Ulrike Schneider*

ΔP as a measure of collocation strength

Considerations based on analyses of hesitation placement in spontaneous speech

https://doi.org/10.1515/cllt-2017-0036

Abstract: This paper explores the proposed benefits of Δ*P* (delta *P*) as a measure of collocation strength. Its focus is on contrasting Δ*P* with other, more commonly used, association measures, particularly transitional probabilities, but also mutual information and Lexical Gravity *G*. To this end, first the strong correlation between Δ*P* and transitional probability is illustrated with the help of two exemplary corpora. This is followed by an analysis of hesitation placement in spontaneous spoken English, based on the assumption that hesitations will not be placed within strong collocations. Results show that, despite their strong similarity, in some contexts Δ*P* is more predictive of hesitation placement than transitional probability. Yet neither Δ*P* nor any of the other association measures emerges as the universally best predictor. On the basis of these results, it is suggested that studies should always rely on several association measures.

Keywords: collocations, association measures, Δ*P* (delta *P*), bigrams, hesitations

1 Introduction

This paper is dedicated to Δ*P* (delta *P*), a unidirectional dependency measure. It first appears in the literature as a means to model the outcome of judgment tasks (cf. e.g. Jenkins and Ward 1965; Ward and Jenkins 1965; Allan 1980), but, as we will see below, recent studies suggest that it may also be used for modelling language learning and for collocation strength testing. In this paper, I will focus on the latter application and evaluate the benefits of Δ*P* as a measure of collocation strength. Of course, there are already a number of other measures of collostruction strength which have been systematically compared elsewhere (cf. e.g. Evert 2004; Kapatsinski 2005; Jurafsky and Martin 2008; Wiechmann

***Corresponding author: Ulrike Schneider,** Department of English and Linguistics, University of Mainz, Philosophicum, Jakob-Welder-Weg 18, 55128 Mainz, Germany,
E-mail: ulrike.schneider@uni-mainz.de

2008; Pecina 2010), yet ΔP promises great potential at relatively low computational cost.

In essence, ΔP is a measure of cue validity, i.e. it measures how strongly two events are linked. Thus it evaluates how reliably a specific $event_1$ triggers another specific $event_2$, simultaneously taking into account how likely $event_2$ is to occur after any other $event_n$. In terms of the most famous contingency learning experiment, namely Pavlov's dogs, ΔP aims to help us understand the conditions necessary to train the dog to salivate at the sound of a bell: according to ΔP, the more likely a reward of food is to follow the sound of a bell and the less likely food is provided under any other circumstances, the more strongly the bell and food will be cognitively linked for the dog and thus the more likely the experiment will be successful (see also explanation of a "good cue" in Ellis and Ferreira-Junior 2009: 197). Humans may develop similar cognitive links between words which tend to occur together – though rather than drool, humans tend to think of another word upon hearing a verbal stimulus.

Ellis (2006) advocates that ΔP be used in models of L1 and L2 acquisition, based on evidence that ΔP theory accurately predicts human estimation experiments (Ellis 2006: 11–12; cf. also Shanks 1995) and that it closely models the associative learning processes in humans and animals. Ellis and Ferreira-Junior (2009: 202) show that this also holds true for language learning by providing evidence that ΔP almost perfectly predicts which verbs learners will use first in a newly acquired construction.

Gries (2013) goes one step further and proposes that ΔP be used as a measure of collostruction strength. He sees ΔP's great potential for corpus linguistics in its unidirectionality, its ease of calculation (for both see the discussion of its calculation below) and the existing evidence that it reflects psychological and psycholinguistic reality (Gries 2013: 143–144). This proposal to lift ΔP from its accustomed domain of experimental studies and to apply it to corpus linguistic research entails transferring ΔP from the analysis of processes to analyses of states, i.e. while studies on contingency learning tend to focus on the stages or sequences in the learning process, studies of collocations are generally totally synchronic in that they determine collocation strengths at a single point in time. As corpora are usually much larger than experimental data, this transfer furthermore entails a huge increase in the size of the data sets ΔP is applied to. This paper assesses the performance of ΔP in corpus data by comparing it to the performance of other measures of association.

The paper is structured as follows: Section 2 first discusses the link between collocation strength and linguistic models of the mind. It briefly outlines the kinds of processing effects that have been modelled with other measures of collocation strength before investigating whether ΔP has the potential to

compete with these measures. To this purpose, it explains the formula for the calculation of ΔP and – based on its application to two exemplary corpora – demonstrates how the measure is influenced by the size of the database. Results illustrate that ΔP differs little from transitional probability when calculated for a large database. Thus the question arises whether we can do away with ΔP in cognitive corpus linguistics and stick to tried and trusted transitional probability. The section concludes by showing that previous studies fail to answer this question. Section 3 presents the data and method of a set of analyses which aim to fill this gap by comparing the performance of ΔP in predicting hesitation placement in spoken American English to the performance of a set of other measures of collocation strength, including transitional probability. These analyses will be presented in Section 4 and discussed in Section 5.

2 Corpus, collocation and cognition

2.1 Collocations as cognitive structures

According to the simplest definitions, a collocation is not much more than a group of words which occur together frequently or at least more frequently than expected by chance (see, for example, the definition of a "lexical bundle" in Biber et al. 1999: 989; for a more detailed discussion see Gries 2013: 138–139). In this respect, ΔP is just the latest method to assess such co-occurrence rates. However, collocations are rarely extracted purely for the purpose of statistical assessment.

Usage-based and connectionist models of the mind assume that linguistic items are connected on various levels and that these connections are created and strengthened through usage so that a network emerges (cf. e.g. Eikmeyer et al. 1999; Langacker 2000; Croft 2001; Fillmore et al. 2003; Fried and Östman 2004; Beckner et al. 2009; Bod 2010; Bybee 1998). Depending on the model, strong collocations can either be represented by strong connections between the nodes representing the collocates (cf. e.g. Phillips 1983; Rumelhart and McClelland 1986; Elman 1990; Brezina et al. 2015) or (additionally) by a node which represents the entire collocation as a "prepackaged unit" (Bybee and McClelland 2005: 384; Biber et al. 1999; Wray 2002), which is then often referred to as a "chunk" (e.g. Bybee 2006; Bybee 2010) or a "construction" (e.g. Goldberg 2005).[1]

1 For a discussion of the differences and (potential) mutual benefits of chunking and non-chunking models, see Perruchet and Pacton (2006).

Thus a collocation as an instance of language use is both a reflection of these mental representations as well as new input, which may alter connection strengths or representation strengths. This is evident from a number of production and comprehension effects. It has been shown that phonetic reduction increases in frequent or predictable collocations (cf. Jurafsky et al. 1998; Gregory et al. 1999; Bybee 2002; Bell et al. 2003; Bybee and Scheibman 2007; Bresnan and Spencer 2013). Strong collocations are also less likely to be disfluent (cf. Beattie and Butterworth 1979; Shriberg and Stolcke 1996; Bybee 2007b; Schneider 2014; Schneider 2016) and are read faster (cf. Frisson et al. 2005; Reali and Christiansen 2007). Furthermore, in experiments participants are faster to judge a strong collocation as acceptable than a weak one (cf. Ellis et al. 2008; Arnon and Snider 2010), but with increasing collocation strength it becomes harder for them to identify the individual words in the collocation (cf. Vogel Sosa and MacFarlane 2002; for more complex findings see also Kapatsinski and Radicke 2009).

