On the computation of collostruction strength: Testing measures of association as expressions of lexical bias

DANIEL WIECHMANN

Abstract

Collostruction strength, i.e. the degree of attraction that a word C_j exhibits to a construction C_k, has been argued to be exploited in processes of online comprehension, for example, to parse ambiguous structures. There are, however, many ways to express this quantity and a large body of candidate measures can be found in the computational and corpus linguistic literature. The present study provides a comprehensive empirical evaluation of 47 competing (variants of) measures of association in order to assess their usefulness for models of sentence comprehension. To that end, the degree of adequacy of a given measure is evaluated against its performance in a task of predicting human behavior in an eye-tracking experiment that investigated the reading of a local syntactic complementation ambiguity (Kennison 2001). The analysis shows that individual measures in fact arrive at different estimations of degrees of attraction between verbs and the relevant complementation patterns, and hence differ in their power to predict human reading behavior. On the basis of the obtained results, it is suggested that minimum sensitivity (Pedersen and Bruce 1996) is best suited as an expression of collostruction strength.

Keywords: *collostruction strength; association measures; language processing; sentence comprehension.*

1. Introduction

> *The brain has to decide upon actions in a competitive, chance driven world, and to do this well it must know about and exploit the non-random probabilities and interdependencies of objects and events signaled by sensory messages.*
> Barlow (2001: 241)

Corpus Linguistics and Linguistic Theory 4–2 (2008), 253–290
DOI 10.1515/CLLT.2008.011
1613-7027/08/0004-253

The last twenty years have led to slow but steady changes regarding many core assumptions in the study of language, both theoretically and in terms of methodology. The advent of what might broadly be called the *experience-* or *usage-based view* has provided new answers to central questions including what language might be, how it should be studied, and what a linguistic theory should be able to account for. Specifically, instead of viewing language as an autonomous cognitive system, best studied by means of introspective data, and attempting to develop a minimal, i.e. most economical, description of what is referred to as 'core grammar' (cf., e.g. Chomsky 1995), many researchers today view language as deeply grounded in general cognition, best studied using empirical methodologies, and attempting to develop a maximalist grammar, i.e. a grammar describing "the full set of particular statements representing a speaker's grasp of linguistic convention, including those subsumed by general principles" (Langacker 1987: 46).

Close inspection of actual language use has revealed the gradient nature of virtually all linguistic categories and eventually has led researchers to cast into doubt the usefulness of categoricity as a desirable property of linguistic theories (Bod et al. 2002 presents probabilistic approaches to linguistics). Hence, most researchers in this approach highlight the probabilistic nature of linguistic knowledge and develop corresponding non-categorical theories. The ample body of available empirical evidence surely gives us some confidence in asserting that frequency information influences not only the acquisition and processing but also the shaping of grammar over historical time (cf. Diessel 2007 for a comprehensive survey of frequency effects in language). If we accept the idea *that* linguistic knowledge comprises information about frequency of use, it seems required that we next turn to questions addressing *what* psychological mechanisms are susceptible to frequency information and *how* to express the corresponding biases produced by these mechanisms in models of linguistic competence.

The present study aims at providing empirical evidence to shed some light on the latter issue. While some scholars today still defend metrics that take raw or relative co-occurrence frequencies of a pair of forms to be the most informative kind of quantitative information (e.g. Krug 1998; Goldberg 1999; Goldberg et al. 2004), it has been argued that such estimations are problematic and that many language related parameters are better expressed in terms of measures of association (for discussion cf. Manning and Schütze 1999: Ch. 5; Gries et al. 2005 also present empirical evidence for the superiority of collocational measures over raw frequencies). However, as a large number of measures of association has been suggested in the literature, the main goal of this study is to evaluate this wide array of measures and disclose which measure best captures the information used to anticipate syntactic patterns in language comprehension. To this end the measures' estimations of association strength between lexical items and syntactic patterns they can occur

in is evaluated against data from a recent on-line eye-tracking experiment (Kennison 2001).

The study is structured as follows. The remainder of this section is dedicated to providing the necessary theoretical background: section 1.1 exposes the cognitive relevance of measures of association, specifically their connection to processes of on-line comprehension. Sections 1.2 and 1.3 describe the general logic of association measures, the type of data needed, and their role in studies of collocativity and by extension 'collostruction strength', i.e. the relationship of lexical items and syntactic patterns as viewed from a construction grammar perspective. Section 2 describes the corpus data that were used in the study (section 2.1), introduces the tested measures (section 2.2.) and presents their application in the attempt to estimate verb subcategorization preferences (section 2.3). Section 2 closes with an illustration of how the measures' outputs relate to each other using a hierarchical agglomerative cluster analysis (section 2.4). Section 3 is dedicated to offering an empirical evaluation of the measures' performance: the measures are evaluated against a task to predict human behavior in a psycholinguistic experiment, specifically fixation times from an eye-tracking experiment reported in Kennison (2001). Finally, Section 4 concludes the study.

1.1. Human cognition is based on probabilistic processing

> *The general idea is an old one, that any two cells or systems of cells that are repeatedly active at the same time will tend to become "associated,"*
> Hebb (1949: 70)

The idea to view language as a probabilistic system of gradient rules is still uncommon in many areas of linguistics (but cf. Bod et al 2001). However, looking for statistical regularities in the environment, in this case the ambient language, and their exploitation for human language processing seems to be a very natural move given that a) such mechanisms have long been uncovered in adjacent areas of human cognition that are better studied – such as vision – and b) a great many of the most central tenets in cognitive linguistics can be characterized as rather straightforward transfers from cognitive psychological research – above all research into vision – (cf. Rao, Olshausen, and Lewicki 2002 for an illustration of the strong consensus that human cognition is based on probabilistic processing; cf. Sinha 2007 for a detailed description of cognitive linguistics as it relates to cognitive psychology and cognitive science in general). At least since Mach (1886) and Pearson (1892) – and maybe most prominently with Helmholtz's Gestalt laws (Helmholtz 1925) – it has been held that the brain utilizes environmental regularities for perception and the development of symbolic working models of the external world. Brunswik (1956) suggested that all Gestalt laws could be derived from a set of statistical inferences employed in human perception. So we may conjec-

ture that just as the massive redundancy of visual patterns is exploited by the visual system to improve object recognition (cf., e.g., Elder and Goldberg 1998; Geng and Behrmann 2006 and references therein), statistical regularities in the ambient language are exploited by the linguistic system. As a matter of fact, there is indeed a growing body of evidence for this idea from both language learning and on-line processing (cf., e.g., Redington et al. 1993; Bates and Elman 1996; Seidenberg 1997; Christiansen and Chater 1999; for the relevance of probabilistic information for language processing cf., e.g., Jurafsky 1996; Brants and Crocker 2000; Lapata et al. 2001). The general advantage of probabilistic processing is readily illustrated when we turn to the domain of parsing, i.e. the recovery of the syntactic structure underlying an incoming string of words. Obviously language signals cannot be perceived holistically – all at once – but need to be interpreted in a serial and incremental fashion. In morphologically poor languages like English, this leads to a massive ambiguity problem. Consider (1)–(3):

(1)	The old man the boat.	[POS ambiguity]
(2)	Since Steven always jogs a mile doesn't seem too long.	[clause boundary ambiguity]
(3)	The student knew the answer …	[NP/S-ambiguity]
	(a) … to the problem.	[NP is direct object]
	(b) … was incorrect.	[NP is $\text{subject}_{\text{embedded clause}}$]

In each of these cases, the system faces a so called "local syntactic ambiguity", which is characterized by the fact that a given string of words is temporarily ambiguous at some point before the completion of the sentence, whereas the global structure is well-defined: example (1) is likely to cause difficulties because of its unusual – viz improbable – assignment of part-of-speech (POS) information, i.e. *man* is much more frequent as a noun than as a verb (and this is particularly true when it is preceded by the string *the old*).[1] Consequently, we would surely expect the human language comprehension system to be irritated, would it in fact make use of probabilistic cues as these cues would lead it up a garden path. In example (2) the presentation of the linguistic material following the string *the old man*, i.e. *the boat*, is clearly inconsistent with the hypothesized structure, thus forcing the system to find an alternative interpretation compatible with the input.[2] Similarly, (2) is likely to lead to a disruption of the parsing process because *jog* tends to prefer a monotransitive syntax and the string *a mile* meets the syntactic and semantic requirements of a plausible direct object. The intransitive usage of *jog* in (2) is hence likely to lure its comprehender into missing the correct clause boundary. Finally, (3) illustrates the phenomenon investigated in the present study, the so called NP/S ambiguity, which arises whenever a postverbal NP, in this case *the answer*, could either function as the direct object of the perceived verb (as in (3a)), or as the subject of an embedded

clause, which then, as a whole, functions as the direct object of that verb (as in (3b)).

In all these cases, we might try and assess the biases, and correspondingly the processing preferences, in (1)–(3) by calculating the preferences of *man* being a noun or a verb, the preference of *jog* being used transitively or intransitively, or the preference of *know* to appear with nominal or sentential complements. This general idea to integrate probabilistic knowledge into models of sentence comprehension has been invoked in many accounts of language processing over the last twenty years or so. Maybe the most prominent exponent of such accounts is MacWhinney's Competition Model (cf., e.g., MacWhinney 1987; MacWhinney and Bates 1989). A central notion within this model (and connectionist models in general) is the notion of "cue validity". The general idea here is that human cognitive processing involves the assessment of the information value of a stimulus (or property of a stimulus). Cue validity is conceived of as "an objective property of the [. . .] perceptual environment relative to some orgamismic state". In addition to cue validity (and its constitutive notions), such connectionist models also postulate a subjective property of the organism that is referred to by the term "cue strength". Cue strength expresses the probability an organism attaches to a piece of information to signal another piece of information. Architecturally, cue strength is the connection strength between two units in a network and may be conceived of as being the notion that corresponds best to the quantity of interest here. So, conceptually, the ideas behind association measures (and their application here) are closely linked to the notion of cue validity and related notions.[3] Even though cue validity and cue strength are mathematically well-defined (in terms of conditional probabilities) and their definitions differ from any measure tested in the present study, they are helpful to understand the general approach taken here.

While this general line of argumentation is easily appreciated, we still need to settle on an adequate way of estimating these preferences. One promising way to estimate such biases—the one pursued in this study—is to apply measures of association, which may be conceived of as expressions of the amount of "glue" between units. The next section will provide a closer look at these measures.

