Does frequency in text instantiate entrenchment in the cognitive system?

Hans-Jörg Schmid

Abstract
This paper investigates the relation between observed discourse frequencies of linguistic elements and structures, on the one hand, and assumptions concerning the entrenchment of these units in the minds of speakers, on the other. While it is usually assumed that there is a fairly direct correlation between frequency of use and degree of entrenchment, it is argued that many essential questions concerning this relation have remained unanswered so far: What is the role of absolute frequency of occurrence as opposed to frequency relative to a given reference construction? How are relative discourse frequencies to be captured statistically in such a way that, for instance, rare lexical items that typically occur in certain constructions can be differentiated from frequent lexical items which are more versatile but also observed to occur in the same construction, often with considerable absolute frequencies of occurrence? What are the psychological implications of different combinations of high and low absolute and relative frequencies? While the paper suggests solutions to some of these problems it also points to a number of unresolved issues to be addressed in the future and calls for a more modest and cautious way of interpreting quantitative observations in cognitive terms.

Keywords: discourse frequency, entrenchment, quantitative approaches, collostruction

1. Introduction[1]

It is common practice in corpus linguistics to assume that the frequency distribution of tokens and types of linguistic phenomena in corpora have – to put it as generally as possible – some kind of significance. Essentially, more frequently occurring structures are believed to hold a more prominent place, not only in actual discourse but also in the linguistic system, than those occurring less often.

Cognitively-oriented corpus linguists also subscribe to this assumption, but they tend to go one step further. Given their cognitive leaning, they should be on the hunt for psychologically plausible models of language based on quantitative observations of corpus data. More specifically, they try to correlate the frequency of occurrence of linguistic phenomena (as observed in corpora) with their salience or entrenchment in the cognitive system. A corollary of this assumption is that patterns of frequency distributions of lexico-grammatical variants of linguistic units correspond to variable degrees of entrenchment of cognitive processes or representations associated with them.

Among early investigations pursuing this line of corpus-based cognitive-linguistic reasoning are Rudzka-Ostyn's (1989) study of the polysemy of the English verb *ask*, Dirven's (1991) paper on agree and my own work on *start* vs. *begin* and the polysemy of the noun *idea* (Schmid 1993).[2] During this early period of quantitative Cognitive Semantics, frequency patterns tended to be interpreted in terms of typicality of meaning, with the most frequent usage-patterns being taken to reflect (proto-)typical senses of lexemes.[3] Schematic meanings were also taken into account in these studies, but it seemed less clear how they are related to frequency distributions (cf. Schmid 1993: 218). The work by Rudzka-Ostyn, Dirven and myself was very much inspired by ideas on the network nature of (lexicalized) conceptual categories developed in a non-quantitative framework by Brugman (1981), Lindner (1982), Geeraerts (1983), Herskovits (1986), Lakoff (1987), Schulze (1988) and others in the course of the 1980s.

Tracing back the historical roots of corpus-driven quantitative Cognitive Semantics is not an end in itself or just an homage to the scholars who have built the foundation that made the current work in the field possible. It is important because it reminds us of the ultimate aim of cognitive linguists to come up with linguistic models that actually claim to reflect (what we believe to know about) the way our minds work. In my perception, this aim is in danger of falling into oblivion. Trapped in a numerical maze by the irresistible lure of masses of data, smart corpus queries, long periods of number-crunching and skilful applications of advanced statistical methods, present-day corpus-driven Cognitive Semantics runs the risk of losing sight of the bigger picture hidden in language, and of the cognitive aspects of meaning it purports to unravel, if only by the name by which it presents itself to the wider linguistic community.

In view of this danger, this paper backtracks a number of steps in the development of corpus-driven Cognitive Semantics. It will question one of the main assumptions underlying, more or less tacitly or explicitly, much of the current work in the field. More specifically, the paper brings under scrutiny the hypothesis that frequency in text more or less directly instantiates salience or entrenchment in the cognitive system, a claim most forcefully proposed in the form of my own From-corpus-to-cognition-principle (Schmid 2000: 39) as an extension to Halliday's dictum that "frequency in text instantiate[s] probability in the [linguistic] system" (Halliday 1993: 3). To make this quite clear, the aim of this contribution is not to question the importance and significance of quantitative approaches in Cognitive Linguistics as such, far from it, but rather to point out a number of potential pitfalls and shortcomings that will have to be addressed in the future.

2. The methodology of quantitative Cognitive Semantics: An idealized outline

To locate the issues addressed in this paper in the methodological landscape of quantitative Cognitive Semantics, it is important to spell out the methodological steps typically gone through in studies set in this framework. It should be borne in mind that this is a generalized and idealized version of the state of the art, which may not have been applied this way in any concrete study but still represents the blueprint of a recipe underlying many. The sequence of steps is summarized in Table 1 (cf. also Tummers, Heylen and Geeraerts 2005: 238-245):

Table 1. Idealized version of the methodology of quantitative Cognitive Semantics

1. **Choice of object of study**: Find and define an interesting linguistic phenomenon.
2. **Choice of corpus**: Find either a huge or a tailor-made special-purpose corpus – ideally representative of the kind of language you want to study (whatever that means) – that includes a large number of diverse instances of the linguistic phenomenon.
3. **Formulate corpus queries**: Operationalize the linguistic problem in such a way that it can be formulated as a corpus query or set of corpus queries.
4. **Retrieve and clean up material**: Use the query to retrieve all instances of the phenomenon from the corpus and get rid of clear unwanted hits.
5. **Get material under control**: Analyse the distribution and frequencies of variants of the valid instances of the phenomenon retrieved; annotate them accordingly or insert them in a database.

6. **Analyse material mathematically/statistically:** Capture the frequency distribution mathematically and check for statistical significances.
7. **Interpret quantitative findings:** Interpret quantitative findings in terms of semantic and cognitive organization.

While it is well known in corpus-linguistic circles that not a single one of these steps is trivial, this is easily overlooked. Although a lot more could be said about the other steps and the way they influence the outcome of corpus studies, I will focus my attention on steps 6 and 7, since my main concern is the significance of frequencies. Suffice it to recall at this stage that the apparently so objective quantitative approach has a much larger number of subjective decisions and sources of errors built into its methodological apparatus than most practitioners of the art are willing to admit (cf. Tognini-Bonelli 2001: 178, Mukherjee 2005: 71–72).

3. The constructions serving as case studies: Shell-content construction

The focus of this paper is on methodological issues. Since these should not be discussed in a linguistic vacuum, I will go back to data from my own previous research (cf. Schmid 2000) to illustrate the problems at hand. Four related types of nominal constructions, all consisting of an abstract noun and a complementing *that*-clause or *to*-infinitive, will serve as case studies. The constructions are illustrated in Table 2:

Table 2. Variants of the shell-content construction

a)	N + that-clause: *The fact that abstract nouns are difficult to pin down ...*
b)	N + to-infinitive: *The idea to illustrate the patterns investigated ...*
c)	N + BE + that-clause: *The problem is that there is a lot to study.*
d)	N + BE + to-infinitive: *The solution is to focus on a bunch of examples.*

As the table shows, the four types differ with regard to the form of the complement (*that*-clause *vs. to*-infinitive) and the link between N and complement (direct link as nominal postmodifier vs. link by means of the copula *be*). In my previous work I have referred to the nouns in these constructions as *shell nouns*, because they conceptually encapsulate the complex pieces of information expressed in the clauses (referred to as *shell contents*). The whole constructions are seen as variants of a more schematic *shell-content construction* (Schmid 2007a, Ungerer and Schmid 2006: 248-250).[4]

4. Data source and retrieval

The data re-considered here were originally collected in 1996. The material was taken from the British section of COBUILD's *Bank of English*, amounting at that time to 225 million running words from the following subcorpora: spoken conversation, transcribed BBC recordings, ephemeral texts such as brochures and leaflets, fiction and non-fiction books, magazines (both lifestyle and political, including *The Economist*), broadsheet newspapers (*The Times*, *The Guardian* and *The Independent*), the tabloid *English Today* as well as the science journal *The New Scientist*. It should be borne in mind that with a proportion of about two-thirds of the whole material, texts from media sources make up the bulk of this corpus.