However, so far, there is no "gold standard" to measure the strength of a collocation. Evert (2004), Wiechmann (2008) and Pecina (2010) each compare up to 80 measures of collocation strength which have been used to date. These range from simple co-occurrence frequency to complex comparisons between observed and expected occurrence rates. This means that today each new measure on the market has to compare to the performance of those already in use. Additionally, as per Occam's razor, we should avoid making any superfluous assumptions. Thus the ideal measure of collocation strength is the one which most accurately models processing effects. Should several perform on par, the one based on the fewest assumptions is preferable. Therefore, the purpose of the present paper is to gauge both the complexity and the performance of ΔP as a measure of collocation strength.

2.2 Delta P as a measure of collocation strength

Table 1 and the formulae below explain the calculation of ΔP. Due to it being a unidirectional measure, there are two formulae, one for forward-directed ΔP and one for its backward-directed counterpart.

$$\Delta P_{\text{forward}} = \frac{a}{a+b} - \frac{c}{c+d} \tag{1}$$

$$\Delta P_{\text{backward}} = \frac{a}{a+c} - \frac{b}{b+d} \tag{2}$$

As explained above, from a statistical point of view, the first part of the formula describes the probability of a specific event$_2$ given another specific event$_1$, while

Table 1: Components of measures of association (adapted from Allan 1980; Gries 2013: 140).

	+ word y	– word y	Total
+ word x	*a*	*b*	*a* + *b*
– word x	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d*

the second part describes the probability of event$_2$ in the absence of event$_1$ (cf. Ellis 2006: 11). In the formula for $\Delta P_{backward}$, the roles of the two events are reversed.

> When these [the two parts] are the same, when the outcome is just as likely when the cue is present as when it is not, there is no covariation between these two events and ΔP = 0. ΔP approaches 1.0 as the presence of the cue increases the likelihood of the outcome and approaches –1.0 as the cue decreases the chance of the outcome – a negative association. (Ellis 2006: 11)

Thus we expect values of Δ*P* between –1 and 1. In order to see whether this is the case when Δ*P* is calculated based on corpus data, we will apply the measure to all two-word pairs (i.e. bigrams) in Lewis Caroll's *Alice's Adventures in Wonderland*, first published in 1865, as well as to those in Herman Melville's *Moby Dick*, first published in 1851.[2] Figure 1 shows the results: despite the fact that Δ*P* was calculated across sentence boundaries (where we would expect weak collocations), it only takes on values between –0.07 and 1, and only 3.9% of bigram types receive scores below zero. This is no singularity owing to any particular characteristic of the two novels. It is corroborated by Gries' calculations of Δ*P* for 262 word pairs taken from the spoken part of the British National Corpus (Gries 2013: Figure 2; see also Schmid and Küchenhoff 2013: Table 9 where transitional probabilities between words and constructions, referred to as "reliance" and "attraction", are compared to the corresponding values of Δ*P*).

An interpretation of the formulae from a corpus-linguistic point of view gives a first indication why the data might be heavily skewed towards positive

2 These novels were selected as test corpora due to their comparatively small size (by current corpus standards) which cuts down computation times and makes it possible to plot all datapoints in a single plot without R (R Development Core Team 2009) running into serious difficulties. Furthermore, as Table 2 shows, *Moby Dick* is roughly ten times the size of *Alice's Adventures in Wonderland*, so the effect of corpus size can also be taken into consideration. Both novels were extracted from Baayen's (2009) packet *languageR* for R with punctuation already removed. Capitalisation was ignored for the purposes of calculation.

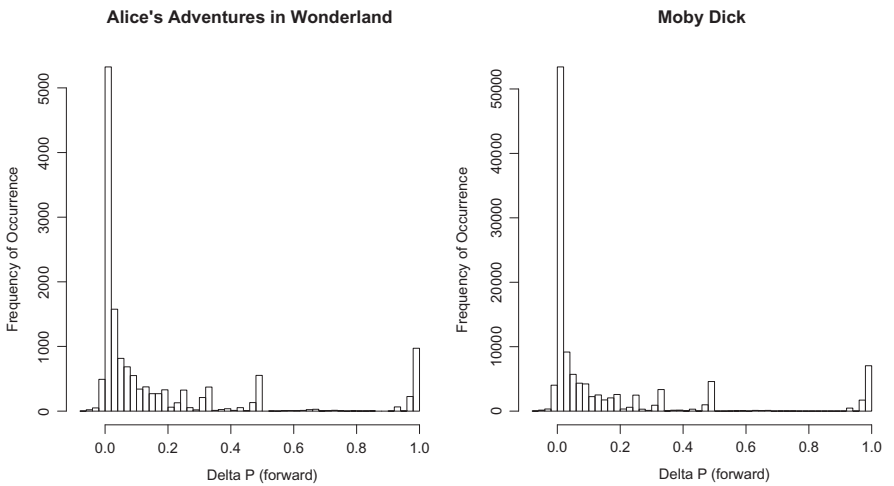

Figure 1: Distribution of values of $\Delta P_{\text{forward}}$ when calculated for all bigrams in the corpora.

Table 2: The two novels used as corpora.

Novel	Word tokens	Word types	Bigram types
Alice's Adventures in Wonderland	27,269	2,992	14,478
Moby Dick	215,994	20,531	116,455

values. From this perspective, the components of the formulae take on the following meanings:

$$\Delta P_{\text{forward}} = \frac{\text{Frequency}_{\text{Bigram}}}{\text{Frequency}_{\text{Word1}}} - \frac{\text{Frequency}_{\text{Word2}} - \text{Frequency}_{\text{Bigram}}}{\text{Corpus Size} - \text{Frequency}_{\text{Word1}}} \quad (3)$$

$$\Delta P_{\text{backward}} = \frac{\text{Frequency}_{\text{Bigram}}}{\text{Frequency}_{\text{Word2}}} - \frac{\text{Frequency}_{\text{Word1}} - \text{Frequency}_{\text{Bigram}}}{\text{Corpus Size} - \text{Frequency}_{\text{Word2}}} \quad (4)$$

It now becomes evident that the first component of each formula is actually the formula for calculating the corresponding transitional probability, i.e. in eq. [3] it is forward transitional probability and in eq. [4] it is backward transitional probability.

We could thus describe ΔP as transitional probability minus an adjustment factor (i.e. $c/(c+d)$ in eq. [1]), which increases the more often word_2 occurs without word_1 (i.e. c in Table 1), as illustrated by the first panel in Figure 2.

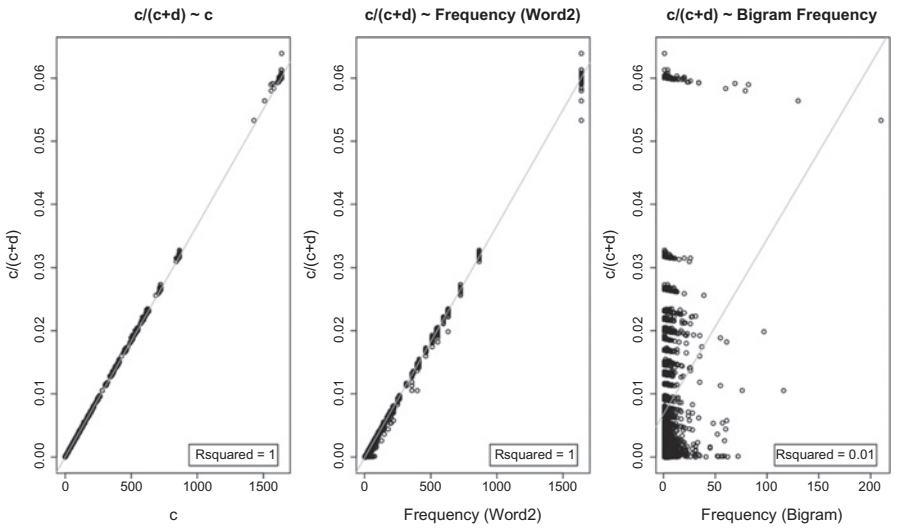

Figure 2: Correlation between the ΔP adjustment factor (i.e. $c/(c+d)$) and its components, calculated for all bigrams in *Alices's Adventures in Wonderland*. Letters in the formula refer to those used in Table 1.