1.2. Measures of association

Measures of association have been employed in many areas of linguistic inquiry, but predominantly in studies of collocativity (cf., e.g., Berry Rogghe 1973; Church and Hanks 1990; Church et al. 1991; Biber 1993; Evert 2004). Identifying collocates has been shown to be valuable in numerous contexts: for example to differentiate between near synonyms in lexicographic or lexical semantic studies; similarly, automatic collocate extraction is helpful in

many data mining tasks and in contexts of machine translation. For the most part, studies examining the degree of collocativity of linguistic units have focused on lexical items, but the idea has also been generalized to other relationships: under the label of 'colligation', scholars in the Firthian tradition have for example investigated the grammatical company a word keeps (or avoids keeping), the grammatical functions that a word prefers (or avoids), or the relative position in a sequence that a word prefers (or avoids) (cf. Firth 1957; Hoey 1998). Irrespective of the types of units under investigation, measures of association can be computed from co-occurrence data expressed in a contingency table presented as Table 1:

Table 1. *Input distribution for measures of association*

	v	$\neg v$	
u	O_{11}	O_{12}	R_1
$\neg u$	O_{21}	O_{22}	R_2
	C_1	C_2	N

Table 1 schematically represents the number of observed instances of occurrence of the items under investigation (u, v) in all logically possible scenarios. These observed frequencies of co-occurrence are labeled O_{11}, O_{12}, O_{21}, and O_{22}, respectively. $R_{1,2}$ and $C_{1,2}$ denote the row and column totals while N denotes the sum of all observed frequencies (= our sample size). The information expressed in such tables can be summarized by informationally equivalent quadruples of the form $\{O_{11}; R_1; C_1; N\}$, which we may – again following Evert (2004) – label the 'frequency signature' of u, v. Thus degrees of collocativity between two units u,v can be computed from their frequency signatures. Table 2 illustrates how we can derive the expected frequencies (E_{11}, E_{12}, E_{21}, E_{22}), i.e. the frequencies of occurrence we would expect if u and v were statistically independent (= under the null hypothesis).

Table 2. *Expected frequencies (assuming statistical independence)*

	v	$\neg v$	
u	E_{11} $= R_1 C_1 / N$	E_{12} $= R_1 C_2 / N$	R_1
$\neg u$	E_{21} $= R_2 C_1 / N$	E_{22} $= R_2 C_2 / N$	R_2
	C_1	C_2	N

Generally speaking, measures of association are geared to compare what is observed with what is expected (but cf. Section 2.2 for a qualification of this statement). Their output is an 'association score' (AS) that expresses the degree to which two units are attracted to each other. The next section explains how this general logic can be used to inform research into the interplay of lexis and syntax.

1.3. Associations between words and syntactic frames: collostructional analysis

In traditional collocational analysis, a pair u,v consists of two linguistic units of type WORD, e.g. the English adjective-noun pairs *rich*, *prick* or *poor, schmock* and we can apply some measure of association to estimate the degree of collocativity (or attraction) between the members of that pair. Recently, this general idea has been extended to investigate the syntax/lexis-interface in what has been termed 'collostructional analysis' (Stefanowitsch and Gries 2003). This method takes advantage of a central idea embraced in construction grammar that there is in fact no principled distinction between syntactic patterns and lexical items as both can be viewed as signs, i.e. pairings of form and meaning. Grammar, in this view, is a system consisting solely of signs of varying degree of specificity (cf., e.g. Goldberg 2006 for a recent characterization of this approach). Under this view, it is natural to extend the domain of application of collocational analysis and investigate degrees of collocativity between signs at various levels of specificity, say lexical constructions and argument structure constructions. The strength of association between a particular syntactic pattern and its constituent lexical constructions is referred to as 'collostruction strength'. Table 3 shows the data used to assess the degree of attraction between the verb *give* and the ditransitive construction.

Table 3. *Input collostructional analysis*

	ditransitive construction	other constructions	
give	O_{11}	O_{12}	R_1
other verbs	O_{21}	O_{22}	R_2
	C_1	C_2	N

Collostructional analysis has so far been applied not only to describe the semantics of more abstract constructions on the basis of their (strongly attracted) constituent constructions or to distinguish among alleged syntactic alternations (e.g. Gries and Stefanowitsch 2004), but also to tap into psychological dimensions of language: specifically syntactic priming (Gries 2005),

sentence production (Gries et al. 2006) and comprehension (Wiechmann 2008).

Having sketched (a) the general logic of association measures (AM) and (b) the quantity we would like to express, i.e. collostruction strength, the next step would have to be the identification of appropriate candidate measures, i.e. measures suited to express the quantity at hand. At this point, let me briefly show why this is not without its challenges: in their original methodological description of collostructional analysis, Stefanowitsch and Gries have argued – on theoretical grounds – to make use of Fisher's exact test (FET; cf. Petersen 1996 for a detailed discussion). This measure is widely accepted to be well-suited for language data since it does not require the satisfaction of distributional assumptions that do not fit language data and it can handle sparse data and skewed distributions. The FET outputs a *p*-value as an expression of the total probability of the observed distribution and all more extreme distributions, i.e. distributions that diverge more strongly from the expected one, if the null hypothesis of statistical independence were true. However, as *p*-values are sensitive to sample sizes, the FET does not lend itself to straightforward comparisons of degrees of attraction obtained from samples of different size. In order to afford such comparisons, it is hence required to opt for a different measure less prone to variations in sample size.[4] This single example already suffices to acknowledge the fact that it may be difficult to identify a single, universally applicable measure on theoretical grounds alone. But if we accept the idea to use two (or more) measures as potential expressions of the quantity at hand, i.e. collostruction strength, we will have to deal with some immediate consequences: not only must we prepare ourselves to justify the assumption that going from measure A, say Fisher's exact test, to another measure B, say odds ratios, still has us talk about the same referent quantity, but in addition it also seems hard to anticipate the degree to which different measures will present (quantitatively) different estimations of collostruction strength. While the former problem clearly is a theoretical one and will be neglected here, it appears necessary to turn to empirical testing of candidate measures in order to assess the latter. Finally, testing the performance of a given measure in a specific task is required as there is no guarantee that the statistically most sound measure will yield the best results or as Evert puts it: "[t]he statistical soundness of log-likelihood does not always translate into better performance. A conclusive answer can therefore only come from a comparative empirical evaluation of association measures, which plugs different measures into the intended application" (Evert 2004: 113).

Before we turn to a more detailed description of both data and methods in subsequent sections, it seems convenient to provide a sketch of the general proceedings of this study. On the whole, the analysis proceeds in 4 steps:

I Data retrieval and coding:

a. A set of 21 transitive verbs that potentially occur with both nominal and finite sentential complements (and which hence give rise to NP/S ambiguities) is used in the investigation. This set includes the following verbs:
accept, announce, assume, believe, claim, deny, discover, establish, expect, feel, hear, mention, notice, promise, realize, remember, report, say, suggest, understand, and *write*

b. For each of these verbs, a sample is extracted from a balanced corpus of contemporary British English in which the respective verb from the list above is immediately followed by an NP. Each data point, i.e. token of the 'V NP' sequence, is coded manually indicating whether the relevant verb is followed by a nominal or sentential complement.

II Computation of association scores:
Degrees of association between a given verb from the set above and the two relevant complementation patterns are computed using 47(variants of) measures of association suggested in the literature.

III Classification of association measures:
The estimations of the 47 measures are classified with respect to the (dis-) similarity of their estimations by means of a hierarchical agglomerative cluster analysis.

IV Evaluation of association measures:
To evaluate the candidate measures, their respective association scores are compared to observed ambiguity effects (= reading time deltas) from a recent experimental study, Kennison (2001), which investigated the use of verb information during sentence comprehension. Regression models were used to describe the relationship between collostruction strength and processing difficulty.

2. Data and method

2.1. Corpus data

The study is based on a 17 million words sample of the British National Corpus World Edition (BNC) that is isomorphic to the British component of the International Corpus of English (ICE GB), i.e. it was designed so as to resemble the compilation of text types of the ICE GB corpus but is about seventeen times larger (cf. Nelson 1996 for a detailed description of the properties of that corpus).[5] For 21 transitive verbs, all 'V immediately followed

by NP' patterns were extracted by means of reg(ular)ex(pression)-searches that exploited the POS-tagging of the corpus to identify possible noun phrase constellations. The output of the extraction procedure was manually checked to ensure that only true hits entered into the analysis. Since the verbs to be inspected differ considerably in terms of frequency of occurrence, the following procedure was applied:

- for verbs showing a token frequency greater than 3,000 a random 10% sample was used
- for verbs showing a token frequency between 300 and 3,000 a random sample of 300 items was used
- for verbs showing a token frequency lower than 300 (but greater than 100) all occurrences were used

This amounts to a combined set of 6,417 data points which was coded manually for the grammatical role played by the postverbal NP, viz direct object of the main verb, subject of an embedded clause, or 'other'.[6] The resulting distributions were used to calculate the respective association scores that express the collostruction strengths of the verb in question towards the relevant syntactic patterns.

2.2. Candidate measures

The set of candidate measures to be tested comprises of 47 (variants of) association measures, which may be binned into six classes on the basis of their mathematical properties and corresponding conceptual commonalities and differences (cf. Evert 2004 for a comprehensive discussion of these properties). All measures can be expressed in terms of the concepts used to describe their frequency signatures:

1. Likelihood measures (equate the amount of evidence against the null hypothesis of independence (H0) with the probability (or likelihood) of the observed co-occurrence frequencies under H0)

$$\textbf{multinomial-likelihood} = \frac{N!}{N^N} \cdot \frac{(E_{11})^{O_{11}} \cdot (E_{12})^{O_{12}} \cdot (E_{21})^{O_{21}} \cdot (E_{22})^{O_{22}}}{O_{11}! \cdot O_{12}! \cdot O_{21}! \cdot O_{22}!}$$

$$\textbf{binomial-likelihood} = \binom{N}{O_{11}} \left(\frac{E_{11}}{N}\right)^{O_{11}} \left(1 - \frac{E_{11}}{N}\right)^{N-O_{11}}$$

$$\textbf{Poisson-likelihood} = e^{-E_{11}} \frac{(E_{11})^{O_{11}}}{O_{11}!}$$

$$\textbf{Poisson-Stirling}_{\textbf{log}} = O_{11} \cdot (\log O_{11} - \log E_{11} - 1)$$

$$\textbf{hypergeometric-likelihood} = \frac{\binom{C_1}{O_{12}} \cdot \binom{C_2}{R_1 - O_{11}}}{\binom{N}{R-1}}$$

2. Exact hypothesis tests (compute a *p*-value expressing the total probability of all possible outcomes that are similar to or more "extreme" than the observed one)

$$\textbf{binomial} = \sum_{k=O_{11}}^{N} \binom{N}{K} \left(\frac{E_{11}}{N}\right)^k \left(1 - \frac{E_{11}}{N}\right)^{N-k}$$

$$\textbf{Poison} = \sum_{k=O_{11}}^{\infty} e^{-E_{11}} \frac{(E_{11})^k}{k!}$$

$$\textbf{Fisher} = \sum_{k=O_{11}}^{\min\{R_1, C_1\}} \frac{\binom{C_1}{k} \cdot \binom{C_2}{R_1 - k}}{\binom{N}{R_1}}$$