Table 3. Corpus queries and numbers of matches (Schmid 2000: 44-45)

Query statement	Number of matching lines in the 225m corpus
Pattern N-cl:	
NN+that/CS (NN = noun, CS = conjunction)	280,217
NN+to+VB (VB = base form of verbs)	560,148
Pattern N-be-to	
NN+is+to	28,463
NN+was+to	12,728
NN+has+been+to	962
NN+will+be+to	960
NN+would+be+to	1,421
NN+would+have+been+to	133
Pattern N-be-that	
NN+is+that	37,155
NN+was+that	9,104
NN+has+been+that	433
NN+will+be+that	178
NN+would+be+that	264
NN+would+have+been+that	19

In the corpus queries, the nominal slots were defined by a part-of-speech dummy (NN), while the complement clauses were identified by means of the complementizers *that* and *to*. The queries aimed at the constructions containing the copula included different morphological variants. Table 3 gives a summary of the query statements and the number of matching lines in the 225 million-word corpus. In addition, analogous patterns with *wh*-clauses were retrieved (e.g. *the question why he didn't come* or *the problem how to define the task*), as well as the highly frequent anaphoric instances of shell nouns, but neither of these types is under consideration here.

After half-manual data-cleaning, the following numbers of valid tokens and noun types represented in the corpus remained for further investigation.

Table 4. Valid hits in the corpus study conducted by Schmid (2000)

CONSTRUCTION	Tokens	Noun types
N + that-clause	141,476	350
N + to-infinitive	228,165	200
N + BE + that-clause	30,992	366
N + BE + to-infinitive	21,876	162

5. Capturing the data

Following the rationale outlined in Table 1, the next step is to arrange the material in such a way that patterns of distribution become visible. This step is often combined with step 6, a first attempt to capture the distribution of the data mathematically.

Applied to the present material, the aim of steps 5 and 6 is to come up with interesting, data-driven observations concerning the interaction between the four types of constructions and the types of nouns occurring in them. More specifically the following questions are of concern:

- To what extent do the different types of constructions attract specific types of nouns?
- To what extent do certain nouns depend on one or more of the constructions for their occurrence in actual discourse?
- Is there a semantic affinity between the frequency distribution of types of nouns and the types of constructions?

In what follows, two different attempts to answer these questions will be sketched out: the attraction-reliance method proposed by Schmid (2000) and the collostructional method introduced by Stefanowitsch and Gries (2003).

5.1 Simple arithmetic: attraction-reliance method

In the late 1990s, when the material for the study reported in Schmid (2000) was collected and analysed, a number of linguistico-statistical measures, which were designed to answer questions of this type, were of course firmly established. The most commonly applied ones were *t*-score and *Mutual Information* (see Church and Hanks 1990), which were both used at COBUILD (Clear 1993) and for analyses of the *British National Corpus*. However, these measures were so technical that even linguists who had applied them with some success admitted they were not able to see behind the formulas and to interpret the actual linguistic significance (cf. also Stubbs 1995 for a critical discussion). Moreover, these formulas are designed to be calculated for a set of potential collocators of the so-called node within a defined span, without taking into consideration their relation to the node. This procedure, which may be useful and necessary if one is concerned with co-occurrence tendencies in general, would have unnecessarily blurred issues that were perfectly clear in the case of shell-content constructions.

In view of these disadvantages of the established instruments in the statistical toolbox, I decided to come up with much simpler but more transparent arithmetic ways of capturing the interaction between nouns and construction. The measures were called *attraction* and *reliance* and calculated as represented in Figure 1:

$$\text{Attraction} = \frac{\textit{frequency of a noun in a pattern x 100}}{\text{total frequency of the pattern}}$$

$$\textit{Reliance} = \frac{\text{frequency of a noun in a pattern x 100}}{\text{total frequency of the noun in the corpus}}$$

Figure 1. Calculating the measures of attraction and reliance (Schmid 2000: 54)

As the figure shows, *attraction* is calculated by dividing the frequency of occurrence of a noun in a pattern by the frequency of the pattern in the corpus. The result of this division measures the degree to which a pattern attracts a particular noun. Since the denominator of the fraction is the same for all nouns which occur in a pattern, the scores for this value are directly proportional to the raw frequencies of nouns. The measure facilitates the comparison of the relative importance of individual nouns for a pattern. While capturing the relation between nouns and pattern, this is very much a paradigmatic way of looking at the nominal slot in the pattern. Differences in attraction are illustrated in Figure 2 for the frequency scores found for the two nouns *fact* and *idea* in the patterns N + *that*-clause:

$$\text{Attraction}_{\textit{fact that}} = \frac{\mathit{26106 \text{ x } 100}}{141476} = \mathit{18.45\%}$$

$$\textit{Attraction}_{\text{idea that}} = \frac{4812 \text{ x } 100}{141476} = \mathit{3.40\%}$$

Figure 2. Exemplifying differences in attraction

As shown in Figure 1, *reliance* is calculated by dividing the frequency of occurrence of a noun in a pattern by its frequency of occurrence in the whole corpus. This measure expresses the proportion of uses of nouns in the patterns vis-à-vis other usage-types of the same noun. As the denominator of the fraction varies with the overall frequency of a noun in the corpus, scores for reliance are not proportional to their frequency of occurrence in the constructions. Viewed from the nouns' perspective, reliance is a syntagmatic rather than a paradigmatic measure, since it accounts for combinations of nouns with types of patterns. Figure 3 illustrates differences in reliance scores in the pattern N + *that*-clause for the nouns *fact* and *realization*.

$$\text{Reliance}_{\textit{fact that}} = \frac{\mathit{26106 \text{ x } 100}}{68472} = \mathit{38.13\%}$$

$$\textit{Reliance}_{\text{realization that}} = \frac{820 \text{ x } 100}{1185} = \mathit{69.20\%}$$

Figure 3. Exemplifying differences in reliance

While *fact* was found much more often in the pattern N + *that*-clause than *realization*, the latter can boast a much higher score for reliance because its overall frequency of occurrence in the corpus is much lower, too.

The attraction-reliance method of capturing the data allows for two different types of information about frequency of occurrence to be made. On the one hand, focusing on how the nominal slots in the constructions are filled, different noun types can be ranked according to their scores for both attraction and reliance in a certain pattern. An illustrative extract from such a ranking for the pattern N + *that*-clause is given in Table 5, where the three columns on the left-hand side provide the ranking for attraction and the other four columns the one for reliance.

Table 5. Ranking of attraction and reliance scores for the construction N + *that*-clause (top 20; 141,476 tokens in the corpus)

Noun	Freq. in pattern	Attraction	Noun	Freq. in pattern	Freq. in Corpus	Reliance
fact	26,106	18.45%	*realization*	820	1,185	69.20%
evidence	5,007	3.54%	*proviso*	111	250	44.40%
idea	4,812	3.40%	*assumption*	1,391	3,151	44.14%
doubt	4,010	2.83%	*assertion*	596	1,492	39.95%
belief	3,696	2.61%	*belief*	3,696	9,344	39.55%
view	3,532	2.50%	*insistence*	796	2,069	38.47%
hope	2,727	1.93%	*fact*	26,106	68,472	38.13%
news	2,572	1.82%	*premise*	274	765	35.82%
feeling	2,511	1.77%	*misapprehension*	44	123	35.77%
impression	2,279	1.61%	*suggestion*	2,033	5,854	34.73%
possibility	2,232	1.58%	*dictum*	84	249	33.73%
claim	2,194	1.55%	*stipulation*	48	145	33.10%
suggestion	2,033	1.44%	*misconception*	91	284	32.04%
speculation	1,922	1.36%	*truism*	47	150	31.33%
knowledge	1,794	1.27%	*reminder*	812	2,688	30.21%
sign	1,738	1.23%	*notion*	1,655	5,713	28.97%
notion	1,655	1.17%	*coincidence*	627	2,196	28.55%
point	1,511	1.07%	*speculation*	1,922	6,778	28.36%
warning	1,460	1.03%	*supposition*	46	164	28.05%
fear	1,432	1.01%	*impression*	2,279	8,206	27.77%

As Table 5 shows, the rank list for attraction is dominated by fairly common, i.e. more or less frequent, nouns. The reason for this is that the overall frequency of a noun in the corpus does of course have an effect on the likelihood of its occurring in any construction. From a purely statistical point of view, frequent nouns have a better chance than less frequent ones. In contrast, the formula used for calculating reliance takes the total frequency of a noun in the corpus into consideration. As a result, the rank list is headed by fairly infrequent nouns which, however, are highly specialized, so to speak, for occurrence in the given pattern. Intuitively, the semantic affinity between these nouns and the constructions seems to be particularly strong.