The second and third panels of Figure 2 show that the adjustment is also strongly correlated with the frequency of the second word but not with the frequency of the collocation (i.e. the bigram). Yet, no matter how frequent the second word or the entire collocation is, the adjustment never reaches values of more than 0.065. This is due to the overpowering denominator: the corpus size, which is part of the denominator, will always be vastly higher than any other number in the formula and thus keep the adjustment factor small.

In summary, ΔP is transitional probability minus a small adjustment, which "punishes" pairs whose second word also frequently occurs in other combinations. As transitional probabilities can only take on values between 0 and 1 (i.e. a probability between 0% and 100%), and the adjustment factor varies between 0 and 0.065 when calculated over an entire corpus, ΔP varies between −0.065 and 1.

As a result, ΔP and transitional probability are strongly correlated. Table 3 shows that the correlation is almost perfect, reaching 0.999 (Person correlation coefficient calculated in R). The table further illustrates that such very strong correlations between measures of collocation strength are rare. These findings raise the question whether ΔP has any additional value compared to transitional probability.

Table 3: Pearson product-moment correlations between different measures of collocation strength (calculated in R).

	$TP_{for.}$	$TP_{ba.}$	*MI*	*G*	$\Delta P_{for.}$	$\Delta P_{ba.}$
Freq.	−0.021	0.028	−0.087	0.602	−0.026	0.023
	−0.022	−0.016	−0.068	0.392	−0.025	−0.019
$TP_{for.}$		−0.176	0.437	0.067	0.999	−0.166
		−0.134	0.406	0.029	0.999	−0.125
$TP_{ba.}$			0.446	0.058	−0.163	0.999
			0.404	0.033	−0.125	0.999
MI				0.150	0.459	0.465
				0.129	0.424	0.424
G					0.067	0.060
					0.030	0.033
$\Delta P_{for.}$						−0.153
						−0.116

Notes: Calculations are based on the entire corpora. Upper and lower values represent results for *Alice's Adventures in Wonderland* and *Moby Dick*, respectively. Formulae for all measures of collostruction strength are provided in Section 3. (Abbreviations: Freq.: frequency; TP: transitional probability; MI: Mutual Information score; *G*: Lexical Gravity *G*; for.: forward; ba.: backward.)

2.3 Previous applications of Δ*P* in cognitive corpus linguistics

Despite the fact that there are tools such as *GraphColl* (Brezina et al. 2015) and *Coll. analysis* (Gries 2014) which calculate both Δ*P* and more conventional measures of collocation strength for the user, surprisingly little work discussing Δ*P* has been published (but see Schmid and Küchenhoff 2013; Gries 2015a; Gries 2015b). To date, Wahl (2015) appears to be the only study which explicitly contrasts the performance of Δ*P* with that of other predictors in a corpus-based analysis.

Wahl (2015) compares six measures of collocation strength in order to determine which best predicts intonation unit boundaries. His study is based on the assumptions that (1) strong collocations are stored as holistic chunks in the mind, and (2) there is a link between these chunks and intonation units in speech, i.e. that speakers do not split one mental unit into several intonation units (Wahl 2015: 192–193). It thus follows that the more strongly two words collocate, the more likely they are mentally chunked and the less likely an intonation unit boundary should fall between them.

Wahl (2015: 198, 200, 202) calculates the *t*-score, mutual information, log likelihood as well as $\Delta P_{forward}$, $\Delta P_{backward}$ and $\Delta P_{maximum}$ for 32,000 two-word collocations which are interrupted by an intonation unit boundary and 168,000 two-word collocations which occur within the same intonation unit in the Santa Barbara Corpus of Spoken American English. Wahl (2015: 209) finds that $\Delta P_{forward}$ only correctly predicts 0.1% of intonation unit boundaries, while $\Delta P_{backward}$ predicts 8.2% correctly – a similar performance to that of the bidirectional measures used in his study. A second study confirms that there is a strong positive correlation between $\Delta P_{backward}$ and the likelihood of collocations being part of the same intonation unit (Spearman's *rho* = 0.97). In the case of $\Delta P_{forward}$, however, the correlation is weak and actually negative (*rho* = –0.24), confirming the "aberrant behaviour" of this measure (Wahl 2015: 213). Drawing on Onnis and Thiessen (2013), Wahl (2015: 209) concludes that "the leftward inter-word relationship is typically the more predictive direction in languages like English", the reason being that "large open classes of content words are more predictive of small numbers of closed-class function words that precede them than vice-versa" (Wahl 2015: 209).

Thus, the results suggest that only $\Delta P_{backward}$ has any practical value in corpus-based approaches to the processing of English. Unfortunately, Wahl does not address the other question at hand, namely whether ΔP performs better than transitional probability. The following section tests this question by using both ΔP and a set of further measures of collocation strength to predict the location of hesitations in spoken English. The relationship between collocation strength and hesitation placement should be the same as that described by Wahl for intonation units: the stronger two words collocate, the more likely they are mentally chunked (or strongly connected) and the less likely a hesitation, such as *um* or *like*, should be placed between them.[3]

3 Data and method

3.1 The corpus

The following analyses are based on the Switchboard NXT corpus of American English (NXT Switchboard Corpus Public Release 2008; Calhoun et al. 2010).

3 If strong collocations are unlikely to be split by either intonation unit boundaries or hesitations, it follows that both should be more likely to fall where collocations are weak(er). These (expected) similarities in the locations of intonation unit boundaries and hesitations should lead to an observable correlation between the two and, indeed, we have some evidence that this is the case (cf. Clark and Fox Tree 2002).

This subset of the larger Switchboard corpus (Godfrey et al. 1992) comprises annotated transcripts of 642 telephone conversations. The callers are previously unacquainted adults representing all dialect areas of the United States who converse about a wide variety of topics. The corpus totals roughly 830,000 words. The quality of the part-of-speech (POS)-tagging and the time alignment of the transcript make the corpus ideally suited for the present purpose.[4]

3.2 Hesitations

The set of hesitations which were analysed consists of unfilled pauses, the fillers *uh* and *um* as well as the discourse markers *well, like, you know* and *I mean*. These discourse markers were included because it has been shown that among other functions, they can be used to mark ongoing lexical and content search or to announce repair sequences (cf. Jucker 1993: 447; Müller 2005: 189; Levey 2006; Fung and Carter 2007: 418). Unfilled pauses were calculated from the given start and end times of words in Switchboard NXT. Based on Goldman-Eisler (1968: 12), pauses shorter than 0.2 s were not considered discontinuities and therefore disregarded. The maximum pause length considered for analysis was 1 s, because beyond this limit chances increase that pauses are no hesitations, but actually due to the speaker being interrupted. Repetitions, self-corrections and drawls were excluded for the sake of data homogeneity.

3.3 Method

The analysis is limited to hesitations placed within three common types of prepositional phrases which have previously been analysed by Maclay and Osgood (1959). This restriction to comparable contexts allows for analyses which go beyond simple comparisons of fluent versus hesitant collocations in that it allows for comparisons of the competing forces in a given syntactic context.

Table 4 shows the phrase types considered for analysis and the number of tokens per phrase. The scripting language R (R Development Core Team 2009)

4 I have occasionally met with skepticism concerning the calculation of word and string frequencies in smallish corpora, yet Wahl (2015: 198–199) finds that studies similar to the present one yield the same results whether frequencies are derived from smaller or larger corpora. Furthermore, larger spoken corpora are often not as accurately POS-tagged and thus provide a different source of noise.

Table 4: Phrase types and number of data-points.