3. Asymptotic hypothesis tests (use a simpler equation than the exact test to approximate the *p*-value)

$$\textbf{z-score} = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

(z-score for O_{11} against E_{11})

$$\textbf{t-score} = \frac{O_{11} - E_{11}}{\sqrt{E_{11}}}$$

(t-score for comparison of O_{11} against E_{11})

$$\textbf{chi-squared}_{\,i} = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$$\textbf{chi-squared}_{\,h} = \frac{N(O_{11}O_{22} - O_{21})^2}{R_1 R_2 C_1 C_2}$$

$$\textbf{chi-squared} = \frac{N(O_{11} - E_{11})^2}{E_{11}E_{22}}$$

$$\textbf{chi-squared}_{\,h,\,corr} = \frac{N(|O_{11}O_{22} - O_{12}O_{21}| - \frac{N}{2})^2}{R_1 R_2 C_1 C_2}$$

$$\textbf{log-likelihood} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

$$\textbf{log-likelihood}_{\textbf{ratio}} = -2 \log \frac{\max P(\vec{X} = \vec{O} \mid N \wedge \pi = \pi_1 \cdot \pi_2}{\max P(\vec{X} = \vec{O} \mid N)}$$

$$\textbf{log-likelihood}_{\textbf{Dunning}} = -2 \log \frac{L(O_{11}, C_1, r) \cdot L(O_{12}, C_2, r)}{L(O_{11}, C_1, r_1) \cdot L(O_{12}, C_2, r_2)}$$

$$L(k, n, r) = r^k (1 - r) 1^{n-k}$$

$$r = \frac{R_1}{N}, r_1 = \frac{O_{11}}{C_1}, r_2 = \frac{O_{12}}{C_2}$$

(Dunning's log-likelihood ratio (one-sided))

4. Point estimates of association strength (are maximum likelihood estimates (MLE) for various coefficients of association strength)

$$\textbf{MI} = \log \frac{O_{11}}{E_{11}}$$

(MLE of pointwise mutual information)

$$\textbf{odds-ratio} = \log \frac{O_{11} O_{22}}{O_{12} O_{21}}$$

$$\textbf{odds-ratio}_{\textbf{disc}} = \log \frac{(O_{11} + \frac{1}{2})(O_{22} + \frac{1}{2})}{(O_{12} + \frac{1}{2})(O_{21} + \frac{1}{2})}$$

(MLE of discounted (log) odds ratio)

$$\textbf{relative risk} = \log \frac{O_{11} C_2}{O_{12} C_1}$$

(MLE of (log) relative risk coefficient)

$$\textbf{MS} = \min \left\{ \frac{O_{11}}{R_1} \cdot \frac{O_{11}}{C_1} \right\}$$

(MLE of minimum sensitivity coefficient)

$$\textbf{Jaccard} = \frac{O_{11}}{O_{11} + O_{12} + O_{21}}$$

$$\textbf{Dice} = \frac{2 O_{11}}{R_1 + C_1}$$

(MLE of Dice coefficient)

$$\textbf{gmean} = \frac{O_{11}}{\sqrt{R_1 C_1}} = \frac{O_{11}}{\sqrt{N E_{11}}}$$

(MLE of geometric mean coefficient)

5. Conservative estimates of association strength

$$\textbf{MI}_{\textbf{cont},\,\alpha} = \log \inf \left\{ \mu > 0 \mid e^{-\mu E_{11}} \sum_{k=O_{11}}^{\infty} \frac{(\mu E_{11})^k}{k!} \geq \alpha \right\}$$

(conservative estimate for the base 10 log of mutual information; association score is the lower endpoint of a two-sided confidence interval for MI at the significance level α)[7]

6. Measures from information theory (are motivated by the information-theoretic concept of mutual information)

$$\textbf{MI} = \log \frac{O_{11}}{E_{11}}$$

(see above)

$$\textbf{local-MI} = O_{11} \cdot \log \frac{O_{11}}{E_{11}}$$

(contribution to average mutual information of full co-occurrence data)

$$\textbf{average-MI} = \sum_{ij} O_{ij} \cdot \log \frac{O_{11}}{E_{11}}$$

(average mutual information between indicator variables)

7. Heuristic measures

$$\textbf{frequency} = O_{11}$$

(raw co-occurrence frequency)

$$\textbf{MI}^2 = \log \frac{(O_{11})^2}{E_{11}}$$

(MLE of MI with numerator squared)

$$\textbf{MI}^3 = \log \frac{(O_{11})^3}{E_{11}}$$

(MLE of MI with numerator cubed)

Random (no equation can be given)

(random scores between 0 and 1 included as baseline/sanity check)

2.3. Estimations of collostruction strength

In order to assess directly the relative preference of a given verb towards the two complementation patterns involved in the NP/S-ambiguity, the present study employs a method termed 'distinctive collexeme analysis' (DCA; Gries and Stefanowitsch 2004). Table 4 presents the data that need to be collected for each verb to assess its preference for a particular complementation pattern relative to a specified number of well-defined alternatives:

Table 4. *Input data for distinctive collexeme analysis*

	nominal complements	sentential complements	
verb *v*	O_{11}	O_{12}	R_1
other verbs	O_{21}	O_{22}	R_2
	C_1	C_2	N

The calculation of the association scores for each verb was conducted by submitting the relevant data, i.e. the frequency signatures, to statistical analysis. All association scores were calculated using the implementations in the UCS toolkit (Version 0.5). UCS comprises a set of libraries and tools based on two programming languages, Perl and R, and was specifically designed to compute and further manipulate association scores.[8]

Association scores from different measures cannot be directly compared as they are expressed on different scales. Comparisons of different measures

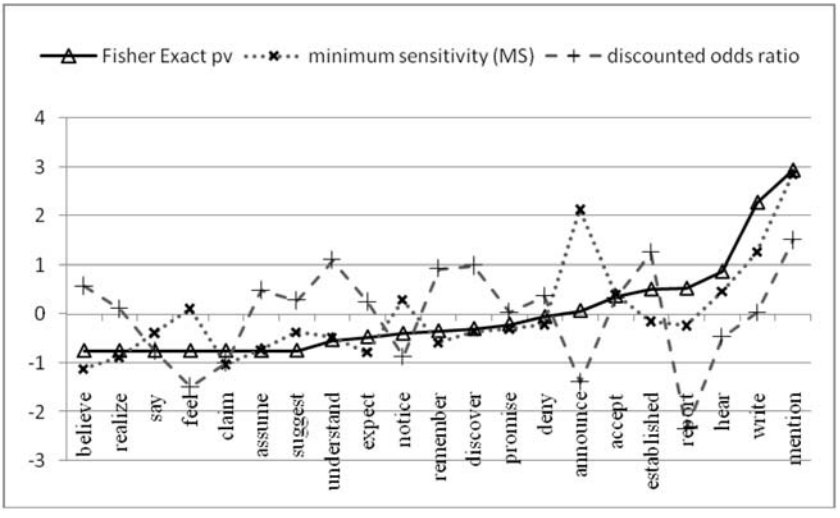

Figure 1. *z-standardized AS (selection)*

are, thus, often performed on the basis of the respective ranks that correspond to their quantitative output. However, going from quantitative output to mere ranking obviously involves a considerable loss of information. In order to allow for direct comparisons without losing this information, the output of all measures was z-standardized to align all AS on a unifying scale.[9] Figure 1 illustrates graphically a selection of the (already standardized) output and shows the estimations produced by the standard measure, Fisher's Exact test (solid line), minimum sensitivity (dotted line), and the discounted odds ratio (dashed line).

The graphs in Figure 1 denote the estimations of the collostruction strengths between the nominal complementation pattern and each of the 21 verbs that have been analyzed here. Due to the z-standardization the interpretation of these association scores is a little tricky; we can, however, say that the greater the score, the stronger the association between a given verb and the nominal pattern (as compared to the finite sentential complementation pattern). We see that for some verbs (e.g. *say* or *claim*) the measures seem to arrive at quite similar estimations, whereas for other verbs (e.g. *report* or *announce*) they differ quite strongly in their estimation of the relationship in question. In order to get a clearer idea of how all tested measures relate to each other in terms of their estimations of collostruction strength, the next section will present a numerical taxonomy of the measure's outputs as computed by means of a hierarchical agglomerative cluster analysis.

2.4. Classification of association measures

The last section has presented a selection of the tested association measures and has shown that individual measures may very well arrive at different estimations of the relationship between a particular pair of units. As we have seen in section 1.3, knowing how different measures relate to each other is vital in cases when our empirical design requires us to find a substitute for the de facto standard test, Fisher's exact test (cf. Gries 2006; Wiechmann 2008). In other words, we want to know which measures produce outputs that can be considered similar enough so as to invite potential substitution should the task at hand require it. The goal of this section is thus to identify the structure underlying the results presented in the last section and to provide a principled classification of the tested association measures. This classification is achieved on the basis of a hierarchical agglomerative clustering (HAC) analysis and subsequent diagnosis of the clustering solutions.

Cluster analysis can be conceived of as a family of techniques that aim at allocating objects to groups (or clusters) on the basis of their (dis)similarity. Simply put: objects that are judged by the clustering algorithm to be very similar are allocated to the same group and dissimilar objects are put in different groups. The variant of cluster analytic techniques that is made use of

here, hierarchical agglomerative clustering, starts off with as many clusters as there are objects to be classified – in our case 48 – and then iteratively classifies these object into larger groups until a single aggregation has been reached.[10] An HAC involves two parts: first, individual objects are grouped together with other objects based on their degree of (dis)similarity. Second, once several objects have been merged, we need to find a way to link them all together to complete the hierarchical structure and include all objects. There are numerous ways to measure a) the distance between objects and b) how to link them and when.[11] Hence, the difficulties for clustering techniques involve questions regarding what measure should be used to a) derive a (dis)similarity matrix relating the objects involved and b) decide how the amalgamation should be performed. For the present purposes, a correlational distance measure (Kendall's Tau) was used to measure degrees of dissimilarity and 'complete linkage' was used for the amalgamation of clusters. 'Kendall' was chosen as a measure of (dis)similarity because it was judged by the author to be more interesting to see how different measures vary in terms of the curvature than in terms of the strength of SUBCAT preferences they assign to a given verb. 'Complete linkage' (or 'furthest neighbor inter-cluster dissimilarity') was chosen because it "tends to produce very compact clusters" (Kaufman and Rousseeuw 1990: 48) and is particularly efficient when the objects form natural distinct groups, which need not necessarily be of equal size.

