is unlikely to yield psychologically plausible measures of entrenchment.

5.3.2.2.3 Statistical Measures

The third way of estimating the entrenchment of complex units also takes their complexity into account but measures usage intensity in terms of measures derived from contingency tests—either test statistics such as G^2 from the log-likelihood test or χ^2 from the chi-square test, or the p values of exact tests such as Fisher–Yates or the binomial test (see Dunning, 1993; Gries, 2012; Pedersen, 1996; Stefanowitsch & Gries, 2003, for discussion; see Evert, 2004, for a comprehensive overview of such measures).

Like probability-based measures, these association measures take into account the co-occurrence frequency of the elements relative to their individual frequencies, but unlike probability-based measures, they also take into account the frequency of co-occurrence relative to the overall size of the corpus. Thus, on the one hand, if two words that are independently frequent also co-occur together frequently, this co-occurrence will be treated as less important than the co-occurrence of words that mainly occur together—in this sense, statistical measures are better than frequency-based ones. On the other hand, if the combination is frequent, then a given relationship between the co-occurrence frequency and the individual frequencies will be treated as more important than the same relationship in a rare combination; in this sense, statistical measures are better than probability-based measures. Put simply, statistical association measures combine the strength of frequency-based measures and probability-based measures and are thus likely to be the best corpus-based approximation of entrenchment.

Table 5.6 shows the most strongly associated adjective–noun pairs according to the widely used G^2 statistic.[3]

It seems highly plausible that these are among the most strongly entrenched adjective–noun combinations (for British speakers in the early 1990s). All combinations are either compounds (and titles, such as *Prime Minister*, or proper names,

3 Calculated as

$$G^2 = 2\sum_i O_i \cdot \ln\left(\frac{O_i}{E_i}\right),$$

where O is the observed frequency and E the expected frequency of each cell of a two-by-two contingency table containing the frequency of the bigram w_{n+1}, the frequency of w_n outside of the bigram, the frequency of w_{n+1} outside of the bigram, and the frequency of all bigrams containing neither w_n nor w_{n+1}.

Tab. 5.6: Statistical Association of Adjective–Noun Pairs in the *British National Corpus*

Rank	Bigram	*n* (Bigram)	*n* (Adjective)	*n* (Noun)	G^2
1	*Prime Minister*	9,461	11,954	23,394	152,595.99
2	*hon. friend*	4,099	10,548	15,867	59,728.61
3	*Soviet Union*	3,895	10,679	16,436	55,762.05
4	*Labour Party*	4,257	13,084	39,680	51,626.13
5	*great deal*	4,021	44,335	10,434	49,481.33
6	*hon. gentleman*	2,908	10,548	5,070	47,890.81
7	*local authorities*	4,028	44,121	12,855	47,561.11
8	*other hand*	5,566	135,478	32,513	45,315.17
9	*local authority*	3,530	44,121	18,189	37,751.86
10	*wide range*	2,743	11,002	19,411	35,568.47

Note. $N = 112{,}092{,}864$.

such as *Soviet Union*) or compound-like (*local authorities*), or they are (part of) fixed phrases (*great deal*, *[on the] other hand*). Frequent nonfixed phrases such as *long time* or *young people* are also treated as strongly entrenched, but not as strongly as they would be under the frequency-based measure; likewise, combinations with a high cue validity are treated as strongly entrenched if they are frequent (such as *ulcerative colitis*, Rank 161 according to G^2), but not if they are rare (like *twin-elliptic harmonograph*, which does not even make it into the top 10,000).

5.3.2.2.4 Discussion

Of course, it is ultimately an empirical issue which measures best approximate entrenchment (or, more precisely, which corpus-linguistic operationalizations of entrenchment correlate with which psycholinguistic operationalizations of entrenchment). The relative merits of measuring the entrenchment of multiword expressions in terms of token frequency and transitional probability have been discussed in the literature, for example, with respect to their ability to predict univerbation phenomena (e.g., cliticization of the negative particle after modals and copulas after pronouns; see Bybee, 2001; Bybee & Scheibman, 1999; Krug, 1998, 2003).