Phrase type	Example	Tokens
Prep Noun	*about baseball*	921
Prep Det Noun	*of the cowboys*	1,078
Prep Det Adj Noun	*in a nice neighbourhood*	395

was used to extract all instances of these phrases from the data which were either preceded or interrupted by a hesitation. The word preceding each phrase was also extracted and will be referred to as "X". Phrases containing a hesitation in more than one position were excluded. Figure 3 shows that hesitation placement varies and that there is no universally preferred location.

In a second step, R was used to calculate collocation strengths. For this purpose, all hesitations listed in Section 3.2 were removed from Switchboard NXT, because testing the hypothesis requires us to know the collocation strength of, for instance, *about baseball*, but not the strengths of *about uh* and *uh baseball*. Furthermore, the corpus was converted to lower case. POS-tags and information about sentence boundaries were kept, but any further markup, including punctuation marks, was excluded. Maintaining sentence boundaries not only ensured that only two-word pairs occurring within the same sentence were

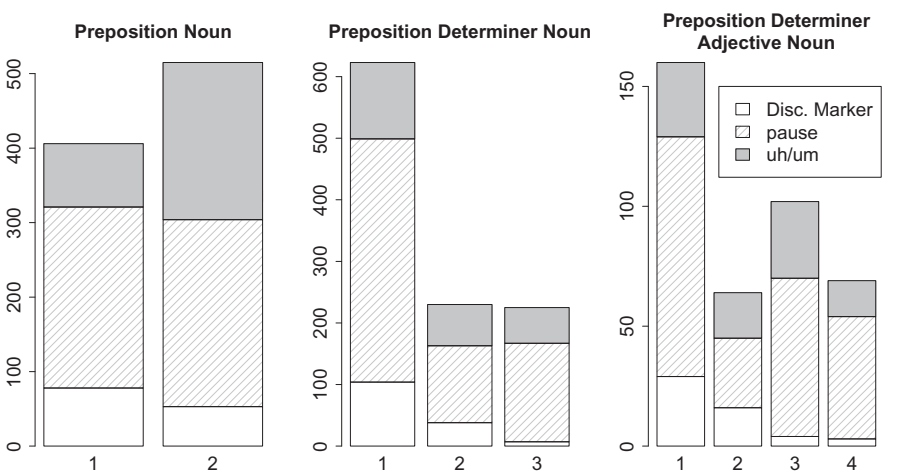

Figure 3: Distribution of hesitations across prepositional phrases. "1" indicates placement before the first word in the phrase (i.e. the preposition), "2" means placement before the second word, etc.

statistically treated as collocations, but also that collocation strengths were never calculated across turn boundaries.

3.4 Measures of collocation strength

In addition to ΔP, the following measures were selected. All of them were calculated for immediately adjacent two-word strings (bigrams). The formulae below refer to the cell names of Table 1 in order to better illustrate that measures are based on different sections of the table and thus on different proportions of available information.

$$\text{Frequency} = a \tag{5}$$

Co-occurrence frequency is the measure that relies on the fewest assumptions. It is commonly used as a simple measure of chunking strength (cf. e.g. Bybee 2007a).

$$\text{Transitional probability}_{\text{forward}} = \frac{a}{a+b} \tag{6}$$

$$\text{Transitional probability}_{\text{backward}} = \frac{a}{a+c} \tag{7}$$

Forward transitional probability measures how likely the first word is to be followed by the second, while backward transitional probability measures how likely the second word is to be preceded by the first. Due to these two perspectives being separated into two distinct measures, transitional probabilities are referred to as unidirectional (just like ΔP). They are frequently employed in cognitive corpus linguistics (cf. e.g. Gregory et al. 1999; Kapatsinski 2005; Tily et al. 2009). However, their use is criticised by Ellis and Ferreira-Junior (2009: 194; see also Wahl's [2015: 204] discussion of their arguments) who argue that Rescorla's (1968) experiments in classical conditioning show that successful conditioning not only depends on the chance of a stimulus being followed by a specific event (as represented by transitional probabilities), but also on the chance of the event occurring without the stimulus (which is incorporated in ΔP). In other words, Ellis and Ferreira-Junior argue that ΔP is a much better predictor of associative learning than transitional probability. Consequently, ΔP should also be a better predictor of collocation strength. Yet, as shown in Section 2.2, we also know that there is only a marginal difference between the two measures when calculated on the basis of corpus data.

$$\text{Mutual Information} = \log\left(\frac{a}{\frac{(a+b)\cdot(a+c)}{a+b+c+d}}\right) \tag{8}$$

The Mutual Information score (*MI*) assesses how strongly two words attract by comparing their actual co-occurrence rate to a chance co-occurrence rate (cf. Oakes 1998; Manning and Schütze 1999). It is a bidirectional measure because, unlike transitional probability and ΔP, it takes associations from left to right as well as from right to left into account. Gries and Mukherjee (2010: 527–528) point out that *MI* has a potential shortcoming: due to syntactic and semantic restrictions, natural language is not based on chance rates of co-occurrence, yet *MI* – like many other measures of collocation strength – is based on this assumption.

$$\text{Lexical Gravity } G = \log\left(\frac{a \cdot (a+b)_{\text{types}}}{a+b}\right) + \log\left(\frac{a \cdot (a+c)_{\text{types}}}{a+c}\right) \quad (9)$$

Lexical Gravity *G*, devised by Daudaravičius and Marcinkevičienė (2004), remedies *MI*'s shortcomings by comparing actual co-occurrence rates to the likelihood of co-occurrence among all *possible* combinations of words. To achieve this, *G* makes use of type frequencies (all other measures of collocation strength only use token frequencies), namely the number of types occurring after the first word (here expressed by $(a+b)_{\text{types}}$) and the number of types occurring before the second word (here expressed by $(a+c)_{\text{types}}$). Despite its comparatively complex calculation, *G* correlates strongly with co-occurrence frequency (see Table 3).

3.5 Statistical evaluation with random forests

The data will be analysed with the help of random forests using the *cforest* command which is part of the *party* package for R (cf. Hothorn et al. 2006; Strobl et al. 2007; Strobl et al. 2008). These algorithms "grow" trees through recursive binary partitioning of the data. For each tree, the algorithm selects a random subset of the data. It then uses the predictor variables to create "branches", i.e. subgroups of the data which are more homogeneous in terms of the dependent variable than their parent groups (cf. Baayen 2008: 148–149; Strobl et al. 2009).

Figure 4 shows an exemplary tree, based on a random sample of 500 hesitations placed in or before phrases of the type "Preposition Determiner Noun". The algorithm partitions the data twice, resulting in a tree with three terminal leaves. The first split indicates that hesitations are predominantly placed at the prepositional phrase boundary (labelled position 1, see Node 2) if the forward transitional probability from the word before the prepositional phrase to the preposition (*TPf*.1) is 7.7% or lower. If it is higher, another factor comes into play, namely Lexical Gravity *G* of the same pair of words (*G*.1). If this

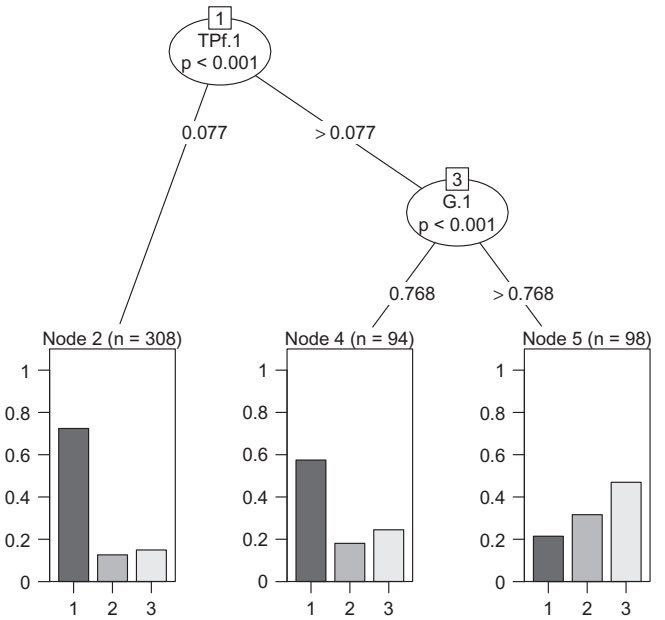

Figure 4: Exemplary tree based on a random 500-word subset of the "Preposition Determiner Noun" data, generated with the *ctree* command in R.

is low, hesitations are still predominantly placed at the prepositional phrase boundary (see Node 4). If *G*, on the other hand, exceeds 0.768, hesitations predominantly occur before the noun (labelled position 3, see Node 5).