The goal of the cluster analysis is to assess to what degree our association measures differ in their estimation of the SUBCAT-preferences for our 21 verbs, or – in construction grammar parlance – the collostruction strengths between the 21 lexical constructions and the nominal complementation construction. Thus, the input for the HAC analysis is the set of column vectors (or one-dimensional arrays) that consist of 21 z-standardized association scores. In short, the HAC was computed to compare 48 such vectors so as to decide which measure is how similar to what other measure.[12]

2.4.1. Results and discussion: Hierarchical agglomerative clustering analysis

The results of a HAC are conveniently expressed by means of a dendrogram. While the compared objects are aligned on the horizontal axis, the vertical axis denotes the 'height', i.e. the linkage distance, and for each node in the tree we can determine the distance at which the respective objects were linked together into a new object.

The dendrogram represents the measures that exhibit the highest intra-cluster similarity and the lowest inter-cluster similarity as the ones amalgamated early in the tree. Thus, the tree representation in Figure 2 allows for a convenient comparison of our candidate measures. We can see for example that – at least for the present data – the output of the binomial likelihood

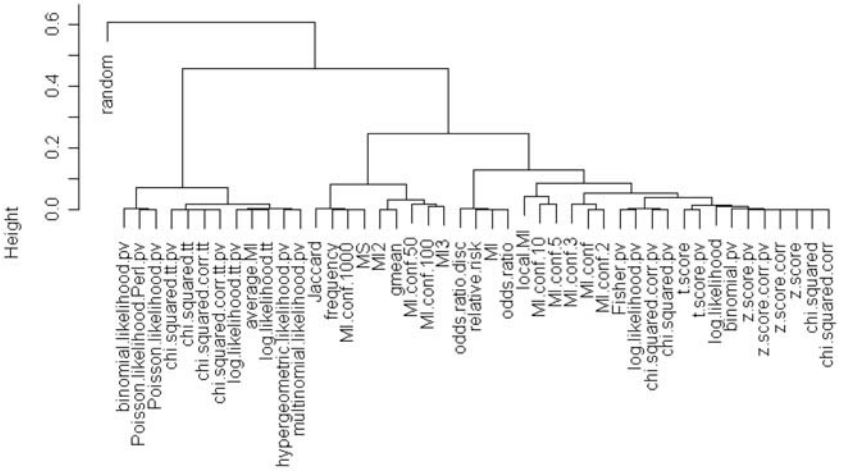

Figure 2. *Classification of association measure*

measure on the very left is most similar to that of (the Perl implementation of) the Poisson likelihood measure and most dissimilar to the corrected chi squared measure (chi.squared.corr).

The dendrogram in Figure 2 provides a rather detailed picture of the structure underlying the data. While this surely is desirable for some purposes, it also introduces new problems: it is certainly far from obvious at what level (i.e. height) we observe the most pronounced structural properties of the data. Let me quickly illustrate the problem. At a height value of approximately .46 we get the most coarse-grained contrast separating all measures to the right of the multinomial likelihood measure from all those to the left of it. The further we go down the tree structure, the higher the resolution of cluster solutions becomes. Now, quite generally, whenever our goal is to partition a set of objects into k clusters, it is desirable to identify (in a data-driven fashion) the ideal value for k. In order to identify the optimal partitioning of the data, we made use of a diagnostic technique that compares all cluster solutions with respect to their 'average silhouette width' (cf. Roossseuw 1987; Kaufman and Roosseeuw 1990: Chapter 2).

Average silhouette width (ASW) is a coefficient that allows the evaluation of cluster solutions on the basis of the optimal ratio of the intra-cluster dissimilarity of the objects within their clusters and the dissimilarity between elements of objects between clusters.[13] The diagnostics of the cluster solutions were conducted using 'Cluster.eval 0.9', a script for R for windows.[14] The script computes the average silhouette width for all partitioning solutions beginning with the minimal one that consists of just two groups to the

most detailed one, which consists of $N_{objects} - 1$, here $48 - 1 = 47$, solutions. Figure 3 presents the results of the diagnostics.

The *y*-axis denotes the average silhouette width of a given clustering solution (ranging from 0 to 1), while the *x*-axis denotes the number of clusters in the solution. Following, Kaufman and Rousseeuw (1990), the ASW-score might be interpreted as follows:

I	0.71–1.00	A strong structure has been found (excellent split)
II	0.51–0.70	A reasonable structure has been found
III	0.26–0.50	The structure is weak and could be artificial
IV	≤ 0.25	No substantial structure has been found (horrible split)

On the basis of these results, three clustering solutions were judged as most adequate: the three groups solution, the seven groups solution, and the eleven-groups solution. Figure 4 presents these three splits.

The top-most split at a height value of roughly 0.35 partitions the objects into 3 groups, indicated by the respective boxes around their constitutive members: the first group consists of only one member, namely the random number "measure", the second one includes everything from the binomial likelihood measure to the multinomial likelihood measure, while the third group consists of all candidates to the right of that measure. While according to Kaufman and Rousseeuw's interpretation of the ASW coefficient this surely is an excellent split (ASW = 0.78), it is obviously also rather coarse-grained. The '7 groups split' (at height $\sim$ 0.85) provides a more fine-grained picture of the underlying structure (ASW = 0.71) and is presented here because it constitutes a local high in the cluster validation graph (cf. Figure 3).[15] Since the seven groups solution still leaves us with a rather large block incorporating all measures from local MI to the corrected chi squared measure, Figure 4 also presents the groupings of the eleven groups split (ASW = 0.77), which divides that group into five constituents and is – due to its high ASW value – preferable over solutions that divide the data into 8, 9, or 10 groups (cf. Figure 3). At this point, we may remind ourselves that cluster analysis in general is an exploratory technique (and not a hypothesis testing one) and so – as all three solutions can be considered excellent splits – we need not decide on a single ideal clustering solution but are well advised to learn from all of them.

Having conducted the HAC, we have expanded our understanding in so far as we now not only know *that* the choice of association measure can lead to different results (cf. section 2.3), but also *how* (dis)similar their respective outputs are. Note, however, that while the clustering provides us with an idea of how scores from different measures relate to each other, it does not tell us anything about the degree of adequacy of a given measure. In order to identify the most adequate one, we need to turn to empirical evaluation. Since the principle goal of this study is to identify the most adequate measure to express

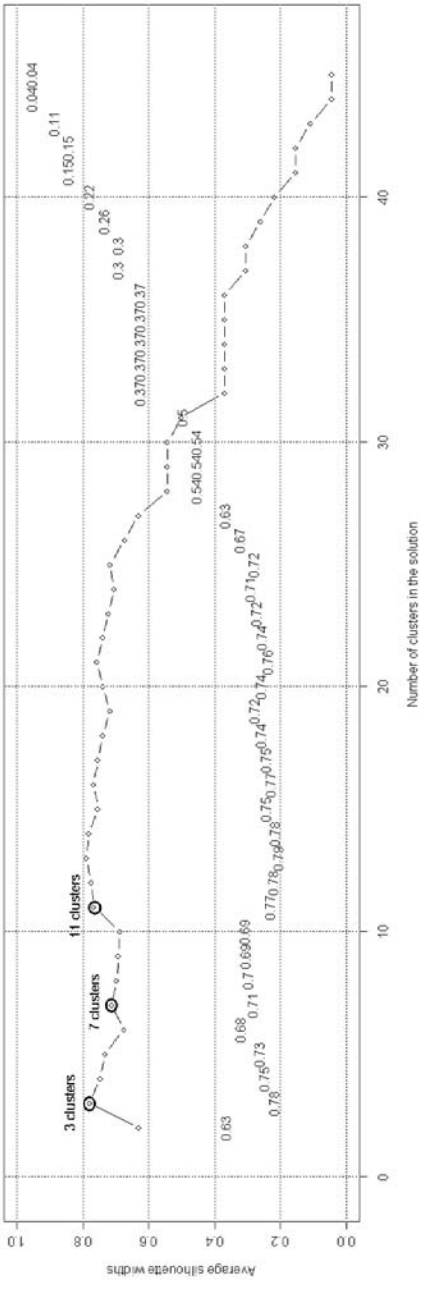

Figure 3. *Cluster validation graph*

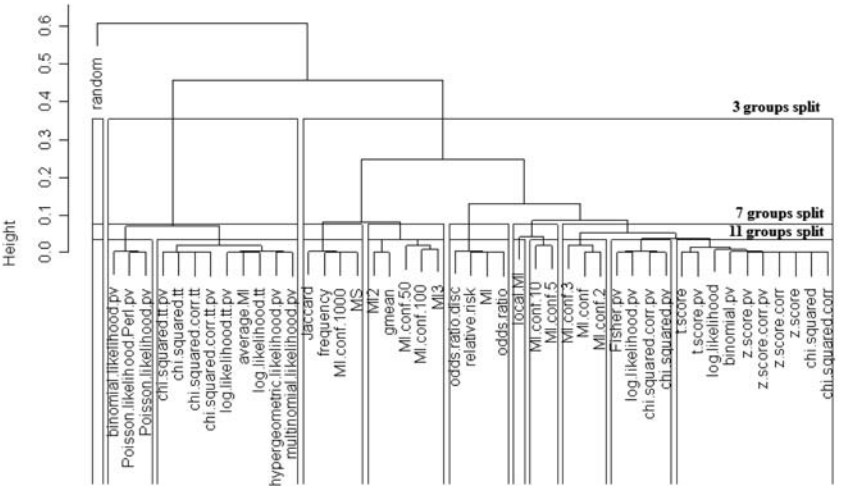

Figure 4. *Cluster diagnostics*

the probabilistic information used in processes of on-line comprehension, we need to evaluate the candidate scores against data from on-line processing.

3. Evaluating association measures

The previous section has demonstrated that we can expect to arrive at different estimations of collocativity (or collostruction strength) depending on the measure we use to express it and it provided us with a classification of competing measures on the basis of the similarity of their respective outputs. This section is dedicated to putting our candidate measures to the test and evaluates them against experimental data from on-line sentence comprehension. Under the assumption that collostruction strength can be viewed as an approximation of the predictiveness of different verbs as indicators of likely syntactic continuations, we can test different measures of association by observing how well they can be used to predict human behavior in an on-line reading task involving the NP/S ambiguity. Before we turn to the experimental data used for the comparison of AM, let me briefly present some necessary prerequisites concerning current theories of language processing.