A second way of exploiting the notion of reliance is to provide reliance profiles for individual nouns. These give information on the recurrent colligations entered into by a given noun. Table 6 collects a small number of examples:

Table 6. Reliance profiles for the nouns *idea, finding* and *temerity*

Noun	N-to	N-BE-to	N-Th	N-BE-Th	N-Wh	N-BE-Wh	Th-N	Th-BE-N	Freq. in corpus	Compiled reliance
idea	1271	1141	4812	790	752	13	1674	325	46,654	23.10%
finding			96	32			254	7	586	66.38%
temerity	118								160	73.75%

Legend: N-to = N+*to*-infinitive; N-BE-to = N+BE +*to*-infinitive; N-Th = N+*that*-clause; N-BE-Th = N+BE+*that*-clause; N-Wh = N+*wh*-clause; N-BE-Wh = N+BE+*wh*-clause; Th-N = demonstrative determiner+N; Th-BE-N = demonstrative pronoun+BE+N

Table 6 includes absolute scores for frequency of occurrence in the four constructions focused on in this paper as well as in four others in which shell nouns are typically found, two containing *wh*-clauses and two containing demonstrative determiners or pronouns respectively with anaphoric function: N + *wh-clause* (*the reason why ...*), N + BE + *wh*-clause (*the question is why ...*), *this/that* + N (*this problem ...*) and *this/that* + BE + N (*that's the problem ...*). The table shows that *idea* is a highly versatile noun that was found to occur in all eight constructions investigated. However, its score for compiled reliance in the four patterns is below 25%, which means that not even a fourth of its occurrences in the corpus were found in the patterns. *Finding* and *temerity*, on the other hand, boast fairly high scores for compiled reliance, but are less versatile or, to put it more positively, show a much stronger affinity with individual constructions: *finding* is primed (cf. Hoey 2005) for occurrence in anaphoric uses and, to a lesser extent, with *that*-clauses, while *temerity* was only found to occur in the pattern N + *to*-infinitive, but with a very high reliance score of almost three-quarters of its 160 total instances in the corpus.

The attraction-reliance method thus provides a way of gauging the reciprocal interaction between nouns and constructions. It captures to some extent the intuition that some nouns are more important for certain constructions than others, and that some constructions are more important for certain nouns than others. As we will see in Section 7, however, the method has a number of shortcomings with regard both to its rather crude arithmetic and to the interpretation of the output it produces. As new statistical tools for assessing the attraction of lexemes by constructions have been proposed since the publication of Schmid (2000), it will be worth looking into these more advanced statistical techniques before we reflect on the significance of frequency in Section 7.

5.2 Less simple arithmetic: Collostructional Analysis

In a series of papers, Stefanowitsch and Gries introduced a set of so-called "collostructional" methods designed to capture in quantitative terms the mutual attraction of lexemes and constructions.[5] Unlike the attraction-reliance method described in Section 5.1, the collostructional techniques do not simply rely on counts of observed frequencies. Instead they measure the degree of likelihood that the patterns of observed frequencies are due to chance. This can be done by comparing observed frequencies to expected frequencies, which can be calculated using additional scores derived from the corpus.

As the following quotation shows, the test case at hand lends itself very readily to what is known as Collostructional Analysis:

> Collostructional analysis always starts with a particular construction and investigates which lexemes are strongly attracted or repelled by a particular slot in the construction (i.e. occur more frequently or less frequently than expected). (Stefanowitsch and Gries 2003: 214)

The actual measure chosen to gauge the degree of attraction is the p-value of a statistical test known as *Fisher-Exact*.[6] Technically speaking, given a certain set of observations in a corpus, the p-value indicates the probability of obtaining this distribution or a more extreme one, assuming the 'zero-hypothesis' that the distribution was the result of chance. Couched in everyday terms, and as applied to collostructions by Stefanowitsch and Gries, the smaller the p-value, the higher the probability that the observed distribution is not due to coincidence and the higher the strength of the association between lexeme and construction.

Four frequency scores are needed to calculate expected frequencies of lexemes (L) and constructions (C), as well as *p*-values (Gries and Stefanowitsch 2003: 218):

1. the frequency of L in C,
2. the frequency of L in all other constructions,
3. the frequency of C with lexemes other than L, and
4. the frequency of all other constructions with lexemes other than L.

The typical output of the test is a list of 'collexemes' of a construction together with 'their' *p*-values indicating the degree of association. More often than not *p*-values are so small that their significance resides in the number of decimal places, usually expressed as scores to the power of minus x.[7] To simplify things, a logarithmic transformation of these scores can be given, which indicates the number of decimal places. The score illustrated in Note 7 would then simply read '20'.

Attractive as this method is, it is not without its pitfalls. While discussing these shortcomings is beyond the scope of this paper (see Kilgarriff 2005 as well as Schmid and Küchenhoff forthc. for a more detailed critique), one serious hurdle for unbiased applications must be mentioned here: the problem of how to determine the score numbered 4 in the list above (i.e. the frequency of all other constructions with lexemes other than L). This frequency score serves as a mathematical reference point which is necessary for calculating the expected frequencies in the 2-by-2 contingency tables serving as input to Fisher-Exact (or other, simpler zero-hypothesis tests such as the more familiar *Chi-square* test). However, this decision is not simply a mathematical but, more importantly, a linguistic one. The only passage where Stefanowitsch and Gries explicitly address this problem occurs in connection with the construction 'N is waiting to happen' (2003: 218):

> the total number of constructions was arrived at by counting the total number of verb tags in the BNC, as we are dealing with a clause-level construction centering around the verb *wait*.

What this quotation clearly indicates is that in order for the Fisher-Exact test to make sense linguistically, and not just mathematically, it is necessary that the construction investigated and the total number of constructions be paradigmatically related. In a sense, the 'total number of constructions' gives the number of constructions which could potentially also have occurred instead of the construction under investigation. But this paradigmatic relation is not unproblematic. For one thing, the constructions under investigation only occur in the progressive form, so it would have made sense to choose only verbs in the progressive form as reference constructions. Furthermore, *is waiting to happen* is a fairly specific type of construction consisting of a verb complemented, or at least followed, by a *to*-infinitive, and this again might have called for a more narrowly defined type of reference construction.

Analogous problems arise in the application of the collostructional method to the nominal constructions serving as case study here. At first sight, two extreme choices suggest themselves as solutions: one would be to use the total number of noun tags in the corpus (ca. 60,000,000); the other extreme would be to insert only the number of other occurrences of shell-content constructions (i.e. 422,509 minus the number of tokens of the intended type). The latter choice would have the advantage of emphasizing the strong paradigmatic relations in this system, but neglects the fact that other nouns or nominal constructions could occur instead of shell nouns.

As it turns out, neither of these choices is particularly satisfying. If a score of 60 million, representing all nouns in the corpus, is entered in the formula for Fisher-Exact, the calculations will be so demanding that they go way beyond the capacity of normal computing systems, thus yielding a *p*-value of 0 (i.e. infinite likelihood). There is not just a problem with computing power, however, but also one related to the nature of statistical significance testing, as an increase in the size of the sample, i.e. the corpus, investigated also raises the degree of confidence that the differences between observed and expected frequencies are significant and robust, thus rendering even arbitrary associations significant (Kilgarriff 2005: 266). Using the score of 422,509 minus x, on the other hand, does not seem to do justice to the substantial size of the total reference corpus, which, after all, provides many more opportunities for constructions comparable to shell-content constructions to occur.