The predominant outcome in a terminal node becomes the tree's prediction for all data-points in the node. This means that prediction accuracy is over 70% in Node 2, but just under 50% in Node 5. Overall, it is 64.6%. Prediction accuracy can be evaluated by comparing the numbers of correct and false predictions to the results of a *baseline classifier* which simply generalises from the predominant outcome to all data-points (cf. Baayen 2008: 153). Given that, in the present case, 59.6% of hesitations occur at the prepositional phrase boundary, the baseline classifier predicts that all hesitations occur in this position, resulting in 298 correct predictions. A chi-square test reveals that the tree's predictions constitute a significant improvement over the baseline ($\chi = 5.19$, $p < 0.05$).

Large assemblies of such trees make up a forest. In contrast to the types of trees which are reported as models in their own right, trees in the forest are random, i.e. they each utilise only random subsets of the data and the predictors (cf. Strobl et al. 2009: 332–333) and may thus come to very different results. The overall prediction for each data-point is then determined by vote. Each tree

submits its prediction and the one with the most votes becomes the prediction of the forest (cf. Strobl et al. 2009: 334). Users can set how many predictors may be considered at any one time (*mtry*), how many trees they want in the forest (*ntree*) and save the *random seed*, a number which controls all random parameters, to ensure that a forest can be replicated.

Most analyses are not only interested in the number of correct predictions but also in the strength of the link between the predictors and the dependent variable. In a forest, the latter is obtained through random permutation of predictors, which are then scored based on how much prediction accuracy decreases after permutation. Negative scores indicate that the predictors caused only noise (cf. Strobl et al. 2009: 335, 343).

Both individual classification trees and random forests can handle multinomial outcomes and complex interactions as well as large numbers of predictors (cf. Tagliamonte and Baayen 2012: 161, 171). Forests, however, are particularly apt at handling collinear predictors due to the fact that the number of competing predictors is limited and correlated predictors thus often do not get to perform in a model together (cf. Strobl et al. 2009: 333).

For the present purpose, this procedure has a further advantage: binary splitting of numeric data entails that the data are rank-ordered and that only the rank – not the absolute value – matters, i.e. if a data-point ranks above the splitting point, it is placed in the "above-split" group, whereas it would be placed in the other group if it ranked below the splitting point. As we know, transitional probabilities and corresponding Δ*P*s are almost perfectly correlated, but the correlations are slightly lower when only ranks are considered, as can be seen in Table 5. This means that rank ordering makes transitional probabilities and Δ*P*s slightly more different. Therefore, if Δ*P* has any advantages over transitional probabilities in a corpus-based study, this kind of analysis has a greater chance of recognising them than a regression analysis.

Table 5: Correlations between transitional probabilities and Δ*P*s in the "Preposition Determiner Noun" data set.

	X Prep	Prep Det	Det N
$TP_{for.} \sim \Delta P_{for.}$	1.000	0.998	1.000
	0.891	0.951	0.931
$TP_{ba.} \sim \Delta P_{ba.}$	0.979	0.996	0.999
	0.547	0.920	0.962

Notes: Upper values were determined with the help of (parametric) Pearson correlations, the lower values with non-parametric Kendall's tau.

4 Analyses and results

The following analyses assess how well the measures of association are able to predict where a hesitation is placed. Hesitation placement in each of the three phrase types is analysed separately. In all models, the dependent variable is the location of the hesitation. Model parameters are set as follows. Forests consist of 5,000 trees each (i.e. *ntree* = 5,000) and are permitted to select from a random subset of five predictors per split (*mtry* = 5).

Table 6 shows the number of correct predictions per model. As described in Section 3.5, model performance is evaluated by comparing it to the number of correct predictions of a simple baseline classifier, which predicts that *all* hesitations in phrase type "Preposition Noun" are placed before the noun and that all hesitations in the other two phrase types are placed before the preposition (see Figure 3). Baselines and forests are compared with the help of 2x1-chi-square tests in which baseline performance is defined as expected values and forest performance as observed values. The tests show that all models perform very highly significantly above the baseline, indicating that collocation strengths have an influence on the placement of hesitations.

Table 6: Prediction accuracy of the forest models.

	n	Baseline accuracy	Model accuracy	χ^2	*p*
Prep Noun	921	515 55.9%	764 82.9%	273.1	< 0.001
Prep Det Noun	1,078	623 57.8%	787 73.0%	102.3	< 0.001
Prep Det Adj Noun	395	160 40.5%	272 68.9%	131.8	< 0.001

To investigate how much each predictor contributes to model performance, their variable importance scores are compared. The three panels in Figure 5 separately show the scores the predictors receive from each forest. I added a line at the 0.01 mark in each panel, which serves two functions. First, it shows that absolute values of variable importance cannot easily be compared across models as they are influenced by a variety of factors (cf. Strobl et al. 2009: 336). Here, the more predictors are considered in a model, the lower the scores. Second, in all three cases there appears to be a gap around the 0.01 mark which separates