The psycholinguistic literature to this point presents a large body of theories and models of sentence comprehension. Maybe the most important dimensions along which the accounts vary concern two questions: 1) is parsing a serial or parallel process and 2) what informational sources are tapped at what time (cf. Pickering 1999; Crocker et al. 2006 for an overview of architectures and mechanisms of language processing). In recent years, lexically

driven parsing mechanisms have gained support from both experimental studies (cf., e.g., Adams et al. 1998; MacWhinney and Bates 1989; MacDonald et al.1994; Altmann 1998; Spivey and Tanenhaus 1998) and computational accounts (cf., e.g., Jurafsky 1996; Narayanan and Jurafsky 1998; Manning and Schütze 1999). With respect to the resolution of NP/S ambiguities, these accounts assume that processing a verb with a strong (item specific) subcategorization preference towards some pattern will lead to processing difficulties when the corresponding expectation derived from these lexical preferences is not met. In other words, the recognition of a verb that strongly prefers NP-complementation causes the system to anticipate the presentation of a nominal continuation, leading to processing disruptions when the verb in fact occurs with a sentential complement and vice versa. The idea of a lexically driven parsing mechanism is, however, not without controversy: on the basis of an eye-tracking study, Kennison (2001) suggests that a preference for the nominal complementation can be found even when the preceding verb shows a SUBCAT-bias towards the alternative pattern. Hence, the data in Kennison (2001) are considered to present evidence against an item-specific parsing mechanism. However, the central hypothesis that comprehenders are sensitive to lexically specific biases should not be rejected prematurely. In fact it has been suggested that "the inaccuracies or apparent discrepancies arise because bias is best taken as a reflection of the comprehender's awareness of meaning-structure relationships at the level of individual verb senses" (Hare et al. 2004: 5). In accordance with the findings in Hare et al. (2003, 2004), Wiechmann (2008) concluded that early parsing is driven by sense-contingent preferences and that corpus-based estimations of these biases can be used to predict reading time behavior. Clear effects could only be observed when the preferences were calculated on the level of individual verb senses. Verb general preferences turned out to be insufficient to allow for the prediction of human reading behavior.

Despite the fact that in light of the available evidence, it appears more appropriate to assume that verb-sense specific preferences guide early parsing decisions, the present study will still focus on verb general, form-based preferences. This is partly due to very practical considerations: first, they are a lot easier to derive from corpora since no semantic disambiguation has to be conducted to identify different word senses and, second, the present study is primarily interested in comparing the relative performance of a large number of measures that are potentially suitable as expressions of collostruction strength. It is expected that applying these measures to more fine-grained data, i.e. verb-sense specific distributions, will result in overall improvements of performance. Nevertheless, it is assumed that verb general preferences factor into the systems decision, as they constitute important generalizations at a formal level. This is to say that regardless of whether more informative, i.e. sense-contingent, or more coarse-grained, i.e. verb general, data are used, the

most suitable measure is expected to make the best predictions of processing difficulty relative to its competitors.

The data used in Kennison (2001) are well suited for such a comparison because the test sentences have been presented in isolation, i.e. without any context that might provide the hearer with some cue regarding the most likely sense expressed by a given verb. Hence, if they use item-specific preferences in early parsing at all, hearers are restricted to exploit their knowledge of verb general preferences. The next section presents a closer look at the experimental data and the methodology employed to compare different association measures.

3.1. Experimental data: Kennison (2001)

Kennison (2001) presents an eye-tracking experiment investigating the effects of lexical information during the process of on-line language comprehension. To that end, reading times of temporarily ambiguous sentences exhibiting the NP/S-ambiguity were measured. Experimental sentences were designed that incorporated a total of 51 verbs (24 of which preferred S-complementation whereas 27 were biased towards NP-patterning). These verbs occurred in the experiment exclusively in their respective past tense forms. Table 5 presents the twelve conditions in which the verbs were utilized.

Table 5. *Example of stimuli used in Kennison (2001)*

	Region 1 subject	**Reg 2** verb	**Reg 3** comp	**Reg 4** NP	**Reg 5** modifier	**Reg 6** disamb. Item	**Reg 7** post-disamb.
SC							
NP biased	The athlete	revealed	(that)	his problem	(with drugs)	worried his parents	every single moment.
S-biased	The athlete	admitted	(that)	his problem	(with drugs)	worried his parents	every single moment.
DO							
NP biased	The athlete	revealed		his problem	(with drugs)	because his parents	every single moment.
S-biased	The athlete	admitted		his problem	(with drugs)	because his parents	every single moment.

Subjects were asked to read the respective sentences, and two measures of reading time were analyzed at seven regions (R1–R7): 'total reading time', which is defined as the sum of all fixations in a region, and 'first pass reading time', which is defined as the sum of all fixations in a region from when the eye first enters a region to when the eye first exits a region. The twelve conditions resulted from the variation of some crucial parameters (as indicated in Table 5): first, the VP was either headed by a verb which preferred sentential complementation or by one occurring more frequently with simple transitive syntax. Second, both types of verbs were inserted into three different sentence types: an S-continuation with an overt *that*-complementizer (= unambiguous S), a *that*-less S-continuation (= ambiguous S), or an NP-continuation (= ambiguous NP). Finally, the post-verbal NP was either followed by a modifying PP (= long ambiguous region) or not (= short ambiguous region). The next

section will illustrate exactly what experimental data were used and how they were compared with the corpus-based findings.

3.2. Method

Only 21 of the verbs used in Kennison (2001) occurred at least 100 times in the 17 million word sample of the BNC and were used in the present study. This restriction is motivated by the fact that low token frequencies are more likely to result in inaccurate estimations of collostruction strength.[16] For each of these verbs, fixation times were collected from the regions highlighted in examples (4) and (5). The '|' symbol marks the borders between fixation regions:

(4) The athlete | admitted | (that) | his problem | (with prescription drugs) | **worried his parents** | nearly every single moment. . . . . (S-complementation)

(5) The athlete | admitted | his problem | (with prescription drugs) | **because his parents** | worried every single moment. . . . . (NP-continuation)

Looking at these data for all verbs under investigation revealed that the most interesting region is the beginning of the disambiguating region, i.e. Region 6, which contains the material immediately following the ambiguous NP (the region in bold print in the examples). Averaged across verbs, the greatest ambiguity effects could be observed at short NP continuations and first pass reading time.[17] Consequently, the fixations time deltas for each verb in ms (averaged across subjects) were taken from these conditions and compared with the verb preferences as estimated by the 47 association measures.

In order to assess the performance of the measures under investigation, an attempt was made to see which measure provides the best predictions of fixation times. Thus the corpus-derived estimations of lexical preferences and the relevant data from the psycholinguistic experiment (first pass fixation times, nominal complementation, short disambiguation region, aggregated across subjects) were compared. The relationship between collostruction strength and reading behavior was examined on the basis of regression analyses. Each measure was investigated in isolation, i.e. for each of the 47 measures a set of regression models was built until the best overall type of model could be identified. For reasons of exposition, I will present only those measures that have been classified into distinct groups by the HAC analysis and/or that have played a more prominent role in the literature.

Let me quickly illustrate the steps involved in the general procedure using the string frequency measure as an example: the first step involved in the

analysis was to model the relationship between collostruction strength and processing difficulty – expressed in terms of fixation time differences – as a linear one, simply because Occam's razor requires that – ceteris paribus – the least complex model should be preferred over possible alternatives. Only in the event of a significant increase of the model's adequacy, a more complex model should be preferred. It turns out that the relationship of the two quantities in question is in fact better characterized by a non-linear (quadratic) model than by a simple linear one. Figures 5 and 6 below demonstrate how this was disclosed.

The analysis begins by plotting the output of a given AM against the ambiguity effect observed in the experimental study (extreme outliers have been removed)[18]. The y-axis denotes the z-standardized output of a given measure of association – in this case 'string frequency'; on the x-axis, we see the fixation time deltas in ms (also z-standardized). The dashed line is the linear regression line. The solid line is the smooth line (or line of best fit), i.e. the line that best characterizes the relationship. The striking difference between the regression line and the smooth line suggests that the relationship between the relevant quantities is not best characterized as a linear one. Consequently, more complex models (inverse, quadratic, cubic) were built to fit the data.

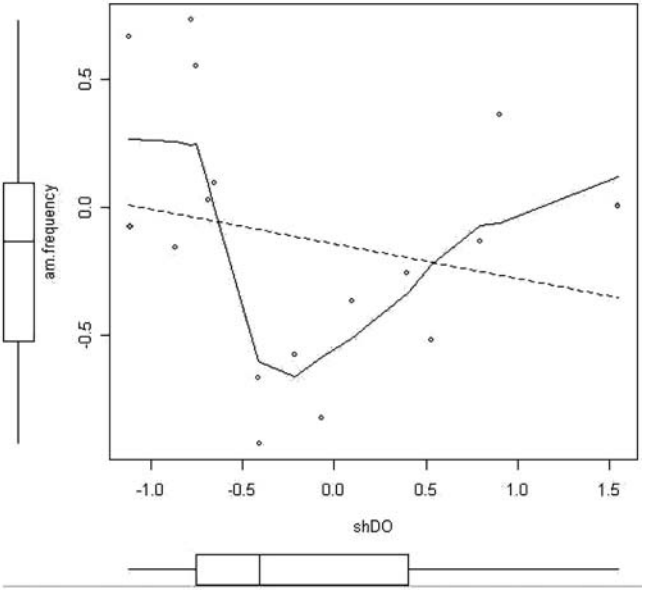

Figure 5. *String frequency versus fixation time*

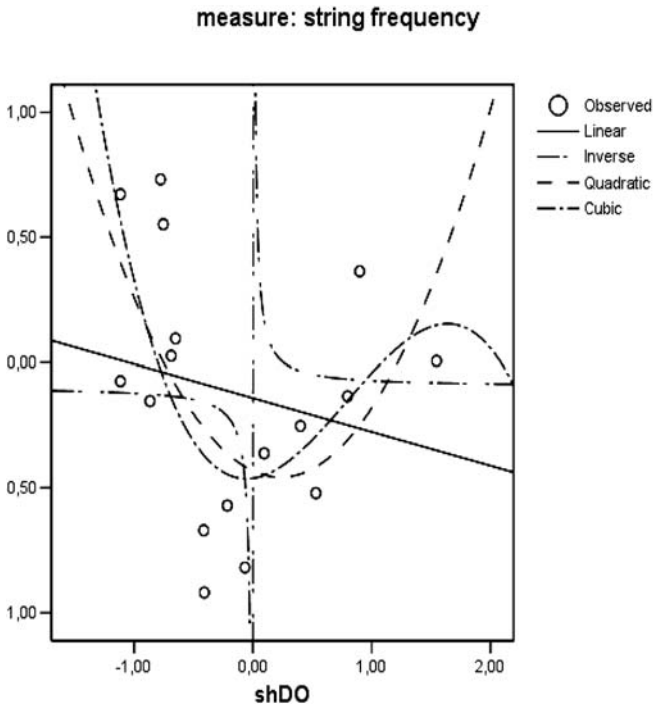

Figure 6. *Comparison of regression models*

Figure 6 shows different types of linear and non-linear regression models that were used (inverse, quadratic, and cubic) fit to the same data. Model evaluations via ANOVA revealed that the quadratic model is best suited to describe the relationship and to predict unobserved cases (linear model: $R^2 = .043$, adjusted R^2 square = −.02, $p = .42$; quadratic model: $R^2 = .34$; adjusted $R^2 = .25$, $p = .04$). Furthermore, none of the other two model types performed better than the quadratic model. Although it was not always the case that the quadratic model performed significantly better than its competitors, it always performed best and so the coefficients of determination from the quadratic models were universally used to evaluate the performance of the association measures.[19]