All three types of measures predict these and related phenomena with an above-chance accuracy, suggesting that all of them are somehow related to entrenchment. However, none of them consistently outperforms the others, suggesting that they measure different aspects of entrenchment that are relevant to different phenomena. Roughly speaking, token frequency measures the usage intensity of the multiword expression as such, corresponding to the entrenchment of the unit as a whole; in contrast, transitional probability measures the usage intensity of the association between the component parts of the expression, corresponding to the entrenchment of the priming relationship between them. Thus, it may be that the more the multiword

expression behaves like a simple unit (i.e., the less likely it is that speakers recognize or are aware of its constituent parts), the better frequency will predict its degree of entrenchment, and the more the multiword expression behaves like a complex unit, the better probability- and/or association-based measures will do.

5.3.2.3 Schematic Units

So far, we have dealt with the dimension of complexity. Let us now turn to the dimension of schematicity, which introduces an additional complication concerning the corpus-based measurement of entrenchment. We illustrate this with two patterns of medium complexity and schematicity: the semifixed expressions [*color* NP ADJ], as in Example 5.1a and 5.1b; and [*drive* NP ADJ], as in Example 5.2a and 5.2b:

Example 5.1 a. *Well*, color me stupid, *because I didn't want to believe he was seeing another woman.* (*Waiting to Exhale*, cit. OED, s.v. colour)
b. "*Well*, color me surprised . . . *not*." (Herald Times, cit. OED, s.v. colour)

Example 5.2 a. *"I don't know how these women cope. It would* drive me crazy." (*BNC* JYB)
b. *"I'm sorry! It's the storm. It's* driving me mad!" (*BNC* CB5).

The phrase [*color* NP ADJ], a (chiefly American English) colloquial expression meaning 'consider me ADJ,' is much less frequent than the second: It does not occur in the *BNC* at all and only 13 times in the 450-million-word *Corpus of Contemporary American English* (*COCA*; Davies, 2009); in contrast, [*drive* NP ADJ] occurs more than 150 times in the *BNC* and more than 1,000 times in *COCA*. Thus, going by frequency, [*drive* NP ADJ] should be much more entrenched than [*color* NP ADJ]. Probability-based measures will lead to the same conclusion: The cue validity of *drive* for [V NP ADJ] is approximately 0.0001, whereas that of *color* for [V NP ADJ] is approximately 0.0000002, that is, 500 times lower.[4]

At first glance, it seems intuitively correct to assign a higher entrenchment to [*drive* NP ADJ] than to [*color* NP ADJ]: Examples like those in 5.2a and 5.2b are likely to be more familiar, and thus more easily and quickly recognized and retrieved, than those in 5.1a and 5.1b. However, what is captured by these measures is not straightforwardly a fact about the patterns [*drive* NP ADJ] and [*color* NP ADJ] because the

4 The exact frequencies needed to perform these calculations are impossible to determine because *COCA* can only be accessed imprecisely using a web interface. We have used estimates of 116,000 occurrences for [V NP ADJ], 6,500 for the verb *color*, 92,000 for the verb *drive*, 12 for the expression [*color* NP ADJ], and 1,200 for the expression [*drive* NP ADJ]; although these are rough estimates, we are confident that they are close to the actual frequencies of occurrence.

raw frequencies confound the frequency of the patterns with the frequency of its specific instantiation(s). The specific expression *drive me crazy*, for example, is vastly more frequent than the specific expression *color me stupid*. This influences the overall raw frequency of the respective patterns, but it does not necessarily influence their entrenchment because it is primarily a fact about the specific expressions and will therefore influence the entrenchment of the specific expressions.

The entrenchment of the schema itself does not depend primarily on its token frequency (i.e., the frequency with which a speaker encounters an instantiation of the pattern) but on its type frequency (i.e., the *number of different instantiations* of the pattern encountered). It is only by virtue of encountering many instantiations of a pattern that a schematic representation emerges from the entrenched representations of individual manifestations (see, e.g., Croft, 2001, p. 28; Croft & Cruse, 2004, p. 309; Diessel, 2004, pp. 29–34; Langacker, 2008, p. 234; Taylor, 2012, p. 285, for the explicit equation of schematic entrenchment with type frequency). Conversely, the entrenchment of a schema is directly related to its productivity, that is, its ability to serve as a template for new instances (see, e.g., Bybee, 2010, p. 67; Croft, 2001, p. 28; Langacker, 2008, p. 234; Lieven, 2010, Taylor, 2012, pp. 173–174, for discussions of the relationship between entrenchment, type frequency, and productivity).