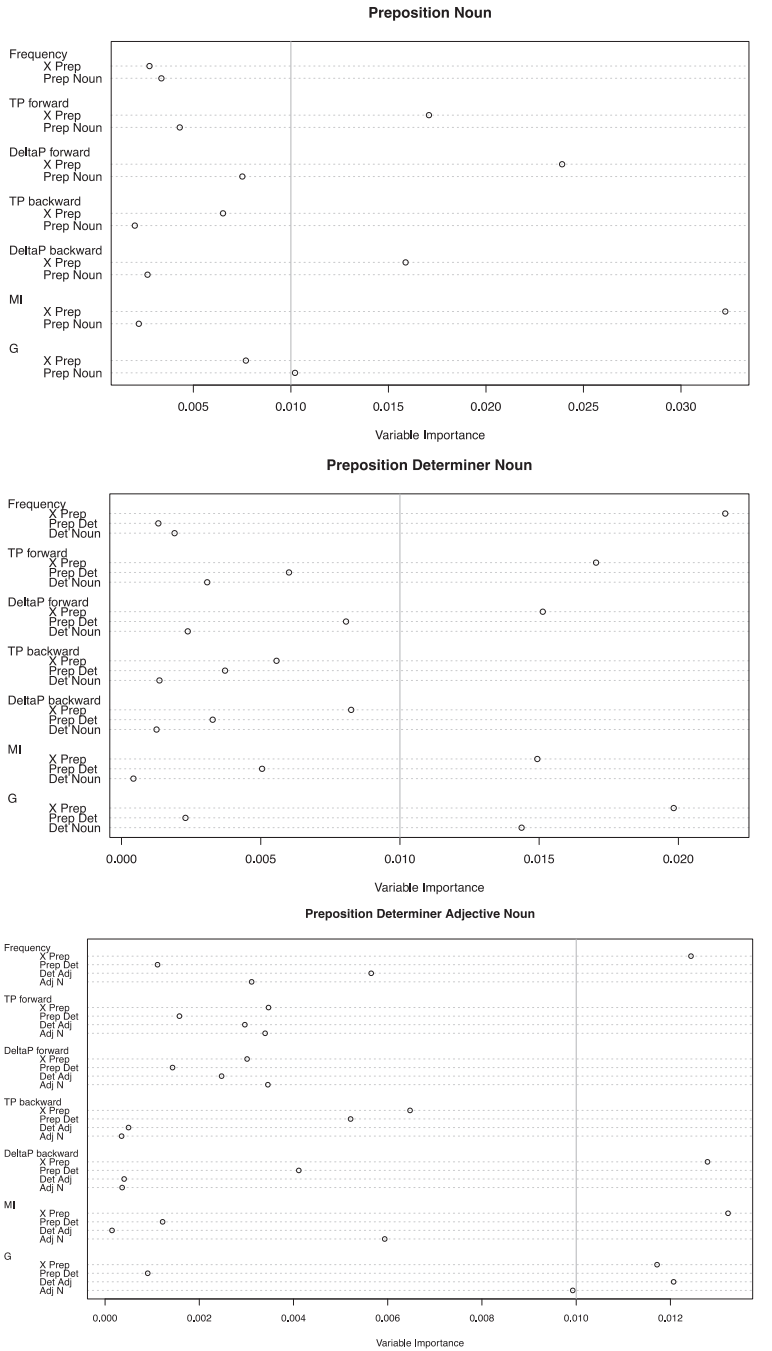

Figure 5: Variable importance scores.

Table 7: List of predictors receiving the best variable importance scores in each model.

Prep Noun		Prep Det Noun		Prep Det Adj Noun	
Predictor	**Var. Imp.**	**Predictor**	**Var. Imp.**	**Predictor**	**Var. Imp.**
MI_{XPrep}	0.032	$Frequency_{XPrep}$	0.022	MI_{XPrep}	0.013
$\Delta P_{for.XPrep}$	0.024	G_{XPrep}	0.020	$\Delta P_{ba.XPrep}$	0.013
$TP_{for.XPrep}$	0.017	$TP_{for.XPrep}$	0.017	$Frequency_{XPrep}$	0.012
$\Delta P_{ba.XPrep}$	0.016	$\Delta P_{for.XPrep}$	0.015	G_{DetAdj}	0.012
		MI_{XPrep}	0.015	G_{XPrep}	0.012
		$G_{DetNoun}$	0.014	G_{AdjN}	0.010

the bulk of the predictors from those which stand out as particularly good ones. Table 7 lists the best predictors in each model in decreasing order of importance. A closer look at the predictors in Table 7 reveals that they are mostly measures of the strength of "X Preposition" collocations, i.e. collocation strengths at the prepositional phrase boundary have a much greater influence on hesitation placement than collocation strengths within the phrase. In this respect, the results from the three trees are very homogeneous. Yet the measures of collocation strength which stand out differ from model to model. In fact, each measure appears in the top group in at least one of the models, giving a first indication that all of them have their merits.[5] We can now address the hypotheses in more detail.

Do ΔPs outperform corresponding transitional probabilities? Particularly forward-directed Δ*P* and forward-directed transitional probability perform very similarly, i.e. both appear in two of the lists of high performing predictors and both perform well in the same syntactic environments. Yet, when there are larger differences between these predictors, it is mostly Δ*P* which outperforms transitional probability (evident in the two upper panels in Figure 5). The case of the backward-directed measures is different, though, as they perform almost identically within the phrase, but at the phrase boundary Δ*P* is consistently the better predictor. Thus, it appears that the minimal differences between Δ*P*s and transitional probabilities under some conditions suffice to bring about a statistical advantage for the former.

Is $\Delta P_{backward}$ a better predictor than $\Delta P_{forward}$? Contrary to Wahl's results, $\Delta P_{backward}$ does not consistently outperform $\Delta P_{forward}$. In fact, it is more often the case that $\Delta P_{forward}$ is the better predictor. Surprisingly, whether backward-

5 Some results at first appear to be a remnant of the randomness of the forests, but this is not the case. Forests of this size lead to very homogeneous results. Furthermore, the variable importance scores reported here are corroborated by mean variable importance scores calculated across 50 forests per phrase type.

directed or forward-directed ΔP performs well, depends not so much on the particular collocation, i.e. its syntactic position, as on the entire phrase type. Only in the model for "Preposition Determiner Adjective Noun" do we find that $\Delta P_{backward}$ outperforms $\Delta P_{forward}$, namely where applied to the phrase boundary and to collocations of the type "Preposition Determiner".

Finally, we can address Gries' (2013) call to pay more attention to directionality effects. He points out that bidirectional measures like *MI* and *G* cannot pick up imbalances in the relation between two words, such as the one between *of* and *course* in the present data set: *course* is highly likely to be preceded by *of*. *Of*, on the other hand, is not strongly attached to *course*. Only unidirectional measures can pick up this imbalance, therefore unidirectional measures, such as ΔP, might better reflect how such imbalanced contexts are processed. Yet there is no evidence in the models that unidirectional measures are better predictors of hesitation placement than bidirectional ones. Of all predictors measuring collocation strength at the phrase boundary, *MI* consistently performs on par or even better than ΔP and transitional probability. In two of the tree models, *G* is also among the top group of predictors for this collocation. Furthermore, *G* most accurately predicts the influence that collocations containing a content word have on the hesitation process.

5 Discussion and conclusion

The analysis presented in this paper has shown that the placement of hesitations in spontaneous speech correlates significantly with collocation strength. The latter was operationalised by means of seven different statistical association measures ranging from simple co-occurrence frequency via unidirectional measures – including relatively new ΔP – to bidirectional *MI* and *G*. The finding that statistically stronger collocations are less likely to be uttered disfluently can be interpreted as evidence for the usage-based tenet that there is a link between usage and sequence storage.

Moreover, preliminary analyses showed that the larger the database, the smaller the difference between ΔP and transitional probability. Moreover, even when calculated for individual novels (which consist of much fewer words than the average modern corpus), ΔP is little more than transitional probability adjusted by at most seven percentage points. Yet, analyses of hesitation placement showed that this minor difference between transitional probability and ΔP in some circumstances suffices to bring about a statistical advantage for the latter. In summary, the analyses provide a tentative indication that Ellis and

Ferreira-Junior's (2009) claim that ΔP is better suited to model associative learning than transitional probabilities, can be transferred to the modelling of speech production.

As ΔP is a directional measure, another focus of the present study was on assessing whether both directions are equally predictive when modelling the processing of English. As a point of comparison, I have drawn on Wahl (2015) who analyses the placement of intonation unit boundaries – another production phenomenon – based also on a usage-based model of language processing and using a similar methodology. Wahl finds an extreme divergence between the performance of $\Delta P_{\text{backward}}$ and $\Delta P_{\text{forward}}$. While the former performs on par with bidirectional measures in his analysis, the latter turns out to be a very poor predictor. He interprets his results as corroborating Onnis and Thiessen's (2013) finding that backward associations are more informative in English than forward associations, because, in English, typically a closed class item is followed by an open-class item (Wahl 2015: 204–205, 214).

In language, closed classes are typically small(ish), but contain very versatile elements, which co-occur with a large variety of different words. Open classes, on the other hand, are typically large, with members that are more restricted in use. Therefore, if we know a function word, like a preposition or a determiner, it is still difficult to guess which content word, e.g. noun, will follow, i.e. forward transitional probability will be low. If we are given a noun, on the other hand, predicting the element preceding it is typically much easier, i.e. backward transitional probability will be higher. Additionally, we have reason to assume that planning in these contexts is also done from right to left as speakers might plan heads of noun phrases before any other words in the phrase.