In view of these difficulties, I have decided to use two different reference scores in applications of the collostructional method in this paper. One is the score 422,509 minus x, because this score at least seems to have some kind of linguistic justification. For a second reference score, the completely arbitrary number of 10,000,000 was chosen, since it was large enough to reflect the massive size of the corpus but is still manageable to some extent as regards capacity. While the choice of an arbitrary number may seem rather odd, from a statistical point of view it is no problem as long as the same score is used for all lexical items tested in one construction. As the application of *p*-values as a measure of attraction strength is controversial anyway (Stefanowitsch and Gries 2003: 239; cf. Schmid and Küchenhoff forthc.), and as, therefore, the ranking of items is much more important than the actual size of the *p*-value, there is not much to be said against such a procedure. Table 7 lists the 10 top-ranking nouns each for attraction and reliance from Table 5 above and gives their *p*-values for both reference scores. The nouns are ordered according the *p*-values in the column on the far right.

Table 7. Attraction, reliance and *p*-value scores for selected nouns in the construction N + *that*-clause.

	Freq. in pattern	Freq. in corpus	Attraction	Reliance	*p*-value, reference score 10,000,000	*p*-value, reference score 281,033
fact	26,106	68,472	18.45%	38,13%	0	0
evidence	5,007	34,391	3.54%	14,56%	0	0
idea	4,812	46,654	3.40%	10,31%	0	0
view	3,532	37,468	2.50%	9,43%	0	0
hope	2,727	16,663	1.93%	16,37%	0	0
news	2,572	49,736	1.82%	5,17%	0	0
feeling	2,511	14,392	1.77%	17,45%	0	0
possibility	2,232	12,075	1.58%	18,48%	0	3,48E-276
doubt	4,010	17,322	2.83%	23,15%	0	2,71E-166
realization	820	1,185	0,58%	69,20%	0	3,01E-139
belief	3,696	9,344	2.61%	39,55%	0	1,52E-41
assumption	1,391	3,151	0,98%	44,14%	0	1,45E-36
impression	2,279	8,206	1.61%	27,77%	0	4,39E-25
assertion	596	1,492	0,42%	39,95%	0	1,25E-07
insistence	796	2,069	0,56%	38,47%	0	1,14E-06
proviso	111	250	0,08%	44,40%	2,95E-134	0,00036
suggestion	2,033	5,854	1,44%	34,73%	0	0,0118
premise	274	765	0,19%	35,82%	5,39E-297	0,16743
misapprehension	44	123	0,03%	35,77%	3,94E-49	0,6329
dictum	84	249	0,06%	33,73%	9,92E-90	0,95

The juxtaposition of the two systems allows for a number of interesting observations. Firstly, at least with the online statistics lab used for calculating the Fisher-Exact test, it is impossible to capture differences in association strength for a considerable number of the nouns included in the table. This is true for calculations with either reference score, and is due to the large size of the corpus used and the resulting high scores for total frequency reached by many nouns (cf. Gries 2005: 278–279, Kilgarriff 2005: 272–273). Secondly, comparatively high p-values close to 1, which indicate low strengths of attraction, are produced by the test especially for nouns with a fairly low overall frequency of occurrence, even if their reliance scores are quite high (cf. the scores for *proviso, premise, misapprehension* and *dictum*). Thirdly, high-frequency nouns with rather low reliance scores such as *view* or *news* leave the test with the same score, i.e. 0, as high-frequency nouns with much higher reliance scores (e.g. *fact*) (see Section 6.3 below). It will be useful to keep these observations in mind when we now turn to a discussion of the cognitive aspects of frequency counts in corpora.

6. Does frequency really instantiate entrenchment?

6.1 Background

In line with the terminological decisions made in Schmid (2007b),[8] the notion of entrenchment is defined as "the degree to which the formation and activation of a cognitive unit is routinized and automated". Within Cognitive Linguistics, both this notion of entrenchment and the idea that entrenchment correlates with frequency of occurrence can be traced back to Langacker. According to him, there is 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. (Langacker 1987: 59)

As this indicates, Langacker conceives of entrenchment as being fostered by repetitions of cognitive events, i.e. by "cognitive occurrences of any degree of complexity, be it the firing of a single neuron or a massive happening of intricate structure and large-scale architecture" (1987: 100). This seems highly convincing, not least in view of the considerable body of evidence from psycholinguistic experiments suggesting that frequency is one major determinant of the ease and speed of lexical access and retrieval, alongside recency of mention in discourse (cf., e.g., Sandra 1994: 30–31, Schmid 2008, Knobel, Finkbeiner and Caramazza 2008). As speed of access in, and retrieval from, the mental lexicon is the closest behavioural correlate to routinization, this indeed supports the idea that frequency and entrenchment co-vary.

Nevertheless, it is not easy to transfer Langacker's idea of 'massive happenings of intricate structure' to larger and complex linguistic units, since it does not seem to take into consideration that the different components of complex linguistics structures may in fact activate each other. For example, it is not unlikely that shell nouns may trigger their recurring shell content clauses, or that the clauses may trigger certain shell nouns. The firing Langacker is talking about may therefore not take place in one go but rather in a cascade-like fashion, with one element triggering one or more other elements.

Another problem with Langacker's view is that he apparently conceives of frequency in a vacuum. However, as Geeraerts, Grondelaers, and Bakema (1994) argue, it is not frequency of use as such that determines entrenchment, but frequency of use with regard to a specific meaning or function, in comparison with alternative expressions of that meaning or function.[9] Like Brown (1965: 321), Rosch (Rosch *et al*. 1976: 435) and Downing (1977: 476) before them, Geeraerts, Grondelaers and Bakema pursue lexicological rather than grammatical goals and investigate the relation between the privileged basic level of categorization and the frequency with which objects are named with terms on this level as opposed to more general superordinate or more specific subordinate terms. This relative frequency is indicative of what they call "onomasiological salience".

As already mentioned in the introduction to this paper, the "From-corpus-to-cognition principle" somewhat daringly proposed in Schmid (2000: 39) was inspired by Halliday's claim that "frequency in text instantiate[s] probability in the [linguistic HJS] system" (Halliday 1993: 3). Partly responding to legitimate objections that the implications of the from-corpus-to-cognition principle were far from clear (cf. Esser 2002: 208), Mukherjee takes up the catchphrase *from corpus to cognition* and tries to refine it (cf. Mukherjee 2005: 67, 91, 247): "From a cognitive point of view, frequency in usage should be best regarded as a quantitative signpost of the degree of entrenchment" (2005: 225). More precisely, what frequency counts in a corpora reflect more or less directly are degrees of *conventionalization* of linguistic units or structures. Conventionalization, however, is a process taking place first and foremost in social, rather than cognitive, systems, and it requires an additional logical step to assume that degrees of conventionalization more or less directly translate into degrees of entrenchment. The crucial link of course is frequency of usage and exposure, which on the one hand reflects degrees of conventionalization in the speech community and on the other hand enhances entrenchment in individual minds (see Schmid forthc. for more details).[10]

All these attempts to correlate frequency with entrenchment have two things in common: they presuppose rather than explicitly question knowledge about the nature of frequency and they treat frequency as a monolithic concept. More or less the same goes for the considerable body of literature in grammaticalization theory that tries to relate the frequency of linguistic units to their propensity to grammaticalize.[11] One notable exception, which will be taken up in the next section, is Hoffmann's (2004) paper on the grammaticalization of low-frequency complex prepositions, which emphasizes the need to be clearer about "what exactly is meant by 'frequency'" and "what is the relationship between frequency and salience" (2004: 189).

6.2 Types of frequency

Hoffmann (2004) distinguishes two kinds of frequency, one with two subtypes. The first type is called *conceptual frequency* and is reminiscent of Geeraerts, Grondelaers and Bakema's notion of *onomasiological salience* mentioned in Section 6.1. As Hoffmann notes (2004: 190), this type is difficult to come to grips with. Its operationalization would require knowledge of the full range of paradigmatic competitors with regard to one function and/or meaning. While this is possible in the lexicon it seems hardly viable to take into consideration all alternative ways of linguistically encoding the function served by a particular lexico-grammatical construction. Since, at least in this respect, grammar – and discourse – are much more open-ended than the lexicon, it does not appear feasible and fruitful to pursue *conceptual frequency* any further in the present study.