Because the type frequency depends to some extent on token frequency (the more tokens, the more opportunities for different types to occur), the two must be put into some kind of relationship. The simplest measure suggested in the literature is the type/token ratio (i.e., N_{types}/N_{tokens}), which is the percentage of tokens that are different from each other.[5]

Let us return to the example of the abstract patterns [*color* NP ADJ] and [*drive* NP ADJ]. Both expressions have a number of variable slots, including the object (which is most often the pronoun *me* in the case of [*drive* NP ADJ], and near-exclusively so in the case of [*color* NP ADJ]), and the adjectival object complement, which we focus on here to illustrate schematic entrenchment. Exhibit 5.2 shows the adjectives occurring in this slot of the two constructions in a 700-million-word slice of the *ENCOW* corpus (Schäfer & Bildhauer, 2012) together with their frequency.

Clearly, the two patterns are very different from each other as far as the distribution of their instantiations is concerned: Although [*drive* NP ADJ] is instantiated more than 20 times more frequently than [*color* NP ADJ] (1,028 vs. 46), it has fewer different instantiations (24 vs. 31). In other words, although its token frequency is higher, its type frequency is lower. The type/token ratios show the differences between the two patterns even more clearly: For [*drive* NP ADJ], it is just above 2% (24/1028 = 0.0233),

5 Note that type/token ratios are not the only way to quantify the entrenchment of schematic patterns. A substantial literature discusses and tests different ways of measuring morphological productivity (for an overview, see Baayen, 2009). Because of the close relationship between schematic entrenchment and productivity, this literature is highly relevant to the discussion of corpus-based entrenchment measures even if the term *entrenchment* is rarely used there.

Exhibit 5.2 Adjectives occurring in the patterns [drive NP ADJ] and [color NP ADJ]

drive NP ADJ (*N* = 1,028)
crazy (495), *mad* (293), *insane* (127), *wild* (29), *bonkers* (19), *batty* (16), *nuts* (10), *mental* (8), *potty* (6), *crackers* (5), *bananas* (5), *loopy* (2), *silly* (2), *ballistic* (1), *berserk* (1), *buggy* (1), *daft* (1), *delirious* (1), *demented* (1), *frantic* (1), *loony* (1), *nutty* (1), *rowdy* (1), *scatty* (1)

***colo(u)r* NP ADJ** (*N* = 46)
unimpressed (8), *skeptical* (6), *cynical* (2), *disappointed* (2), *jealous* (2), *amazed* (1), *blue* (1), *curious* (1), *dubious* (1), *envious* (1), *excited* (1), *fundamentalist* (1), *green* (1), *happy* (1), *hyper-paranoid* (1), *impressed* (1), *inflammable* (1), *innocent* (1), *interested* (1), *Marxist* (1), *naive* (1), *old-fashioned* (1), *pathetic* (1), *perfect* (1), *shocked* (1), *simple* (1), *slow* (1), *strange* (1), *unconvinced* (1), *unsurprised* (1), *wrong* (1)

whereas for [*color* NP ADJ], it is almost 70% (31/46 = 0.6739). In other words, although the specific expressions *drive sb crazy*, *drive sb mad*, and *drive sb insane* are vastly more entrenched than the specific expressions *color me unimpressed*, *color me cynical*, or any other instance of this pattern, the schematic pattern [*color* NP ADJ] is more entrenched than the schematic pattern [*drive* NP ADJ]. This may seem counterintuitive given the vast difference in token frequency between the two patterns, but note that it is also supported by the qualitative differences in productivity: The instances of [*drive* NP ADJ] are all filled by adjectives meaning "insane" and/or "angry" (i.e., synonyms of *crazy/mad*), whereas the instances of [*color* NP ADJ] are filled by a semantically heterogeneous set of adjectives.

The type/token ratio (or other measures of productivity/schematic entrenchment) can also be applied to simple schematic expressions (e.g., word classes) or fully schematic expressions (e.g., the pattern [ADJ N]), yielding measures that are generally interpretable in terms of entrenchment. For example, the tagged version of the *BROWN* corpus (Francis & Kucera, 1979) contains 7,631 distinct items tagged as (uninflected) adjectives, occurring a total of 68,588 times. Thus, the type/token ratio for the word class adjective is 0.11. Nouns have a somewhat lower but similar type/token ratio of 0.08 (13,130:164,892). In contrast, prepositions have a type/token ratio of 0.001 (132:122,620) and determiners one of 0.0004 (51:136,193), more than 100 times lower than those of nouns and adjectives. Thus, although many individual *members* of the word classes preposition and determiner are more entrenched than even the most frequent individual nouns or adjectives, the word classes noun and adjective themselves are much more entrenched than the word classes preposition and determiner. This corresponds most obviously with the fact that the word classes noun and adjective are open classes, whereas preposition and determiner are closed classes. Open classes such as noun and adjective have a high productivity: Their schematic representation is entrenched enough to allow the easy addition of new members. In contrast, closed classes have a low or even nonexistent productivity: Their schematic representations are so weakly entrenched relative to their individual members that they allow the