3.3. Results and discussion

Figure 7 provides a graphical overview of the results, i.e. the coefficients of determination (adjusted R^2) from the respective quadratic models that approximate the strength of the relationships between the output of association

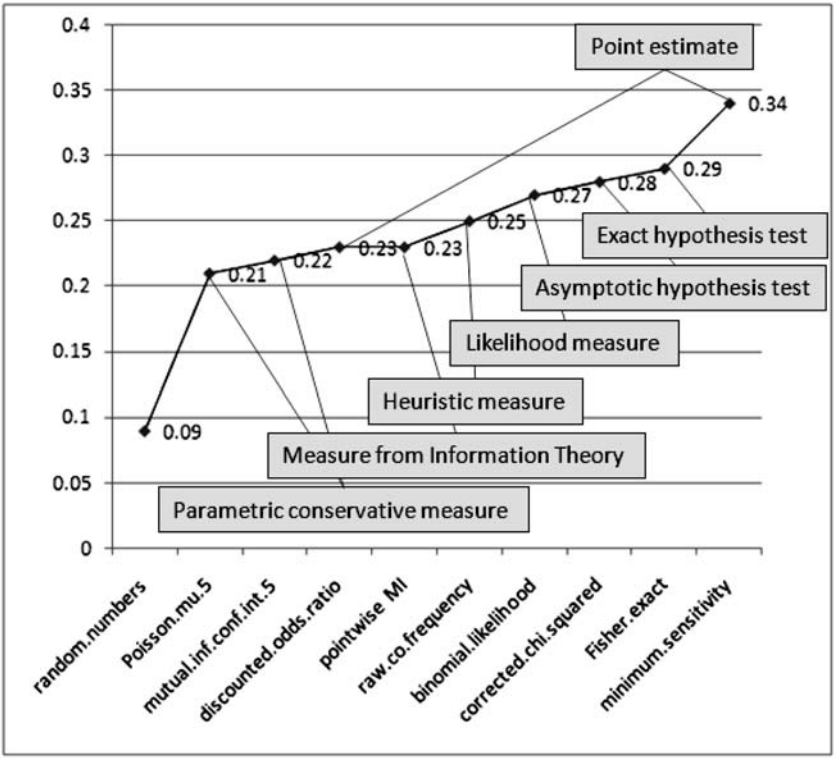

Figure 7. *Coefficients of determination (R^2) across AM (selection)*

measures and the magnitude of fixation time deltas at the disambiguating region in the crucial condition (nominal complementation with short ambiguous region). For expository reasons, Figure 8 presents only a subset of measures, where inclusion into this set was dependent on three criteria:

I include measures from all theoretical groups (cf. section 2.2)
II include measures from each cluster detected by the HAC analysis (cf. section 2.4)
III ceteris paribus, established members of a given theoretical group or cluster were preferred over relatively unknown ones

The interpretation of Figure 7 is straightforward: the higher the coefficient of determination, the better the performance of the respective measure. We can see that the choice of measure influences the strength of the relationship between collostruction strength and reading time behavior as estimated by

the regression model. The best score (adjusted R^2 = .34) was reached by a point estimate (maximum likelihood estimate), namely 'minimum sensitivity' (MS). The exact hypothesis test, 'Fisher's exact test', occupies second place (adjusted R^2 = .29) followed by its approximation, Pearson's chi squared measure (adjusted R^2 = .28) and the binomial likelihood measure (adjusted R^2 = .27). Surprisingly, raw co-occurrence frequency outperforms the measure from Information Theory, 'pointwise mutual information', and also the parametric measures 'mutual.inf.conf.int.5', i.e. the lower endpoint of the two-sided confidence interval (at significance level alpha = 1E-5).

At this point a few caveats are in order. It should be noted that the absolute magnitude of the coefficients is only of secondary interest here for three reasons. Firstly, the models are far from optimal as the sheer amount of models (n = 48) prevents any serious attempt to find the most adequate characterization of the relationships in question (cf. footnote 17). Secondly, we must recognize that both fixation time data and estimations of association strength constitute very sensitive types of data and so relatively low coefficients are to be expected at this point as it is always very difficult to predict the outcome of on-line experiments on the basis of off-line data.[20] Reading rates for adults generally lie between 150 and 400 words per minute, thus a ten word sentence is usually read in about 1 to 3 seconds (Carpenter and Just 1977), and the duration of a fixation may be as little as 160ms (cf. Tinker 1965). And thirdly (and maybe most importantly), the scope of this study is – of course – rather limited and so the results need to be interpreted with caution.

A natural last step in the analysis is testing the differences between the coefficients for statistical significance. This was done in analogy to a method suggested in Wuensch et al. (2002): all coefficients were first transformed to approximate a standard distribution by means of Fisher's Z-transformation (cf. Fisher 1918) and ordered by magnitude.[21] It is then possible to compute a test statistic that outputs a z-value, which can be used to obtain the corresponding p-value[22]. The pairwise comparisons do however not even approach conventional levels of significance. Even a direct comparison of the top-ranked and the lowest-ranked measures ($R^2_{\text{minimum sensitivity}}$ = .37 and $R^2_{\text{Poisson.mu10}}$ = .14) yields a p-value of .439 and is hence judged by the test to be far from significant. At this point, the limitations of the present study become obvious. Since the proposed test for the comparison of coefficients of determination clearly judged the different measures to be not statistically significantly different from one another, we cannot deduce that any measure is suited any better than any other measure. Simply put, the result that the difference between AMs is not statistically significant does rather suggest that nothing depends on the choice of measure. However, I believe there are in fact good reasons why we should not infer from the outcome of the null hypothesis test that we can turn to an arbitrary measure of association: first,

the statistics that compares the measures' performance is rather insensitive and would require the sample sizes to be roughly six times larger than they actually are in order to judge the same difference between .37 and .14 as significant[23]. And second, we need to bear in mind that the present study rests on an analysis of roughly 6,500 manually coded data points that distribute across 21 verbs. The constraining number is thus not the number of verb tokens but rather the number of types investigated. Hence, the required sample size is hard to come by not only because manual inspection of (6 × 6,500 =) 39,000 data points is quite costly, but also because there is – to the best of my knowledge – no on-line experimental study that has tested as many as 120 verbs. In other words, the reason why the differences between the coefficients of determination did not turn out to be statistically significant is not obvious. It is possible that the test is simply suited to serve the task required here. However, given the fact that the coefficients exhibit a (crude) range of > .2, it appears likely that researcher A might doubt the relevance of collostruction strength for processes of on-line sentence comprehension had he used the Poisson.mu.10 measure (R^2 = .14), whereas researcher B, who conducted the exact same study but with minimum sensitivity as a measure of collostruction strength, might infer from an R^2 = .37 that verb specific preferences do in fact play a role in parsing.

Regardless of how significant the differences should be considered to be, we are still left with a ranked order and I believe that at least two findings should be commented on: firstly, the fact that raw co-occurrence frequency performed so well, and secondly, the finding that a point estimate (minimum sensitivity) outperformed the statistically sound FET. Of these two issues the former appears more striking and unexpected than the latter. After all, if co-occurrence frequency outperforms measures from information theory and conservative parametric statistics and if the differences between all these scores cannot be shown to be statistically significant, then Occam's razor tells us to use this simpler heuristic. However, we have reason to believe that such an argument can be avoided here by paying closer attention to the data. Again, let me quickly point out what I believe is the main reason for the surprisingly good performance of the raw frequency measure. As described in section 3.2, the regression models that were fitted to characterize the relationship between the collostruction strength estimations and the reading time behavior do not take into account extreme cases: data points that diverge drastically from the mean have been removed (cf. footnote 16). The number of outliers varies across measures between 0 and 4, and the data points that were removed were not always identical. In case of the raw frequency measure, the types deleted were the verbs *notice*, *feel* and *say*, which as illustrated in Figure 8, clearly show the greatest difference to the FET measure and must be considered rather crude overestimations by the raw frequency measure. Removing these three types has the greatest impact on the curvature of the vector and thus

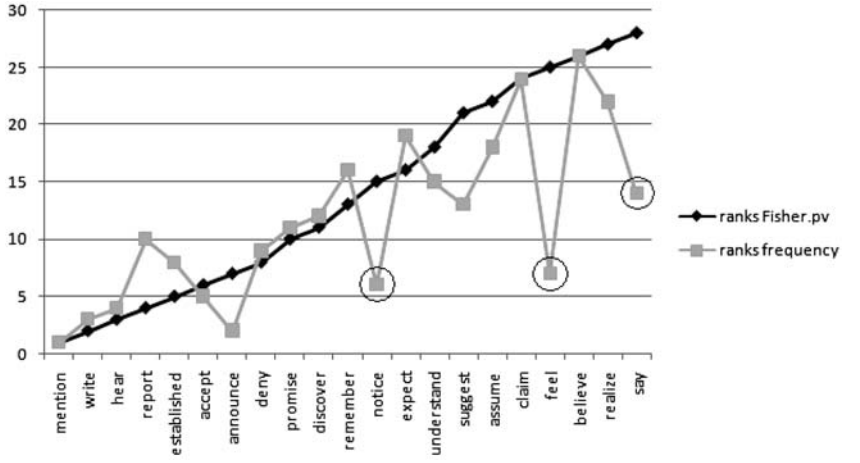

Figure 8. *Ranked output raw frequency versus FET*

can be held responsible for the surprisingly good fit of the corresponding regression model.

More generally, it may be the case that the SUBCAT preferences relevant for parsing are well approximated by string frequency in cases where the set of possible patterns is relatively small and the token frequency of a particular verb is relatively high. In these cases, raw co-occurrence, O_{11}, is particularly informative. Again, given the soundness of the argumentation that brought us from string frequency to more sophisticated measures in the first place, I believe we should not throw out the baby with the bath water. Instead we should try and determine exactly under what circumstances string frequency is more informative than theoretically more pleasing measures (or maybe just informative enough). Future research may pursue these issues.