The second type is called *lexical* or *textual frequency* and is further subdivided into *absolute* and *relative frequency*. Hoffmann leaves no doubt as to which of the two he finds more important for corpus-linguistic studies:

> [F]requency information for an individual linguistic item only becomes meaningful as a diagnostic tool if it is compared with the frequency of occurrence of related linguistic phenomena. (Hoffmann 2004: 190)

Like Langacker, however, Hoffmann focuses in particular on what Krug (2003) calls "string frequency" and takes little notice of the possibility of assessing the frequency of one component of a construction, say the verb in the ditransitive construction or the shell noun in a shell-content construction, in relation to another to which it is syntagmatically, rather than paradigmatically, related (i.e. the complements). Quite clearly this type of 'relative' frequency differs from the one envisaged by Hoffmann. It is therefore necessary to adapt Hoffmann's classification to the needs of this study.

In line with Hoffmann's proposal, the first type can be called *absolute frequency*, even though it will still be measured as the relative frequency of occurrence of a linguistic phenomenon in a given corpus. This is the only feasible way of operationalizing absolute frequency, since even this measure needs some kind of reference score and has to be quantified. With regard to shell-content constructions, it is possible to measure the absolute frequency of six types of linguistic entities in a corpus:

– Absolute frequency of 1) tokens and 2) types of nouns
– Absolute frequency of 3) tokens and 4) types of complements ('shell contents')
– Absolute frequency of 5) tokens and 6) types of constructions (i.e. nouns in patterns)

The second type of frequency is *relative frequency* (defined in a way different from Hoffmann's). Relative frequency can be approached from the two complementary perspectives introduced in Section 5:

– *Attraction*: the relative frequency of tokens of noun type vis-à-vis the frequency of tokens of construction types.
– *Reliance*: the relative frequency of tokens of noun type in a construction vis-à-vis tokens of the same noun type in other constructions.

Seen from the perspective of the noun, the two types of relative frequencies are relative to other nouns occurring in the same construction (*attraction*) and relative to occurrences of the same noun in other constructions (*reliance*).

With these distinctions in place we are now in a position to examine the relations between different types of frequency and degrees of entrenchment and evaluate the attempts sketched in Section 5 and 6 to quantify this relation.

6.3 Types of frequency, degrees of entrenchment and arithmetic modelling

In view of the overwhelming evidence in cognitive (neuro-)psychology (see Section 6.1) it seems safe to accept that repeated patterns of neuronal activity foster the entrenchment, routinization and activation of the corresponding cognitive events. If this seems plausible enough, then it will also make sense to acknowledge that the absolute frequency of occurrence of types of linguistic entities will show a relationship to the degree of entrenchment of their cognitive and neurological correlates (whatever these may be). This follows from the assumption that frequency of occurrence in discourse relates to frequency of processing in the minds of the members of the speech community.[12]

This is not the whole story, however. In fact, psycholinguistic evidence also exists which suggests that this relation may apply to linguistic forms **irrespective of their function and meaning**. Thus, in a classic study, Swinney (1979) demonstrated that during lexical access, i.e. roughly the first third of a second after being confronted with a word-form, test subjects activate both contextually appropriate and contextually inappropriate meanings of homonyms such as *bug* ('insect' vs. 'overhearing device'). More recent production experiments have even suggested that low-frequency forms (e.g. *nun*) profit with regard to their speed of activation from the existence of high-frequency homophones (*none*), and this in spite of the fact that the two forms represent two different lexemes whose meanings are totally unrelated (Jescheniak and Levelt 1994, Jescheniak, Meyer and Levelt 2003).[13] The frequency determining the ease of lexical access may possibly not be the word-specific frequency (e.g. of *nun* as opposed to *none*) but the cumulative frequency of all the homophonic forms (i.e. frequency of *nun* plus frequency of *none*).

This finding from lexical access studies can be transferred to the problem at hand. If it is true that even homophonic (but not homographic) forms influence each other with regard to entrenchment, then it would also seem very likely that different usage-patterns of one and the same noun lead to cumulative entrenchment. This in turn would suggest that the overall token frequency of nouns in the corpus (in all environments) will have an effect on their entrenchment, both in a certain shell-content construction and in all other environments in which they occur. For example, high-frequency nouns like *time, point* or *way* are most likely more entrenched than less frequent ones (like *disinclination* or *unwillingness*), irrespective of their actual linguistic environment. In stark contrast to received opinion in state-of-the-art (cognitive) corpus linguistics, epitomized for instance in the passage from Hoffmann (2004) quoted in the previous section, this means that there is after all an absolute, *cotext-free* type of entrenchment, which correlates with absolute frequency of occurrence.

The question now is whether *cotext-free entrenchment* is integrated in quantitative accounts of attraction strengths. In the attraction-reliance framework introduced in Section 5.1, absolute frequency is included as a factor, albeit tacitly rather explicitly, since the formula used for calculating attraction scores does **not** include the overall frequency of a given noun in the corpus. As a result, the rank lists for attraction tend to be headed by nouns with fairly high absolute frequencies. Mathematically ill-informed as this clearly is, it may in fact have a certain degree of cognitive plausibility, as it allows absolute frequency scores to influence and even supersede relative frequencies in the patterns. In the collostructional framework (cf. Section 5.2), absolute frequency constitutes an integral part of the calculation of *p*-values, since it is entered in the contingency tables and thus automatically and deliberately taken into account. The scores given in Section 5.2 even suggest that the test exaggerates the effect of absolute frequency mathematically, as higher absolute frequencies increase the confidence in the assessment of the dataset and thus automatically result in lower *p*-values, i.e. higher attraction scores.

If absolute frequency translates into the cognitive system as *cotext-free entrenchment*, it seems reasonable to think of its relative counterpart as reflecting *cotextual entrenchment*. Very much in line with Hoey's (2005) idea of lexical priming, cotextual entrenchment can be seen as the tendency of one linguistic element or unit to trigger the (co-)activation of one or more other linguistic units or structures in language users' minds, if the former significantly co-occurs with the latter in actual discourse. Elements co-occurring frequently are intuitively held to be more cotextually entrenched vis-à-vis each other than elements rarely found in each other's company.

The trouble with cotextual entrenchment is that, as we have just seen, its strength will inevitably be influenced by the effects of cotext-free entrenchment. Still worse, the strength of this effect is difficult to gauge. Theoretically, the full range of combinations of cotext-free and cotextual entrenchment are possible. To facilitate further discussion of the interaction of cotext-free and cotextual entrenchment, combinations of extreme values are cross-tabulated in Table 8. They are illustrated with hand-picked everyday examples as well as typical cases of shell-content constructions (which are accompanied by relative frequency scores for occurrence in the respective pattern and absolute frequency scores in the corpus).

Table 8. Theoretical combinations of extremes of cotext-free and cotextual entrenchment

		Cotext-free entrenchment	
		high	low
Cotextual entrenchment	high	*get up*; *fact* + *that*-clause (26,106 out of 68,472)	*with kith and kin*; *disinclination* + *to*-infinitive (45 out of 62)
	low	*get low*; *way* + BE + *to*-infinitive (316 out of 201, 366)	*shopgrift a nouse*; *aphorism* + BE + *that* (1 out of 81)

The bottom right-hand cell is undoubtedly the one presenting the fewest problems. In everyday terms, this cell describes the occurrence of rare words in uncommon uses. So far the verb *shopgrift* ("the activity of purchasing something from a shop, using it, and then returning it within a specific period in order to get a full refund", Maxwell 2006, s.v. *shopgrifting*) and the noun *nouse* ("a pointing mechanism for a personal computer which is activated by movements of the nose", Maxwell 2006, s.v. *nouse*) are hardly established neologisms with a low frequency of occurrence.[1] In addition, their combination is odd, to say the least. In a similar vein, *aphorism* is a fairly rare noun in the COBUILD corpus and was found to occur only once in the pattern N + BE + *that*-clause. While the noun itself may well be entrenched in some people's minds, for example literary scholars or teachers of rhetoric,[15] it can hardly be considered a salient lexeme, neither in this construction nor elsewhere.