addition of new members only occasionally (in the case of prepositions) or not at all (in the case of determiners).

With respect to fully abstract patterns, consider [ADJ N] (the default pattern for nouns modified by adjectives in English) and [N ADJ] (a rare pattern borrowed into English from French during the Middle English period and found in present-day English mainly in job titles such as *Secretary General* or *poet laureate* but also in some otherwise regular noun phrases such as *body politic* or *life eternal*). The tagged version of the *BROWN* corpus contains 23,524 different types of the pattern [ADJ N], occurring a total of 30,142 times; the type/token ratio is thus a very high 0.78, indicating a strong entrenchment of the pattern relative to its individual members (even the most frequent combination, *old man*, occurs only 66 times, accounting for just 0.2% of all tokens). In contrast, there are 22 types of the pattern [N ADJ], occurring a total of 57 times; the type/token ratio is thus a much lower 0.39, indicating a much weaker entrenchment of the pattern relative to its individual members (the most frequent member, *Attorney General*, occurs 18 times, accounting for almost a third (31.6%) of the pattern).

5.4 Corpora and Entrenchment: Further Issues

It is uncontroversial that, as a theoretical concept, entrenchment is causally related to frequency (or, more precisely, usage intensity in its different forms)—as pointed out earlier, this relation was posited by Langacker (1987) as part of the definition of entrenchment. It should also be uncontroversial that linguistic corpora are the most obvious (perhaps the only) source from which different measures of usage intensity can be reliably derived empirically. This seems so obvious that it is taken for granted in much of the corpus-linguistic literature that makes use of the notion (see, e.g., Gries & Stefanowitsch, 2004; Schönefeld, 2012; Stefanowitsch & Gries, 2003; Zeschel, 2008, 2010). It also seems to be taken for granted in experimental psycholinguistics, where stimuli are routinely controlled for frequency.

Nevertheless, the relation between entrenchment and frequency in general, or corpus frequency in particular, has been questioned from a number of perspectives, three of which we discuss in conclusion.

First, it has been argued that entrenchment does not correspond to frequency empirically (Blumenthal-Dramé, 2012). For this criticism to be viable, we would need to have a way of measuring entrenchment directly. However, as pointed out at the end of Section 5.2, entrenchment is a theoretical construct, and any empirical measure of it will be based on operationalizations that capture the phenomenon behind the theoretical construct only partially.

Corpus-based measures will capture overall usage frequency, but they will fail to capture more subtle determinants of usage intensity. The situational salience of an

individual usage event may give it a weight that is disproportionate to its frequency (see in this context Schmid's notion of *contextual entrenchment*; Schmid, 2010, p. 126). For example, the ADJ-N combination *lonely hunter* is not particularly frequent; it occurs three times in the *BNC*, with a cue validity of 0.000018 and a G^2 of 33.22. Nevertheless, it is likely to be highly entrenched for readers of Carson McCuller's *The Heart Is a Lonely Hunter* (note that two of the three uses in the *BNC* occur in mentions of this novel's title). In fact, a linguistic structure may sometimes be salient precisely because it is rare but unusual. Take again the example of *honey-throated haranguer-ess*: It is unlikely that readers of this chapter will have encountered this combination anywhere else (there is not a single hit for *haranguer-ess* on the entire World Wide Web, and only 369 for the adjective *honey-throated*), and it is unlikely that they will encounter it ever again. Still, many are likely to remember it anyway (the authors of this chapter certainly will). Finally, verbal thought likely has an impact on entrenchment, but it will not be captured in corpora unless and until technologies for mind reading become available.

However, psycholinguistic measures are no more likely to capture entrenchment fully accurately. The response time to psycholinguistic stimuli is dependent not just on the kinds of long-term effects of priming that correspond to entrenchment; it also depends on short-term effects (e.g., the recency of the exposure to a linguistic structure, or, again, situational salience).