Even if we can at least partially explain the surprisingly good performance of the string frequency measure, we still need to face the fact that minimum sensitivity outperformed FET. Since MS is the "winner of this competition", I will present a little more information about its theoretical composition (for further details cf. Pedersen and Bruce 1996; Pedersen 1998). Sensitivity, or recall rate, is classically defined as the proportion of true results that agree with the true state:

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false positives}}$$

For co-occurrence phenomena involving pairs w_1, w_2, we compute the sensitivity of both members of our bigram (w_1, w_2) as follows:

$$S_{w1} = \frac{O_{11}}{C_1} = P(w1|w2) \qquad \text{and} \qquad S_{w2} = \frac{O_{11}}{R_1} = P(w2|w1).$$

Minimum sensitivity thus relates two conditional probabilities (*P*(verb|construction) and P(construction|verb)). Both values vary between 0 and 1 and their minimum can be viewed as a measure of association between the members of the pair, such that the greater the minimum, the greater the degree of association. MS is preferable over its competitors for the following reasons:

1. It is free from underlying distributional assumptions that are not met by natural language data.
2. It is computationally less demanding than the FET suggested by Stefanowitsch and Gries (2003) in their original proposal of collostructional analysis.
3. As a point estimate (maximum likelihood estimate) of association strength, its output is less dependent on sample sizes than (exact or asymptotic) statistical hypothesis tests and hence does not necessitate the change of statistics when the task at hand involves comparisons of estimates from samples that differ in size.
4. It is empirically most adequate, as it not only scored best in the present study but it also outperformed its competitors (Pearson's chi-squared measure, t-score, and log-likelihood) in a study judging the degree of collocativity of adjective noun pairs presented in Krenn (2000).

The good performance of MS revives the methodological issue that has been addressed in Gries et al. (2005): in their analysis of the association between verbs and the *as*-predicative construction, Gries and colleagues compare (the predictions of) conditional probabilities to collostruction strength as expressed by FET. On the theoretical side, they point out that both conditional probabilities (P(verb|construction) and P(construction|verb)) do not take into account all the potentially relevant information, i.e. all the information entailed in the frequency signature, and are thus informationally poorer that the exact statistical hypothesis test. In terms of empirical performance – in a task to predict human behavior in a sentence completion task – , Gries and colleagues showed that FET outperforms both conditional probabilities and suggest that this can be interpreted as resulting from the lower informational richness of these measures. Note, however, that MS relates the two conditional probabilities that have been looked at in isolation in Gries et al. (2005), and thereby raises the level of informational richness (maybe to a level that closely approximates what the human processing system is capable of taking into account). Given the limitations of the present study, I shall refrain from deeper speculation here and instead invite future research to further examine this question. However, I would like to emphasize the more general point

that Gries and colleagues put so aptly: "If cognitive linguists aim to be true to their tenet of providing a realistic usage-based picture, then their methods must be geared to their purposes" (Gries et al. 2005: 665).

4. Concluding remarks

So what can be learned from this investigation? I hope to have shown that there are three main lessons that can be learned from this study:

First, we have seen that different measures of association arrive at different estimations of collostruction strength between a given verb and two of their complementation patterns (Section 1). Section 2 has provided a principled methodology of how the outputs of the tested measures relate to each other. The results of the agglomerative hierarchical clustering technique give us some confidence in asserting that – should the task require it – we can go from the theoretically sound Fisher exact test, i.e. from the group of exact statistical hypothesis tests, to a maximum likelihood estimate, say odds ratios, without risking too much quantitative difference in the estimation of association strength. Section 3 has shown that the choice of AM has measurable consequences (and so our choice is not completely unconstrained): the coefficients of determination (adjusted R^2) from the regression models have exhibit noticeable differences ($\Delta R^2 > .2$) even though these differences were not judged to be statistically significant by the hypothesis test employed here.

In conclusion, I have tried to make the point that there is still a strong need for empirical evaluations of competing measures of collocativity (or collostruction strength for that matter): not only can the choice of measure influence decisions relevant for linguistic theory building, e.g. inferences regarding the semantics of an argument structure construction on the basis of its collocates, but it also heavily influences our understanding of the workings of cognitive processes: contemporary models of sentence processing routinely make use of individual SUBCAT preferences and hence we should try and estimate them as precisely as possible.

Bionote

The author received his Magister Artium in English linguistics (with minors in psychology and philosophy) from the University of Hamburg, Germany, in 2005. At current, he is at the University of Jena, Germany, working towards his doctoral dissertation, which presents a quantitative corpus linguistic approach to the processing of English relative clause constructions. His research interests include questions concerning the relationship between competing linguistic theories, the relationship of linguistic theories and psychological theories of language processing, empirical methods in (corpus) linguistics, and philosophical issues in linguistic semantics and pragmatics.

Notes

* I would like to express my thanks to Shelia Kennison for sharing her original fixation time data with me. I thank Stefanie Wulff for letting me use her ICE-isomorphic BNC compilation. I would also like to thank Stefan Evert for his UCS TOOLKIT, Stefan Th. Gries for CLUSTER EVAL 0.9 and all R contributors for their efforts and open source spirit. Finally, I thank Stefan Th. Gries for many valuable discussions as well as Stefanie Wulff and two anonymous reviewers for their helpful comments. All remaining errors are of course my own.

1. According to Leech et. al (2001), *man* occurs 1003 times per million words in its nominal usage in the BNC, whereas its frequency as a verb is lower than 10 per million.
2. For the purposes of this paper, the author remains neutral with respect to the exact mechanism responsible for observable processing difficulties. The description should be viewed on an as-if level and should not be taken as a description of the actual processes involved in early stages of parsing. Wiechmann (2008) argues – in accordance with MacDonald (1993, 1994) and others – that the processing latencies in psycholinguistic experiments are best explained by a constraint satisfaction account of parsing.
3. Within the Competition Model, cue validity is not an atomic notion but is in turn derived from a number of constitutive notions, namely "cue availability", "cue reliability", and "conflict validity" (cf. Bates and MacWhinney 1989: 41ff for a discussion).
4. This is why Gries (2006) turned to odds ratios to compare statistical regularities across different corpora. Similarly, Wiechmann (2008) made use of discounted odds ratios to assess the degrees of attraction between a given verb and a particular complementation pattern.
5. The ICE-isomorphic BNC compilation was generously provided by Stefanie Wulff.
6. At this point, I would like to stress the value of manual inspection of the corpus data. Although a fully automated identification of the type of complementation surely would have been possible, this would have led to a decline of precision. CLAWS, a hybrid automatic tagging system used to annotate the items in the BNC with part of speech (POS) information, reaches high degrees of accuracy (96% to 97% depending on text type) but still has its occasional problems: I discovered numerous errors particularly in connection with the word *that*, which exhibits a high degree of ambiguity with respect to the POS it can instantiate.
7. For the present study, α was pre-set to the value specified in the name of the measure. Values are 2, 3, 5, 10, 50, 100, 1000. So for example, MI.conf.10 computes the two-sided confidence interval at significance level 1E-10.
8. The UCS toolkit is available free of charge from http://www.collocations.de. All calculations were performed on the following platform: Linux kernel 2.4.3, Perl 5.6.1, R 2.5.0.
9. Z-standardization, $Z = (X - M)/SD$, involves two steps: first, the variable is centered so that the mean (M) becomes zero; second, a division by the standard deviation (SD) makes the unit a SD. Z-standardization is a linear transformation and so all relative distances remain intact.
10. Other clustering techniques partition the data into a number of cluster specified by the user (= partitioning) or are divisive, i.e. they start with a single cluster and then split up the aggregate until all object are in different groups. Everitt (1993) provides an approachable introduction presenting many different clustering techniques.
11. A similarity coefficient s_{ij} expresses the strength of the relationship between two elements given a set of p variates common to both. Usually, similarity is treated as a symmetrical notion such that $s_{ij} = s_{ji}$. For an extensive list of similarity coefficients, the reader may consult Gower (1985). Dissimilarity (or distance) measures can be seen as a complement of similarity measures. They usually have the metric property that for all objects i, j, k it holds that $d_{ij} + d_{ik} \geq d_{jk}$. Everitt (1993) discusses some of the most widely used measures that can be used to express d_{ij}.

12. In addition to the analysis of the z-standardized output reported here, all HAC-calculations have also been done on the basis of the ranked output of the association measures. As expected, abstracting away from fine-grained quantitative differences resulted in more coarse-grained groupings with optimal solutions consisting of fewer groups. However, for reasons of exposition, the results were not reported here as they do not add to the overall interpretation of the findings.
13. For all objects i in the data, the silhouette of i $s(i)$ is computed as follows: $s(i)\frac{b(1)-a(i)}{\max\{a(i),b(i)\}}$ [where $a(i)$ = within cluster dissimilarity = average dissimilarity of i to all objects a given cluster A and $b(i)$ = between cluster dissimilarity = min $d(I, C)$ with $C \neq A$; $d(I, C)$ = average dissimilarity of i to all objects of C].
14. 'Cluster.eval 0.9.' was generously provided by Stefan Th. Gries.
15. An ASW value is a local high, if it is greater than its (immediate) neighbors.
16. The threshold level of 100 maybe considered higher than necessary. In order to obtain a larger number of types to rest the analysis on, which would be relevant for later steps of the analysis, the frequency threshold level was lowered to 30. It turned out that while this move caused the estimation of collostruction strength to become less adequate, it did not counter the effect associated with the low type frequency in the evaluation of measures.
17. The region was identified as most interesting on the basis of correlational analyses$_{\text{Kendall}}$ between FET values of a given verb and the associated fixation times of a) different potentially relevant positions (Region 6 $\pm$ 1 position), b) short versus long ambiguous regions, and c) first pass versus total reading times. For reasons of exposition, the respective results will not be reported here in any detail. However, the results identify the relevant region rather clearly with p-values corresponding to the correlation coefficients ranging from .07 for the regions chosen here to > .5 for all other regions.
18. In order to count as an 'extreme outlier', a data point must be at least 1.5 * IQR (= interquartile range) lower than the 1st quartile or 1.5 * IQR higher than the 3rd quartile.
19. I agree with an anonymous reviewer that it is rather surprising to see that the quadratic models achieved the best fit (as many cognitively relevant quantities have been shown to be on a log scale). However, I would like to emphasise that the coefficients of determination are intended here to provide a basis of comparison of the measures under investigation. The quadratic models surely provide better approximations of the actual relationship between the quantities at hand than the linear models, but they are not meant to represent the most adequate type of model. I should be perfectly explicit about the fact that it was not the goal of the regression analyses to find the most adequate model. Doing so for 48 association measures would exceed the scope of the study for the simple reason that statistical modelling requires (tedious) model checking. I am convinced that additional modelling and model-testing of more and larger data sets are needed in order to arrive at a more adequate characterization of the relationship between fixation times and collostruction strength.
20. Off-line data result from techniques that are product-oriented, often involving conscious thinking and reflection. Among these methods are questionnaire studies, sorting and association tasks. In contrast, on-line methods try to tap into processes and usually use reaction times as a dependent variable.
21. Each coefficient was transformed in this fashion: $R^2 = \frac{\log_g(1 + R^2)}{1 - R^2}$.
22. The value of z can be computed as follows: $z = \frac{(R_1^2 - R_2^2)}{\sqrt{\frac{1}{n_1-3} \times \frac{1}{n_2-3}}}$.
23. With n = 120, we get p = 0.0487.