Similarly straightforward, but complementary cases are captured in the top left-hand cell. There can be no doubt that *fact* is deeply entrenched in most adult speakers' minds. The high proportion of uses in the pattern N + *that*-clause (26,106 out of a total of 68,472) also predicts a high level of cotextual entrenchment. In the attraction-reliance framework, this is reflected in a combination of high scores for both attraction and reliance. In the collostructional framework, cases like these are the best candidates for producing *p*-values of 0, which loosely speaking indicates an infinitely high probability that the observed frequencies are not due to chance. In cases of this type, it is a moot point whether the strong scores for attraction are a result of superseding absolute frequency or due to relative frequency. Put rather bluntly, the noun *fact* is entrenched, the N + *that*-clause construction is entrenched, and the lexically filled construction *fact* + *that*-clause is entrenched as well, just as the verb *get*, the particle *up* and the phrasal verb *get up* are entrenched.

The top right-hand cell is a bit more problematic. *Kith* is a very rare linguistic form; therefore it is very likely that it rates low with regard to cotext-free entrenchment. With regard to cotextual entrenchment in the fixed expression *with kith and kin*, however, it clearly rates high, since it has no other habitat to thrive in. Here cotextual entrenchment clearly comes to the fore, as it is not influenced by cotext-free entrenchment. However, extreme cases of this type are more or less restricted to the domain of phraseology. The closest approximation in the area of shell-content constructions is found for low-frequency nouns relying heavily on the pattern N + *to*-infinitive for the occurrence in discourse. The noun *disinclination*, which boasts a reliance score of 73.75% (45 out of 62), is a case in point. Intuitively, examples of this type show the highest degree of semantic affinity with the matching pattern, and thus also of cotextual entrenchment. In the attraction-reliance method, however, this strong affinity only shows up in the reliance scores; in the rank list for attraction, which ranks the 200 nouns found in the pattern N + *to*-infinitive for their frequency in that pattern, *disinclination* occupies rank 193. In the Fisher-Exact test, the *p*-values for cases of this type are astonishingly high, indicating a comparatively low attraction score. The combination *disinclination* + *to*-infinitive yields *p*-values of 2.42e-60 (for a reference score of 10 million) and 0.00321 (for a reference score of 194,244, i.e. all valid tokens of shell-content construction minus the 228,165 tokens of the N + *to*-infinitive construction). The presumably high degree of cotextual entrenchment is not reflected particularly in the second score because – as discussed in Section 6 – low absolute frequencies reduce the confidence of the Fisher-Exact test.

Finally, the bottom left-hand cell is where we can observe how absolute entrenchment can get the better of relative entrenchment. No more than 316 out the mass of 201,366 tokens of the noun *way* in the corpus were found in the construction N + BE + *to*-infinitive, most of them in the more specific patterns *the only way is/was to ...* and *the best way is/was to ...*. Now, while these patterns do sound familiar and are thus most likely cotextually entrenched in most speakers' minds, they are neither typical instantiations of the noun nor of the construction. The construction has a much stronger association with mental and deontic nouns like *aim, intention, ambition, task* and *job* (cf. Schmid 2007a for a semantic analysis of this construction). Here my feeling would be that the enormous cotext-free entrenchment of the semantically highly unspecific noun *way*, which lends itself to uses in a huge range of different patterns, clearly overrides its cotextual entrenchment (which cannot be ignored, however). Do the two quantitative methods capture this effect? The attraction-reliance method lists *way* as ranking 20th (out of 162 types) in terms of attraction, and 135th in terms of reliance. In a sense, this combination of ranks reflects our intuition concerning the effects of cotext-free and cotextual entrenchment, but it does not help a lot in actually quantifying them with any degree of precision. In the Fisher-Exact test, the *p*-values for *way* + BE + *to*-infinitive are in a way complementary to those obtained for *disinclination* + *to*-infinitive: while the latter yielded a decently low *p*-value for a reference score of 10 million and a fairly large one for the closed-system reference score, *way* yields the score 0 for the smaller reference score of 400,624, and 4.86e-9 for the larger one.

Where do we stand now? This comparison of hypothetical patterns of entrenchment and two different attempts to capture them quantitatively has shown that we seem to be quite far from having a good grip on the relation between frequency and entrenchment. This is mainly due to the unclear interaction between absolute and relative frequency, or cotext-free and cotextual entrenchment, respectively. While some patterns of this interaction as manifested in observed frequencies may in fact be reflected quite well in the scores for attraction and reliance, the attraction-reliance method may be criticized for being unable to produce one single score for entrenchment that takes both cotext-free and cotextual entrenchment into account. Even if we accept that the reciprocal attraction of constructions and nouns is a two-dimensional phenomenon that deserves two measures, it still remains a problem that the method works with raw, observed frequencies and does not include any tests of significance. The more sophisticated collostructional method, on the other hand, does exploit statistical tests of significance, relates observed to expected frequencies and also takes absolute frequency into consideration (though possibly with exaggerated effects). However, its application is seriously impeded by the uncertainty concerning the appropriate choice of reference scores, which have a strong effect on the *p*-values indicating the strength of attraction. Furthermore, the exclusive reliance in this method on significance testing risks masking important distributional differences which are very likely also reflections of different entrenchment patterns. This can be illustrated with the help of the fictive examples juxtaposed in Table 9 below.

Table 9. Juxtaposition of different fictive frequency distributions and their reliance scores and *p*-values

High relative frequency – low absolute frequency; reliance = 40.00%		Low relative frequency – high absolute frequency; reliance = 2.26 %	
N in construction	Construction with other nouns	N in construction	Construction with other nouns
40	22,000	113	21,886
100	10,000,000	5000	99,951,41
N in corpus in other constructions	Other constructions with other nouns	N in corpus in other constructions	Other constructions with other nouns
p = 6.74E-72		p = 8.78e-72	

The left-hand side of the table exemplifies a case of high reliance caused by the combination of high relative with low absolute frequency (similar to the type *disinclination* + *to*-infinitive, but less extreme). In contrast, the frequency pattern on the right-hand side shows a moderately frequent noun with a fairly small number of occurrences in the given pattern (reliance score 2.26%). The reference scores in the right-hand cells of the two columns only differ because the grand total (representing the sum of the scores in all four cells) must remain stable in order for the comparison to be correct. The crucial point here is that despite the striking differences in relative vs. absolute frequency, both patterns produce an almost identical *p*-value in the Fisher-Exact test. This means the two fictive nouns would turn out to have identical attraction strengths to the given construction representing identical degrees of cotextual entrenchment, which seems somewhat misleading.

What I have not considered so far in this section are all kinds of combinations of more or less medium values for absolute and relative frequencies, presumably reflecting medium degrees of cotext-free and cotextual entrenchment. While some of the cells included in Table 9 seem to be at least theoretically straightforward, if we are honest we must admit that we know very little about how to deal with these ‘mediocre’ cases. It seems very plausible that combinations with high scores for cotext-free and cotextual entrenchment such as *the fact that ..., the aim is to ..., the problem is that ...* and *the attempt to …* serve metaphorically speaking as conceptual anchors of the respective construction types. They more or less have the status of fixed phrases and are most likely retrieved as one chunk, as suggested by Sinclair's (1991) idiom principle. At the other extreme, combinations of the type *the aphorism that ...* (or even more *shopgrift a nouse*) may in fact catch our attention simply because they are so unfamiliar to us – an effect exploited for rhetorical and stylistic means, e.g. in poetry, journalism and advertising. Whether the huge bulk of combinations between these extremes in fact show the kind of linear proportional correlation between frequency and entrenchment which is usually taken to exist, is an open question.

7. Conclusion

Unfortunately, but also perhaps not surprisingly in view of the preceding discussion, this paper has to end on a somewhat less-than-enthusiastic note. It seems to me that many researchers, inclusing myself, have had a great deal too much confidence in the potential of quantitative methods for the study of aspects of the linguistic and cognitive system. All quantitative methods that I am aware of ultimately boil down to counting the frequencies of tokens and types of linguistic phenomena. What I have tried to show here, however, is that so far we have understood neither the nature of frequency itself nor its relation to entrenchment, let alone come up with a convincing way of capturing either one of them or the relation between them in quantitative terms. This remains true in spite of the indisputable advantages of quantitative methods such as their predictive power, the possibility to falsify models by means of repeat analysis and their enormous capacity when it comes to coming to grips with highly multivariate datasets. Essentially, this failure is caused by the following complications.