It should thus not be surprising if there is no perfect match between experimental and corpus-based measures of entrenchment, nor should mismatches be taken as evidence against the plausibility of corpus-based or experimental operationalizations. Each of them can, and does, contribute to our understanding of cognitive processes independently—the value of corpus methods is not *inherently* dependent on whether its results can be replicated or confirmed by experimental methods (or vice versa).

Still, the two types of operationalization are attempts to measure the same phenomenon and should be thought of as complementary. Experimental and elicitation data primarily measure potential *effects* of entrenchment, although they may, by including training phases, also manipulate the *causes* of entrenchment. In contrast, corpus data measure primarily potential *causes* (based on the corpus-as-input model), although they may also be used to investigate certain *effects* of entrenchment (based on the corpus-as-output model). Thus, it would be surprising (and problematic) if there were no correlation at all between them.

In fact, there is encouraging evidence to the effect that the two perspectives and methodological paradigms do approximate the same phenomenon because their results consistently produce converging evidence on various levels of complexity (e.g., Gries, Hampe, & Schönefeld, 2005, 2010; Wiechmann, 2008; see also Stefanowitsch, 2008, for a corpus-based discussion of "negative entrenchment," i.e., the entrenchment of the *absence* of expected combinations of units in the input; and Ambridge, Bidgood, et al., 2015, for corresponding psycholinguistic evidence). This is not to say that there is general agreement on which particular type of measure best describes or predicts

which type of linguistic unit at which level of granularity (e.g., Bybee, 2010; Gries, 2012, 2015; Küchenhoff & Schmid, 2015; Schmid & Küchenhoff, 2013), but given the complexities involved in measuring frequency/usage intensity, this is hardly surprising. Crucially, these and other studies show that there is a correlation between psycholinguistic measures of entrenchment and text frequency in general.

Second, it has recently been argued that corpora are generally unsuitable for the study of entrenchment. Blumenthal-Dramé (2012, especially Chapter 8) argues that because corpora aggregate the linguistic usage of many speakers, they cannot be used for determining the entrenchment of linguistic structures in a given individual's mind. Similarly, Schmid (2010, p. 117) suggested that corpus-based measures of entrenchment are better thought of as measuring conventionalization.

It is true that a given linguistic corpus is not typically representative of the input, let alone the output of a particular individual. However, this does not constitute an argument against using corpora in the study of cognition because the same is true of experimental measures, which are also averaged across groups of subjects. As in a balanced corpus, these subjects are assumed to be, but never actually shown to be, representative of the speech community. Thus, experiments, like corpora, measure the average entrenchment of a structure in the mind of a typical member of the speech community.

Of course, the cognitive sciences are generally not actually concerned with the mental representations of particular individuals, but if they were, note that it would be much easier to construct corpora representing the input–output of a particular individual than it would be to run valid and reliable experiments on a particular individual (for a corpus-based case study of individual differences in entrenchment, see Schmid & Mantlik, 2015).

It is also plausible to assume that corpus-based measures of entrenchment may be used to measure degrees of conventionalization, but this does not preclude their use in measuring average entrenchment. Conventionalization is a theoretical construct that differs from entrenchment mainly in that it describes established linguistic structures at the level of the linguistic system itself (in syntactic theory or in grammar writing) or at the level of the speech community as an abstract entity (e.g., in sociolinguistics). Entrenchment, in contrast, describes established linguistic structures at the level of an average speaker's mental representation or at the level of the speech community as an aggregate of individuals. This is not to say that entrenchment and conventionalization are the same thing—they differ theoretically in a number of ways. It is to say that they are measured in the same way (or similar) ways—they do not differ empirically. Perhaps we could say that the corpus-as-input view is more amenable to models concerned with entrenchment, whereas the corpus-as-output view is more in line with models interested in conventionality.