Based on this reasoning, backward-directed measures of collocation strength should be good predictors of the chunkiness of collocations in which the first word is a function word and the second is a content word or where the first word is more frequent than the second. Table 8 shows that these two conditions generally coincide.

Interestingly, the contexts where backward-directed measures perform well in my data are not in the "first-frequent" group. In fact, first-frequent contexts are where the backward-directed measures perform worst. We might conclude that transitions so far into the phrase simply have no influence on the placement of hesitations, but the excellent performance of G in these contexts is evidence that they do.

Table 8 furthermore shows that generalising from first-frequent collocations to all two-word pairs in English may be much too simplistic. In the given prepositional phrase contexts, only a third of all collocations follows this pattern, which may be a special property of this particular environment, but it

Table 8: Types of collocations in the three data sets. Percentages indicate which word in the collocation is generally the more frequent.

Type of collocation	Prep. Noun	Prep. Det. Noun	Prep. Det. Adj. N.
X Preposition	Second (84.0%)	Second (81.6%)	Second (82.8%)
Preposition Noun	First (97.9%)		
Preposition Det		Second (76.2%)	Second (81.5%)
Det Noun		First (98.1%)	
Det Adjective			First (99.0%)
Adjective Noun			52.2% vs. 47.3%

cautions that the case may be similar in other contexts as well. Unfortunately, Wahl does not take syntactic factors into account, so we do not know whether syntax and the effects of different association measures interacted in his data. Finally, even though we know that there is a correlation between intonation unit boundaries and hesitation placement (cf. Clark and Fox Tree 2002), they may depend on different factors.

Finally, ΔP – being unidirectional – was compared to *G* and *MI* – both bidirectional. The results provide no evidence that any approach is generally preferable. However, as mentioned above, *G* stands out as the measure most accurately predicting the influence that collocations containing a content word have on the hesitation process. This could be interpreted as evidence that *G* best captures the processing of content words and their immediate surroundings. Yet it must not be seen as an indication that any bidirectional measure is a better predictor of chunking than a unidirectional one. The finding should rather be attributed to the specific properties of *G* itself, because comparing, for instance, *G* and ΔP is not just a question of comparing directionality and non-directionality, but may also reflect the quantity of information contained in the measures, as *G* uses contingency information besides frequency information while ΔP only uses the latter.[6]

In conclusion, I propose that future studies rely on several measures, both bi- and unidirectional, as a predictor's performance depends hugely on the POS it is applied to. Furthermore, this paper addressed only part of "the new ways of studying collocations" proposed by Gries (2013: 159). For instance, the

6 I thank an anonymous reviewer for drawing my attention to this fact.

performance of ΔP for longer collocations still needs to be assessed and compared to that of other association measures such as the ones discussed here.

References

Allan, Lorraine G. 1980. A note on measurement of contingency between two binary variables in judgement tasks. *Bulletin of the Psychonomic Society* 15(3). 147–149.

Arnon, Inbal & Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62. 67–82.

Baayen, R. Harald. 2008. *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. Harald. 2009. LanguageR: Data sets and functions with 'Analyzing Linguistic Data: A practical introduction to statistics'. R package version 0.955. http://CRAN.R-project.org/package=languageR.

Beattie, Geoffrey & Brian L. Butterworth. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22(3). 201–211.

Beckner, Clay, Richard Blythe, Morten H. Joan Bybee, William Croft Christiansen, Nick C. Ellis, John Holland, Ke Jinyun, Diane Larsen-Freeman & Tom Schoeneman. 2009. Language is a complex adaptive system: Position paper. *Language Learning* 59(Supplement 1). 1–26.

Bell, Alan, Daniel Jurafsky, Eric Fosler-Lussier, Cynthia Girand, Michelle Gregory & Daniel Gildea. 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113(2). 1001–1024.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *Longman grammar of spoken and written English*. Harlow: Pearson.

Bod, Rens. 2010. Probabilistic linguistics. In Bernd Heine & Heiko Narrog (eds.), *The Oxford handbook of linguistic analysis*, 633–662. Oxford: Oxford University Press.

Bresnan, Joan & Jessica Spencer. 2013. *Frequency and variation in English subject-verb contraction*. Stanford, CA: Stanford University Department of Linguistics and Center for the Study of Language and Information.

Brezina, Vaclav, Tony McEnery & Stephen Wattam. 2015. Collocations in context. A new perspective on collocational networks. *International Journal of Corpus Linguistics* 20(2). 139–173.

Bybee, Joan. 1998. The emergent lexicon. *Chicago Linguistics Society* 34: *The Panels*. 421–435.

Bybee, Joan. 2002. Phonological evidence for the exemplar storage of multiword sequences. *Studies in Second Language Acquisition* 24(2). 215–221.

Bybee, Joan. 2006. From usage to grammar: The mind's response to repetition. *Language* 82(4). 711–733.

Bybee, Joan. 2007a. *Frequency of use and the organization of language*. Oxford: Oxford University Press.

Bybee, Joan. 2007b. Sequentiality as the basis of constituent structure. In Joan Bybee (ed.), *Frequency of use and the organisation of language*, 313–335. Oxford: Oxford University

Press. (Reprinted from Talmy Givón & Bertram F. Malle (eds.), *The evolution of language out of pre-language*. Amsterdam: John Benjamins. 2002. 107–132.).
Bybee, Joan. 2010. *Language, usage, and cognition*. Cambridge: Cambridge University Press.
Bybee, Joan & James L. McClelland. 2005. Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review* 22. 381–410.
Bybee, Joan & Joanne Scheibman. 2007. The effect of usage on degrees of constituency. The reduction of *don't* in English. In Joan Bybee (ed.), *Frequency of use and the organisation of language*, 294–312. Oxford: Oxford University Press. (Reprinted from *Linguistics* 37(4). 1999. 575–596.).
Calhoun, Sasha, Jean Carletta, Jason Brenier, Neil Mayo, Daniel Jurafsky, Mark Steedman & David Beaver. 2010. The NXT-format switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation Journal* 44. 387–419.
Clark, Herbert H. & Jean E. Fox Tree. 2002. Using *uh* and *um* in spontaneous speaking. *Cognition* 84. 73–110.
Croft, William. 2001. *Radical construction grammar: Syntactic theory in typological perspective*. Oxford: Oxford University Press.
Daudaravičius, Vidas & Marcinkevičienė. Rūta. 2004. Gravity counts for the boundaries of collocations. *International Journal of Corpus Linguistics* 9(2). 321–348.
Eikmeyer, Hans-Jürgen, Ulrich Schade, Marc Kupietz & Uwe Laubenstein. 1999. A connectionist view of language production. In Rolf Klabunde & Christiane Von Stutterheim (eds.), *Representations and processes in language production*, 205–236. Wiesbaden: Deutscher Universitätsverlag.
Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27(1). 1–24.
Ellis, Nick C. & Fernando Ferreira-Junior. 2009. Constructions and their acquisition. Islands and the distinctiveness of the occupancy. *Annual Review of Cognitive Linguistics* 7. 187–220.
Ellis, Nick C., Rita Simpson-Vlach & Carson Maynard. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics and TESOL. *TESOL Quarterly* 24(3). 375–396.
Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14. 179–211.
Evert, Stefan. 2004. *The statistics of co-occurrences: Word pairs and collocations*. Stuttgart: Institut für maschinelle Sprachverarbeitung, University of Stuttgart dissertation.
Fillmore, Charles J., Paul Kay & Mary Catherine O'Connor. 2003. Regularity and idiomaticity in grammatical constructions: The case of *let alone*. In Michael Tomasello (ed.), *The new psychology of language: Cognitive and functional approaches to language structure*, 243–270. Mahwah, NJ: Lawrence Erlbaum.
Fried, Mirjam & Östman. Jan-Ola. 2004. Construction grammar: A thumbnail sketch. In Mirjam Fried & Jan-Ola Östman (eds.), *Construction grammar in a cross-language perspective*, 11–86. Amsterdam/Philadelphia: John Benjamins.
Frisson, Steven, Keith Rayner & Martin J. Pickering. 2005. Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory and Cognition* 31(5). 862–877.
Fung, Loretta & Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28(3). 410–439.