Firstly, frequency of occurrence is a much less objective measure than most proponents of quantitative (cognitive) linguistics seem to realize. The assessment of frequency scores depends not only on what researchers retrieve and count as valid tokens, but also on how they calculate frequency. Even if they show awareness of the need to distinguish absolute from relative frequency (as of course most practitioners do), then it is still unclear how the two interact with each other, since absolute frequency may not be as irrelevant as most corpus linguists think. Secondly, advanced statistical techniques, which take absolute frequencies into consideration in order to gauge the significance of observed relative frequencies, have the problem of determining the reference scores required for the tests and run the risk of obscuring different combinations of absolute and relative frequency of occurrence. Thirdly, even if we accept the plausibility of the general claim that frequency of processing, and thus of occurrence in discourse, correlates with strength of entrenchment, we are still underinformed about the relation between cotext-free and cotextual entrenchment. This is particularly true of the large bulk of cases showing a medium range of association of lexeme and construction. Recent attempts at tallying results from corpus studies with results from experimental methods, for example by Gries, Hampe and Schönefeld (2005, 2010) and Wiechmann (2008; cf. also Gilquin and Gries 2009), point to one direction where additional information may be available. While it must be stressed that psycholinguistic experiments represent just another way of trying to tap into the black box, whose 'real' workings will remain hidden to us for some time, converging evidence produced by different methods is undoubtedly superior to results from either corpus or experimental studies.

Notes

1. I would like to thank Joan Bybee, Susanne Handl, Laura Janda, Adam Kilgarriff, Manfred Krug and John Newman for invaluable comments on earlier versions of this paper. I am also indebted to the participants of the theme session at the Krakow ICLA conference (July 2007), of the workshop on "Chunks in Corpus Linguistics and Cognitive Linguistics" Erlangen/Germany (October 2007) and the attendees of a guest lecture at Freiburg University (May 2008) for their input into this study.
2. It should not go unnoted that in lexicography and descriptive grammar the relevance of the frequency of patterns was recognized very early by John Sinclair and taken into consideration in the design of entries in the first edition of the COBUILD dictionary (Sinclair *et al.* 1987) and the first COBUILD grammar (Sinclair 1990). For an account of further developments in lexicography in the 1990s, see Kilgarriff (1997).
3. In fact, as the recent study by Mukherjee (2005) on ditransitive verbs shows, the idea that frequency of occurrence relates to (proto-)typicality is – more or less explicitly – still going strong, despite the debate in the 1980s triggered by

Rosch's (1975) quantitative approach to prototypicality; cf. Schmid 1993: 27–28). For a critique of this approach, see Gilquin (2006).

4. See Schmid (2000: 301-376) for more details on the shared semantic, textual and cognitive functions of shell-content constructions.
5. Cf. Stefanowitsch and Gries (2003), Gries and Stefanowitsch (2004, 2010), Gries (2006a, 2006b) and Stefanowitsch (2005).
6. The Fisher-Exact test is part of most available statistics programmes such as R or SPSS, but it can also be found online; see Wulff (2005) for a useful survey of sites.
7. The scores are usually expressed as, e.g., 2.345e-20, which reads "2.345 to the power of minus 20", i.e. 0.00000000000000000002345.
8. In contrast to *entrenchment*, the notion of *salience* is not taken to refer to degrees of routinization, but either to temporary activation states of mental concepts (referred to as *cognitive salience*) or to inherent and consequently more or less permanent properties of entities in the real world (i.e. *ontological salience*; cf. Schmid, in print). The relation between the two notions is quite complex: on the one hand, ontologically salient entities attract our attention more easily and thus more frequently than nonsalient ones. As a result, cognitive events related to the processing of ontologically salient entities will occur more frequently and lead to an earlier entrenchment of corresponding cognitive units, i.e. concepts. On the other hand, deeply entrenched cognitive units are more likely to become cognitively salient than less well entrenched ones, because a smaller amount of spreading activation will suffice to activate them.
9. As will be shown in Section 7.3, there is psycholinguistic evidence suggesting that absolute frequency of occurrence may be an important factor after all.
10. In contrast to Langacker (2008: 21, fn.13) I consider it important to keep the notions of entrenchment and conventionalization apart. As pointed out by Langacker, entrenchment is a matter of individual minds whereas conventionality and conventionalization are notions pertaining to speech communities. While these two systems, the cognitive and the social, are intricately intertwined, they are governed by different kinds of structures and processes: association, chunking, automatization, generalization and categorization in the cognitive system, as opposed to innovation, accommodation, diffusion and normation in social systems (cf. Schmid forthc.).
11. An early milestone in this tradition is Bybee (1985). Recent publications looking closer into the relation between frequency, grammaticalization tendency and entrenchment or salience include Wray (1999), Croft (2000), Bybee (2001), (2003), (2006), Krug (2003), Hoffmann (2004) and Mair (2004), as well as the collection of articles edited by Bybee and Hopper (2001).
12. This of course does not imply that highly frequent items or patterns automatically correspond to what are known as *prototypes* in Cognitive Semantics (cf. Gilquin 2006), because there are other factors determining prototypicality, e.g.

perceptual salience or conceptual complexity (and because there is no agreement on how to define the notion in the first place).

13. It should be added that the cumulative frequency effect of homophones has been questioned by other researchers, most notably by Carramazza *et al.* (2001) and (2004).
14. Google searches performed on 10 August 2007 produced no more than 79 hits on English-language pages for the form *shopgrift*, and 174,000 hits for *nouse* (The form *mouse* yielded 117,000,000 hits on that day).
15. This is a reminder of Hoey's (2005) important insight that degrees of lexical priming (and thus cotextual entrenchment) are register- and even speaker-dependent (cf. Schmid 2007c).

References

Brown, Roger
1965 *Social Psychology*. New York: Free Press.

Brugman, Claudia M.
1981 *The Story of Over*. Trier: LAUT.

Bybee, Joan
1985 *Morphology. A study of the relationship between meaning and form*. Amsterdam / Philadelphia: John Benjamins.
2001 Frequency effects on French liaison. In *Frequency and the Emergence of Linguistic Structure*, Joan Bybee and Paul Hopper (eds.), 337-359. Amsterdam / Philadelphia: John Benjamins.
2003 Mechanisms of change in grammaticization: The role of frequency. In *Handbook of Historical Linguistics*, Brian D. Joseph and Richard D. Janda (eds.), 602-623. Oxford: Blackwell.
2006 From usage to grammar: The mind's response to repetition. *Language* 82: 711-733.

Caramazza, Alfonso, Albert Costa, Michele Miozzo, and Yanchao Bi
2001 The specific-word frequency effect: Implications for the representation of homophones. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 27: 1430-1450.

Caramazza, Alfonso, Yanchao Bi, Albert Costa and Michele Miozzo
2004 What determines the speed of lexical access: Homophone or specific-word frequency? A reply to Jescheniak *et al.* (2003). *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30: 278-282.

Church, Kenneth W. and Patrick Hanks

1990 Word association norms, mutual information & lexicography. *Computational Linguistics* 16: 22-29.

Clear, Jeremy

1993 From Firth principles. Computational tools for the study of collocation. In *Text and Technology. In honour of John Sinclair*, Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), 271-292. Amsterdam / Philadelphia: John Benjamins.

Croft, William

2000 *Explaining Language Change. An evolutionary approach*. Harlow: Longman.

Dirven, René

1991 Schema and subschemata in the lexical structure of the verb *agree*. *Cahiers de l'Institut de Linguistique de Louvain* 17: 25-42.

Downing, Pamela

1977 On 'basic levels' and the categorization of objects in English discourse. In *Proceedings of the Third Annual Meeting of the Berkeley Linguistics Society*, Kenneth W. Whistler, *et al*., (eds.), 475–487. Berkeley: Berkeley Linguistics Society.

Esser, Jürgen

2002 Review of H.-J. Schmid (2000), *English Abstract Nouns as Conceptual Shells. From corpus to cognition*. In *Anglistik* 13: 204-208.

Geeraerts, Dirk

1983 Prototype theory and diachronic semantics. A case study. *Indogermanische Forschungen* 88: 1-32.

Geeraerts, Dirk, Stef Grondelaers, and Peter Bakema

1994 *The Structure of Lexical Variation. Meaning, naming, and context*. Berlin / New York: Mouton de Gruyter.