Third, and finally, there is a general criticism of corpus linguistics that is also relevant to the quality of entrenchment measures derived from corpora: Although we can easily imagine a perfect corpus (or different perfect corpora for different research

contexts), actually existing corpora fall well short of such perfect corpora in terms of size, sampling of text types and demographic representation. In Section 5.3, we demonstrated some of the problems caused by the inclusion of specific text types when inferring the average entrenchment of structures. However, sampling issues are not unique to corpus linguistics but are an integral part of any methodology. They must be dealt with in the short term by keeping them in mind when moving from data to model building, and in the long term by reducing them as much as possible. In the case of corpus linguistics, this means making more complete and more creative use of the resources that are already available and that encompass not just established corpora like the *BNC*, the *BROWN*-Family, and the (still expanding) *ICE*-Family, but also specialized corpora such as the Manchester Corpus of the input to and output of children during first-language acquisition (Theakston, Lieven, Pine, & Rowland, 2001) and the vast and varied text archives that are increasingly made available online (and, of course, the Internet with its huge amount of informal everyday language found on message boards, mailing lists, and the social media). Finally, it means constructing larger and more balanced corpora.

References

Ambridge, B., Bidgood, A., Twomey, K. E., Pine, J. M., Rowland, C. F., & Freudenthal, D. (2015). Preemption versus entrenchment: Towards a construction-general solution to the problem of the retreat from verb argument structure overgeneralization. *PLoS ONE*, *10*, e0123723. http://dx.doi.org/10.1371/journal.pone.0123723

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *Journal of Child Language*, *42*, 239–273. http://dx.doi.org/10.1017/S030500091400049X

Andor, J. (2004). The master and his performance: An interview with Noam Chomsky. *Intercultural Pragmatics*, *1*, 93–111. http://dx.doi.org/10.1515/iprg.2004.009

Aston, G., & Burnard, L. (1998). *The BNC handbook: Exploring the British National Corpus with SARA*. Edinburgh, Scotland: Edinburgh University Press.

Baayen, R. H. (2009). Corpus linguistics in morphology: Morphological productivity. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics. An international handbook* (Vol. 2, HSK, 29.2, pp. 899–919). Berlin, Germany: Mouton de Gruyter.

Bates, E., & MacWhinney, B. (1987). Competition, variation, and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–194). Hillsdale, NJ: Erlbaum.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, England: Longman.

Blumenthal-Dramé, A. (2012). *Entrenchment in usage-based theories: What corpus data do and do not reveal about the mind*. Berlin, Germany: De Gruyter Mouton.

Bush, N. (2001). Frequency effects and word-boundary palatalization in English. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 256–280). http://dx.doi.org/10.1075/tsl.45.14bus

Bybee, J. L. (2001). Frequency effects on French liaison. In J. Bybee & P. Hopper (Eds.), *Frequency and the emergence of linguistic structure* (pp. 337–359). http://dx.doi.org/10.1075/tsl.45.17byb

Bybee, J. L. (2010). *Language, usage and cognition*. http://dx.doi.org/10.1017/CBO9780511750526

Bybee, J. L. & Scheibman, J. (1999). The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics*, *37*, 575–596. http://dx.doi.org/10.1515/ling.37.4.575

Chomsky, N. A. (1957). *Syntactic structures*. The Hague, the Netherlands: Mouton.

Croft, W. (2001). *Radical construction grammar. Syntactic theory in typological perspective*. http://dx.doi.org/10.1093/acprof:oso/9780198299554.001.0001

Croft, W., & Cruse, D. A. (2004). *Cognitive linguistics*. http://dx.doi.org/10.1017/CBO9780511803864

Davies, M. (2009). The 385+ million word *Corpus of Contemporary American English* (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, *14*, 159–190. http://dx.doi.org/10.1075/ijcl.14.2.02dav

de Saussure, F. (1916). *Cours de linguistique générale* [Course in general linguistics]. Lausanne, Switzerland: Payot.

Diessel, H. (2004). *The acquisition of complex sentences*. http://dx.doi.org/10.1017/CBO9780511486531

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61–74.

Ellis, N. C., & Wulff, S. (2015). Usage-based approaches in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 75–93). New York, NY: Routledge.

Evert, S. (2004). *The statistics of word co-occurrences: Word pairs and collocations* (Unpublished doctoral dissertation). Universität Stuttgart, Stuttgart, Germany.

Ford, M. A., Davis, M. H., & Marslen-Wilson, W. D. (2010). Derivational morphology and base morpheme frequency. *Journal of Memory and Language*, *63*, 117–130. http://dx.doi.org/10.1016/j.jml.2009.01.003

Francis, W. N., & Kucera, H. (1979). *Brown corpus manual: Manual of information to accompany* A standard corpus of present-day edited American English*, for use with digital computers* (revised and amplified). Providence, RI: Department of Linguistics, Brown University.