References

Agresti, Alan
1999 *Categorical Data Analysis*. New York: John Wiley and Sons.

Adams, Brian C., Charles Clifton Jr. and Don C. Mitchell
1998 Lexical guidance in sentence processing? *Psychonomic Bulletin and Review 5*, 265–270.

Altmann, Gerry T.M.
1998. Ambiguity in Sentence Processing. *Trends in Cognitive Sciences* 4, 146–152.

Barlow, Horace
2001 The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences* 24, 602–607.

Bates, Elizabeth and Jeffrey Elman
1996 Learning rediscovered. *Science* 274, 1849–1850.

Berry-Rogghe, Godelieve L.M.
1973 The computation of collocations and their relevance to lexical studies. In Aitken, A.J., Bailey, R.W., and Hamilton-Smith, N. (eds.), *The Computer and Literary Studies*, Edinburgh, 103–112.

Biber, Douglas
1993 Co-occurrrence patterns among collocations: A tool for corpus-based lexical knowledge acquisition. *Computational Linguistics* 19, 549–556.

Blaheta, Don and Johnson, Mark.
2001 Unsupervised learning of multi-word verbs. In Proceedings of the ACL Workshop on Collocates, Toulouse France, 54–60.

Bod, Rens, Jennifer Hay and Stefanie Jennedy (eds.)
2003 *Probabilistic Linguistics*. Cambridge, MA: MIT Press.

Brants, Thorsten and Matthew Crocker
2000 Probabilistic parsing and psychological plausibility. In *Proceedings of the 18th conference on Computational linguistics* 1, 111–117.

Brunswik, Egon
1956 *Perception and the Representative Design of Psychological Experiments*. Berkeley: University of California Press.

Carpenter, Patricia A. and Marcel A. Just
1977 Reading comprehension as eyes see it. In Carpenter, P.A. and Marcel A Just (eds.) *Cognitive Processes in Comprehension*. Mahwah, NJ: Lawrence Erlbaum.

Chomsky, Noam.
1995 *The Minimalist Program*. Cambridge, MA: MIT Press.

Church, Kenneth and Patrick Hanks
1990 Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, 22–29.

Christiansen, Morten and Nick Chater
1999 Connectionist natural language processing: The state of the art. *Cognitive Science* 23, 417–437.

Church, Kenneth, William A. Gale, Patrick Hanks and Donald Hindle
1991 *Using statistics in lexical analysis, Lexical Acquisition: Using On-line Resources to Build a Lexicon*, Mahwah, NJ: Lawrence Erlbaum, 115–164.

Crocker, Matthew, Martin Pickering and Charles Clifton Jr.
2006 *Architectures and Mechanisms of Language Processing.* Cambridge: Cambridge University Press.

Diessel, Holger
2007 Frequency effects in language acquisition, language use, and diachronic change. *New Ideas in Psychology* 25: 108–127.

Elder, James H. and Rick M. Goldberg
1998 The statistics of natural image contours. In *IEEE Workshop on Perceptual Organisation in Computer Vision 1998.*
Everitt, Brian S.
1993 *Cluster Analysis*. New York: John Wiley and Sons.
Evert, Stefan
2004 The Statistics of Word Cooccurrences: Word Pairs and Collocations. Unpublished doctoral dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.
Firth, John .R.
1957 A synopsis of linguistic theory 1930–55, *Studies in linguistic analysis*, The Philological Society, 1–32.
Fisher, Ronald A.
1918 The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* 52, 399–433.
Geng, Jin J. and Marlene Behrmann
2006 Spatial probability as an attentional bias in visual search. *Perception and Psychophysics* 67(7), 1252–1268.
Goldberg, Adele
1999 The emergence of the semantics of argument structure constructions. In MacWhinney, Brian (ed.), *The Emergence of Language*. Mahwah, NJ: Lawrence Erlbaum, 197–212.
2006 *Constructions at work: The Nature of Generalisation in Language*. Oxford: Oxford University Press.
Goldberg, Adele, Devin Casenhier and Nitya Sethuraman
2004 Learning argument structure generalizations. *Cognitive Linguistics* 14, 289–316.
Gower, John C.
1985 Measures of similarity, dissimilarity, and distance. In: Kotz, Samuel and Norman L. Johnson (eds.). *Encyclopedia of Statistical Sciences*, Vol. 5. New York: Wiley, 397–405.
Gries, Stefan Th.
2005 Syntactic priming: A corpus-based approach. *Journal of Psycholinguistic Research,* 34, 365–399.
2006 Exploring the variability within and between corpora: some methodological considerations. *Corpora* 1, 109–51.
Gries, Stefan Th. and Anatol Stefanowitsch
2004 Extending collostructional analysis: A corpus-based perspectives on 'alternations'. *International Journal of Corpus Linguistics* 9, 97–129.
Gries, Stefan Th, Beate Hampe and Doris Schönefeld
2006 Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16, 635–76.
Hare, Mary L., Ken McRae and Jeffrey L. Elman
2003 Sense and structure: Meaning as a determinant of verb subcategorization preferences. *Journal of Memory and Language* 48, 281–303.
2004 Admitting that admitting verb sense into corpus analyses makes sense. *Language and Cognitive Processes* 19, 181–224.
Hebb, David
1949 *The organization of behaviour: A neuropsychological theory*. New York: Wiley.

Helmholtz, Hermann von.
1925 *Physiological Optics. Volume III. The Theory of the Perceptions of vision (Translated from 3rd German Edition, 1910)*. Washington: Optical Society of America.
Hoey, Micheal
1998 Introducing Applied Linguistics: 25 Years on. *The 31st BAAL Annual Meeting: Language and Literacies*. University of Manchester.
Jurafsky, Daniel
1996 A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science* 20, 137–194.
Kennison, Shelia
2001 Limitations on the use of verb information during sentence comprehension. *Psychonomic Bullettin and Review* 8, 132–138.
Krenn, Brigitte
2000 The Usual Suspect: Data-oriented models for the identification and representation of lexical collocation., volume 7 of Saarbrücken Dissertations in Computational Linguistics and Language Technology. DFKI and Universität des Saarlandes, Saarbrücken, Germany.
Krug, Manfred
1998 String Frequency: A Cognitive Motivating Factor in Coalescence, Language Processing and Linguistic Change. *Journal of English Linguistics* 26, 286–320.
Lapata, Mirella, Frank Keller and Sabine Schulte im Walde
2001 Verb frame frequency as a predictor of verb bias. *Journal of Psycholinguistic Research* 30, 419–435.
Langacker, Ronald
1987 *Foundations of Cognitive Grammar: Volume I: Theoretical Prerequisites*. Stanford, CA: Stanford University Press.
Leech, Geoffrey, Paul Rayson and Andrew Wilson
2001 *Word frequencies in written and spoken English based on the British National Corpus*. London: Longman.
MacDonald, Maryellen C., Neil J.Pearlmutter and Mark S. Seidenberg
1994 The lexical nature of syntactic ambiguity resolution. *Psychological Review* 101, 676–703.
Mach, Ernst
1886 *The Analysis of Sensations, and the Relation of the Physical to the Psychical (Translation of the 1st, revised from the 5th, German Edition by S. Waterlow)*. Chicago and London: Open Court.
MacWhinney, Brian
1987 Toward a psycholinguistically plausible parser. In S. Thomason (Ed.) ESCOL 1986. Columbus, OH: Ohio State University.
MacWhinney, Brian and Elisabeth Bates (eds.)
1989 *The crosslinguistic study of sentence processing*. New York: Cambridge University Press.
Manning, Christopher and Hinrich Schütze
1999 *Foundations of Statistical Natural Language Processing*. Cambridge, MA:MIT Press.
Mitchell, Donald C.
1987 Lexical guidance in human parsing: Locus and processing characteristics. In Coltheart, Max (ed.), *Attention and Performance XII*. Hillsdale, NJ: Lawrence Erlbaum.

Narayanan, Srini and Daniel Jurafsky
2001 A Bayesian Model Predicts Parse Preferences and Reading Times in Sentence Comprehension, *Neural Information Processing Systems*.
Nelson, Gerald
1996 The Design of the Corpus'. In: Greenbaum, S. (ed.) *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press, 27–35.
Pearson, Karl
1892 *The Grammar of Science*. London: Walter Scott.
Pedersen, Ted
1996 Fishing for exactness. In: *Proceedings of the South Central SAS User Group Conference*, Austin, TX.
1998 Dependent Bigram Identification. *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, July 1998, Madison WI.
Pedersen, Ted and Robert Bruce
1996 What to infer from a description. Technical Report 96-CSE-04, Southern Methodist University, Dallas, TX.
Pickering, Martin
1999 Sentence comprehension. In: Garrod, S. and Pickering, M. (eds.). *Language Processing*. Sussex, UK: Psychology Press, 123–153.
Rao, Rajesh P.N., Ohlshausen, Bruno A., and Lewicki, Michael S.
2002 *Probabilistic Models of the Brain: Perception and Neural Function*. Cambridge, MA: MIT Press.
Redington, Martin, Nick Chater and Steve Finch
1993 Distributional information and the acquisition of linguistic categories: A statistical approach. In *Proceedings of the 15th annual meeting of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum, 848–853.
Seidenberg, Mark S.
1997 Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275, 1599–1603.
Sinha, Chris
2007 Cognitive Linguistics, Psychology and Cognitive Science. In Geeraerts, Dirk and Hubert Cuyckens (eds.) *Handbook of Cognitive Linguistics*, Oxford, Oxford University Press, 1266–1294.
Spivey, Michael J. and Tanenhaus M.K.
1998 Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning, Memory and Cognition* 24, 1521–1543.
Stefanowitsch, Anatol and Stefan Th. Gries
2003 Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8, 209–243.
Tinker, Miles A.
1965 *Bases for effective reading*. Minneapolis: University of Minneapolis Press.
Wiechmann, Daniel
2008 Sense-contingent lexical preferences and early parsing decisions: Corpus-evidence from local NP/S-ambiguities. *Cognitive Linguistics* 19, 439–455.
Wuensch, Karl L., Kevin W. Jenkins and G. Michael Poteat
2002 Misanthropy, idealism, and attitudes towards animals. *Anthrozoös* 15, 139–149.

Corpora

The British National Corpus (Version 1.0). 1995. Oxford University Computing Services for the BNC Consortium. Oxford: Oxford University.

Software

*R*2.5.0. R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.
(http://www.r-project.org/)

UCS toolkit V 0.5. Stefan Evert 2004–2006.
(http://www.collocations.de/software.html#UCS/)

Cluster.eval 0.9 – A script for R for Windows. Stefan Th. Gries 2007.
Available upon request from the author
(http://www.linguistics.ucsb.edu/faculty/gries)