Gilquin, Gaëtanelle

2006 The place of prototypicality in corpus linguistics. Causation in the hot seat. In *Corpora in Cognitive Linguistics. Corpus-based approaches to syntax and lexis*, Stefan Th. Gries and Anatol Stefanowitsch (eds.), 159–191. Berlin / New York: Mouton de Gruyter.

Gilquin, Gaëtanelle, and Stefan Th. Gries

2009 Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistics Theory* 5: 1–26.

Goldberg, Adele

1995 *Constructions: A Construction Grammar approach to argument structure constructions*. Chicago: University of Chicago Press.

Gries, Stefan Th.
2005 Null-hypothesis significance testing of word frequencies: A follow-up on Kilgarriff. *Corpus Linguistics and Linguistic Theory* 1: 277-94.

Gries, Stefan Th. and Anatol Stefanowitsch
2004 Extending collostructional analysis: A corpus-based perspective on 'alternations'. *International Journal of Corpus Linguistics* 9: 97-129.
2010 Cluster analysis and the identification of collexeme classes. In *Empirical and Experimental Methods in Cognitive/Functional Research*, Sally Rice and John Newman (eds.), 73-90. Stanford, CA: CSLI.

Gries, Stefan Th., Beate Hampe and Doris Schönefeld
2005 Converging evidence: Bringing together experimental and corpus data on the associations of verbs and constructions. *Cognitive Linguistics* 16: 635-676.
2010 Converging evidence II: More on the association of verbs and constructions. In *Empirical and Experimental Methods in Cognitive/Functional Research*, Sally Rice and John Newman (eds.), 59-72. Stanford: CSLI.

Halliday, M.A.K.
1993 Quantitative studies and probabilities in grammar. In *Data, Description, Discourse. Papers on the English language in honour of John McH. Sinclair*, Michael Hoey (ed.), 1-25. London: Harper Collins.

Herskovits, Anna
1986 *Language and Spatial Cognition: An interdisciplinary study of the prepositions in English*. Cambridge: Cambridge University Press.

Hoey, Michael
2005 *Lexical Priming. A new theory of words and language.* London and New York: Routledge.

Hoffmann, Sebastian
2004 Are low-frequency complex prepositions grammaticalized? On the limits of corpus data – and the importance of intuition. In *Corpus Approaches to Grammaticalization in English*, Hans Lindquist and Christian Mair (eds.), 171-210. Amsterdam / Philadelphia: John Benjamins.

Jescheniak, Jörg D. and Willem J. M. Levelt
1994 Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20: 824-843.

Jescheniak, Jörg D., Antje Meyer and Willem J. M. Levelt

2003 Specific-word frequency is not all that counts in speech production: Comments on Caramazza, Costa, *et al*. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29: 432-438.

Kilgarriff, Adam

1997 Putting frequencies in the dictionary. *International Journal of Lexicography* 10: 135-155.

2005 Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory* 1(2): 263-276.

Knobel, Mark, Matthew Finkbeiner and Alfonso Caramazza

2008 The many places of frequency: Evidence for a novel locus of the lexical frequency effect in word production. *Cognitive Neuropsychology* 25: 256-86.

Krug, Manfred

2003 Frequency as a determinant in grammatical variation and change. In *Determinants of Grammatical Variation in English*, Günter Rohdenberg and Britta Mondorf (eds.), 7–67. Berlin / New York: Mouton de Gruyter.

Lakoff, George

1982 *Categories and Cognitive Models*. Trier: LAUT.

1987 *Women, Fire, and Dangerous Things. What categories reveal about the mind*. Chicago / London: University of Chicago Press.

Langacker, Ronald W.

1987 *Foundations of Cognitive Grammar,* Vol. 1. *Theoretical prerequisites*. Stanford: Stanford University Press.

Lindner, Susan J.

1982 *A lexico-semantic analysis of English verb particle constructions with* out *and* up, PhD Dissertation, San Diego, University of California.

Mair, Christian

2004 Corpus linguistics and grammaticalization theory: Statistics, frequencies, and beyond. In *Corpus Approaches to Grammaticalization in English*, Hans Lindquist and Christian Mair (eds.), 121-150. Amsterdam / Philadelphia: John Benjamins.

Matthiessen, Christian M.I.M.

2006 Frequency profiles of some basic grammatical systems: An interim report. In *System and Corpus: Exploring connections*, Geoff Thompson and Susan Hunston (eds.), 103-142. London: Equinox.

Maxwell, Kerry

2006 *From* al desko *to* zorbing. *New words for the 21st century*. London: Pan MacMillan.

Mukherjee, Joybrato
: 2005 *English Ditransitive Verbs. Aspects of theory, description and a usage-based model*. Amsterdam / New York: Rodopi.

Rosch, Eleanor
: 1975 Cognitive representations of semantic categories. *Journal of Experimental Psychology* 104: 193–233.

Rosch, Eleonor, Caroline B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem
: 1976 Basic objects in natural categories. *Cognitive Psychology* 8: 382–439.

Rudzka-Ostyn, Brygida
: 1989 Prototypes, schemas, and cross-category correspondences: The case of *ask*. *Linguistics* 27: 613-661.

Sandra, Dominiek
: 1994 *Morphology in the reader's mental lexicon*. Frankfurt/Main: Peter Lang.

Schmid, Hans-Jörg
: 1993 *Cottage and Co., Idea, Start vs. Begin. Die Kategorisierung als Grundprinzip einer differenzierten Bedeutungsbeschreibung*. Tübingen: Niemeyer.
: 2000 *English Abstract Nouns as Conceptual Shells. From corpus to cognition*. Berlin / New York: Mouton de Gruyter.
: 2007a Non-compositionality and emergent meaning of lexico-grammatical chunks: A corpus study of noun phrases with sentential complements as constructions. *Zeitschrift für Anglistik und Amerikanistik* 3(3): 313-340
: 2007b Entrenchment, salience and basic levels. In *The Oxford Handbook of Cognitive Linguistics*, Dirk Geeraerts and Hubert Cuyckens (eds.), 117-138. Oxford: Oxford University Press.
: 2007c Review of Hoey (2005). *Lexical priming. A new theory of words and language.* London: Routledge. In *Anglia* 125(2): 339–342.
: 2008 New words in the mind. Concept-formation and entrenchment of neologisms. *Anglia* 126(1): 1–36.
: forthc. Puzzling over entrenchment and conventionalization. In *Constructions – Collocations – Patterns*, Thomas Herbst, Hans-Jörg Schmid and Susen Faulhaber (eds.)

Schmid, Hans-Jörg and Helmut Küchenhoff
: forthc. Looking behind the scenes of collostructional analysis.

Schulze, Rainer
: 1988 A short story of *down*. In *Understanding the Lexicon. Meaning, sense and world knowledge in lexical semantics*, Werner Hüllen and Rainer Schulze (eds.), 395–414. Tübingen: Niemeyer.

Sinclair, John M.
1990 *Collins COBUILD English Grammar*. London: Collins.
1991 *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John M., *et al.*
1987 *Collins COBUILD English Language Dictionary*. London: Collins.

Stefanowitsch, Anatol
2005 New York, Dayton (Ohio), and the raw frequency fallacy. *Corpus Linguistics and Linguistic Theory* 1(2): 295-301.

Stefanowitsch, Anatol, and Stefan Th. Gries
2003 Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics* 8(2): 209-243.

Stubbs, Michael
1995 Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language* 2: 23-55.

Swinney, David A.
1979 Lexical access during sentence comprehension: (Re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior* 18: 645-659.

Tognini Bonelli, Elena
2001 *Corpus Linguistics at Work*. Amsterdam / Philadelphia: John Benjamins.

Tummers, Jose, Kris Heylen, and Dirk Geeraerts
2005 Usage-based approaches in Cognitive Linguistics: A technical state of the art. *Corpus Linguistics and Linguistic Theory* 1(2): 225-261.

Wiechmann, Daniel
2008 On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4(2): 253–290.

Wulff, Stefanie
2005 Online statistics labs. *Corpus Linguistics and Linguistic Theory* 1(2): 303-308.