Glynn, D., & Fischer, K. (Eds.). (2010). *Quantitative methods in cognitive semantics: Corpus-driven approaches*. http://dx.doi.org/10.1515/9783110226423

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. Chicago, IL: University of Chicago Press.

Gries, S. T. (2003). *Multifactorial analysis in corpus linguistics: A study of particle placement*. New York, NY: Continuum.

Gries, S. T. (2012). Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics: Some necessary clarifications. *Studies in Language*, *11*, 477–510. http://dx.doi.org/10.1075/sl.36.3.02gri

Gries, S. T. (2015). More (old and new) misunderstandings of collostructional analysis: On Schmid and Küchenhoff (2013). *Cognitive Linguistics*, *26*, 505–536. http://dx.doi.org/10.1515/cog-2014-0092

Gries, S. T., Hampe, B., & Schönefeld, D. (2005). Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics*, *16*, 635–676. http://dx.doi.org/10.1515/cogl.2005.16.4.635

Gries, S. T., Hampe, B., & Schönefeld, D. (2010). Converging evidence II: More on the association of verbs and constructions. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 59–72). Stanford, CA: CSLI.

Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on "alternations." *International Journal of Corpus Linguistics*, *9*, 97–129. http://dx.doi.org/10.1075/ijcl.9.1.06gri

Gries, S. T., & Stefanowitsch, A. (Eds.). (2006). *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis.* http://dx.doi.org/10.1515/9783110197709

Hay, J. B., & Baayen, R. H. (2005). Shifting paradigms: Gradient structure in morphology. *Trends in Cognitive Sciences*, *9*, 342–348. http://dx.doi.org/10.1016/j.tics.2005.04.002

Hilpert, M. (2013). *Constructional change in English: Developments in allomorphy, word formation, and syntax.* http://dx.doi.org/10.1017/CBO9781139004206

Hoey, M. (2005). *Lexical priming. A new theory of words and language.* Abington, England: Routledge.

Hunston, S. (2002). *Corpora in applied linguistics.* http://dx.doi.org/10.1017/CBO9781139524773

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English.* http://dx.doi.org/10.1075/scl.4

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* http://dx.doi.org/10.1093/acprof:oso/9780198270126.001.0001

Krug, M. (1998). String frequency: A cognitive motivating factor in coalescence, language processing, and linguistic change. *Journal of English Linguistics*, *26*, 286–320. http://dx.doi.org/10.1177/007542429802600402

Krug, M. (2003). Frequency as a determinant in grammatical variation and change. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of grammatical variation in English* (pp. 7–67). http://dx.doi.org/10.1515/9783110900019.7

Küchenhoff, H., & Schmid, H.-J. (2015). Reply to "More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff" by Stefan Th. Gries. *Cognitive Linguistics*, *26*, 537–547. http://dx.doi.org/10.1515/cog-2015-0053

Langacker, R. W. (1987). *Foundations of cognitive grammar: Vol. 1. Theoretical prerequisites.* Stanford, CA: Stanford University Press.

Langacker, R. W. (1991). *Foundations of cognitive grammar: Vol. 2. Descriptive applications.* Stanford, CA: Stanford University Press.

Langacker, R. W. (2008). *Cognitive grammar: A basic introduction.* http://dx.doi.org/10.1093/acprof:oso/9780195331967.001.0001

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in contemporary English: A grammatical study.* http://dx.doi.org/10.1017/CBO9780511642210

Lieven, E. (2010). Input and first language acquisition: Evaluating the role of frequency. *Lingua*, *120*, 2546–2556. http://dx.doi.org/10.1016/j.lingua.2010.06.005

Lieven, E., & Tomasello, M. (2008). Children's first language acquisition from a usage-based perspective. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 168–196). New York, NY: Routledge.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Mahwah, NJ: Erlbaum.

MacWhinney, B. (2008). A unified model. In P. J. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 341–371). New York, NY: Routledge.