Godfrey, John J., Edward Holliman & McDaniel. Jane 1992. SWITCHBOARD: Telephone speech corpus for research and development. *International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1992* 1. I-517–I-20.

Goldberg, Adele. 2005. *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goldman-Eisler, Frieda. 1968. *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

Gregory, Michelle L., William D. Raymond, Alan Bell, Eric Fosler-Lussier & Daniel Jurafsky. 1999. The effects of collocational strength and contextual predictability in lexical production. *Communication and Linguistic Studies* 35. 151–166.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next …. *International Journal of Corpus Linguistics* 18(1). 137–165.

Gries, Stefan Th. 2014. *Coll.analysis 3.5. A script for R to compute perform collostructional analyses*. http://www.linguistics.ucsb.edu/faculty/stgries/teaching/groningen/index.html.

Gries, Stefan Th. 2015a. More (old and new) misunderstandings of collostruction analysis: On Schmidt & Küchenhoff (2013). *Cognitive Linguistics* 26(3). 505–536.

Gries, Stefan Th. 2015b. The role of quantitative methods in cognitive linguistics. In Jocelyne Daems, Eline Zenner, Kris Heylen, Dirk Speelman & Hubert Cuyckens (eds.), *Change of paradigms – New paradoxes. Recontextualizing language and linguistics*. Berlin/Boston: De Gruyter Mouton.

Gries, Stefan Th. & Joybrato Mukherjee. 2010. Lexical gravity across varieties of English: An ICE-based study of n-Grams in Asian Englishes. *International Journal of Corpus Linguistics* 15(4). 520–548.

Hothorn, Torsten, Kurt Hornik & Achim Zeileis. 2006. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics* 15(3). 651–674.

Jenkins, Herbert M. & William C. Ward. 1965. Judgement of contingency between responses and outcomes. *Psychological Monographs* 79(1). 1–17.

Jucker, Andreas. 1993. The discourse marker *well*: A relevance-theoretical account. *Journal of Pragmatics* 19. 435–452.

Jurafsky, Daniel, Alan Bell, Eric Fosler-Lussier, Cynthia Girand & William D. Raymond. 1998. Reduction of English function words in Switchboard. *Proceedings of the International Conference of Spoken Language Processing, Sydney*. 1–4.

Jurafsky, Daniel & James H. Martin. 2008. *Speech and language processing. An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson/Prentice Hall International.

Kapatsinski, Vsevolod M. 2005. Measuring the relationship of structure to use: Determinants of the extent of recycle in repetition repair. *Berkeley Linguistics Society* 30. 481–492.

Kapatsinski, Vsevolod M. & Joshua Radicke. 2009. Frequency and the emergence of prefabs: Evidence from monitoring. In Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali & Kathleen M. Wheatley (eds.), *Formulaic language. Vol. 2: Acquisition, loss, psychological reality, functional explanations*, 499–520. Amsterdam/Philadelphia: John Benjamins.

Langacker, Ronald W. 2000. A dynamic usage-based model. In Suzanne Kemmer & Michael Barlow (eds.), *Usage-based models of language*, 1–63. Stanford, CA: CSLI Publications.

Levey, Stephen. 2006. The sociolinguistic distribution of discourse marker *like* in preadolescent speech. *Multilingua* 25. 413–441.

Maclay, Howard & Charles E. Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15. 19–44.
Manning, Christopher D. & Hinrich Schütze. 1999. *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
Müller, Simone. 2005. *Discourse markers in native and non-native English discourse*. Amsterdam/Philadelphia: John Benjamins.
NXT Switchboard Corpus Public Release. 2008. *Philadelphia: Linguistic Data Consortium*. Catalog #LDC2009T26.
Oakes, Michael. 1998. *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.
Onnis, Luca & Eric Thiessen. 2013. Language experience changes subsequent learning. *Cognition* 162(2). 168–284.
Pecina, Pavel. 2010. Lexical association measures and collocation extraction. *Language Resources and Evaluation* 44(1/2). 137–158.
Perruchet, Pierre & Sebastien Pacton. 2006. Implicit learning and statistical learning: One phenomenon, two approaches. *TRENDS in Cognitive Sciences* 10(5). 233–238.
Phillips, Martin K. 1983. *Lexical macrostructure in science text*. Birmingham: University of Birmingham dissertation.
R Development Core Team. 2009. *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. http://www.R-project.org.
Reali, Florencia & Morten H. Christiansen. 2007. Processing of relative clauses is made easier by frequency of occurrence. *Journal of Memory and Language* 57. 1–23.
Rescorla, Robert A. 1968. Probability of shock in the presence and absence of CS in fear conditioning. *Journal of Comparative Physiological Psychology* 66. 1–5.
Rumelhart, David E. & James L. McClelland (eds.). 1986. *Parallel distributed processing: Explorations in the microstructure of cognition. Foundations*, vol. 1. Cambridge, MA/ London: MIT Press/Bradford.
Schmid, Hans-Jörg & Küchenhoff. Helmut. 2013. Collostructional analysis and other ways of measuring lexicogrammatical attraction: Theoretical premises, practical problems and cognitive underpinnings. *Cognitive Linguistics* 24(3). 531–577.
Schneider, Ulrike. 2014. *Frequency, chunks and hesitations. A usage-based analysis of chunking in English*. Freiburg: NIHIN Studies. https://freidok.uni-freiburg.de/data/9793
Schneider, Ulrike. 2016. Chunking as a factor determining the placement of hesitations. A corpus-based study of spoken English. In Heike Behrens & Stefan Pfänder (eds.), *Frequency effects in language: What counts in language processing, acquisition and change*, 61–89. Berlin/New York: Mouton De Gruyter.
Shanks, David R. 1995. *The psychology of associative learning*. Cambridge: Cambridge University Press.
Shriberg, Elizabeth & Andreas Stolcke. 1996. Word predictability after hesitations: A corpus-based study. *Proceedings of the International Conference on Spoken Language Processing*. 1868–1871.
Strobl, Carolin, Anne-Laure Boulestreix, Thomas Kneib, Thomas Augustin & Achim Zeileis. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9. 307.
Strobl, Carolin, Anne-Laure Boulestreix, Achim Zeileis & Torsten Hothorn. 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8. 25.

Strobl, Carolin, James Malley & Gerhard Tutz. 2009. An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods* 14(4). 323–348.

Tagliamonte, Sali A. & R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. *Language Variation and Change* 24. 135–178.

Tily, Harry, Susanne Gahl, Inbal Arnon, Neal Snider, Anubha Kothari & Joan Bresnan. 2009. Syntactic probabilities affect pronunciation variation in spontaneous speech. *Language and Cognition* 1(2). 147–165.

Vogel Sosa, Anna & James MacFarlane. 2002. Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word. *Journal of Brain and Language* 83. 227–236.

Wahl, Alexander. 2015. Intonation unit boundaries and the storage of bigrams. Evidence from bidirectional and directional association measures. *Review of Cognitive Linguistics* 13(1). 191–219.

Ward, William C. & Herbert M. Jenkins. 1965. The display of information and the judgement of contingency. *Canadian Journal of Experimental Psychology* 19(3). 231–241.

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.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.