Mair, C. (2004). Corpus linguistics and grammaticalisation theory: Statistics, frequencies, and beyond. In H. Lindquist & C. Mair (Eds.), *Studies in corpus linguistics* (Vol. 13, pp. 121–150). http://dx.doi.org/10.1075/scl.13.07mai

Mair, C. (2006). *Twentieth-century English: History, variation and standardization.* http://dx.doi.org/10.1017/CBO9780511486951

Makkai, A. (1972). *Idiom structure in English*. http://dx.doi.org/10.1515/9783110812671

Marslen-Wilson, W. D., Komisarjevsky, L., Waksler, R., & Older, L. (1994). Morphology and meaning in the English mental lexicon. *Psychological Review*, *101*, 3–33. http://dx.doi.org/10.1037/0033-295X.101.1.3

Martinet, A. (1960). *Éléments de linguistique générale* [Elements of general linguistics]. Paris, France: Colin.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge, England: Cambridge University Press.

Pedersen, T. (1996). Fishing for exactness. *Proceedings of the South-Central SAS Users Group Conference* (SCSUG-96), 188–200. Austin, TX.

Perek, F. (2015). *Argument structure in usage-based construction grammar: Experimental and corpus-based perspectives*. Amsterdam, the Netherlands: Benjamins.

Reisberg, D. (Ed.). (2013). *The Oxford handbook of cognitive psychology*. http://dx.doi.org/10.1093/oxfordhb/9780195376746.001.0001

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621. http://dx.doi.org/10.1006/jmla.1996.0032

Schäfer, R., & Bildhauer, F. (2012). Building large corpora from the web using a new efficient tool chain. In N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, . . . S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC–12; pp. 486–493). Istanbul, Turkey: ELRA.

Schmid, H.-J. (2000). *English abstract nouns as conceptual shells: From corpus to cognition*. http://dx.doi.org/10.1515/9783110808704

Schmid, H.-J. (2010). Does frequency in text instantiate entrenchment in the cognitive system? In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 101–133). http://dx.doi.org/10.1515/9783110226423.101

Schmid, H.-J., & Küchenhoff, H. (2013). Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics*, *24*, 531–577. http://dx.doi.org/10.1515/cog-2013-0018

Schmid, H.-J., & Mantlik, A. (2015). Entrenchment in historical corpora? Reconstructing dead authors' minds from their usage profiles. *Anglia*, *133*, 583–623. http://dx.doi.org/10.1515/ang-2015-0056

Schneider, U. (2014). *Frequency, chunks and hesitations. A usage-based analysis of chunking in English* (Doctoral dissertation). Albert-Ludwigs-Universität, Freiburg, Germany.

Schönefeld, D. (2012). Things going unnoticed—A usage-based analysis of *go*-constructions. In D. Divjak & S. T. Gries (Eds.), *Frequency effects in language representation* (Vol. 2, pp. 11–49). Berlin, Germany: De Gruyter Mouton.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, England: Oxford University Press.

Stefanowitsch, A. (2008). Negative entrenchment: A usage-based approach to negative evidence. *Cognitive Linguistics*, *19*, 513–531. http://dx.doi.org/10.1515/COGL.2008.020

Stefanowitsch, A. (2011). Cognitive linguistics meets the corpus. In M. Brdar, S. T. Gries, & M. Žic Fuchs (Eds.), *Cognitive linguistics: Convergence and expansion* (pp. 257–289). http://dx.doi.org/10.1075/hcp.32.16ste

Stefanowitsch, A., & Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, *8*, 209–243. http://dx.doi.org/10.1075/ijcl.8.2.03ste

Taylor, J. R. (2012). *The mental corpus: How language is represented in the mind.* http://dx.doi.org/10.1093/acprof:oso/9780199290802.001.0001

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*, 127–152. http://dx.doi.org/10.1017/S0305000900004608

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Traxler, M. J., & Gernsbacher, M. A. (2006). *Handbook of psycholinguistics*. Amsterdam, the Netherlands; Boston, MA: Elsevier/Academic Press.

Wiechmann, D. (2008). On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory*, *4*, 253–290. http://dx.doi.org/10.1515/CLLT.2008.011

Wulff, S. (2008). *Rethinking idiomaticity: A usage-based approach.* London, England: Continuum.

Wulff, S. (2013). Words and idioms. In T. Hoffmann & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 274–289). Oxford, England: Oxford University Press.

Zeschel, A. (2008). Lexical chunking effects in syntactic processing. *Cognitive Linguistics*, *19*, 427–446. http://dx.doi.org/10.1515/COGL.2008.016

Zeschel, A. (2010). Exemplars and analogy: Semantic extension in constructional networks. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 201–219). Berlin, Germany: de Gruyter Mouton.