

## Abstract

*We investigate asymmetry in corpus-derived and human word associations. Most prior work has studied paradigmatic relations, either derived from free association norms or from large corpora using measures of statistical association and semantic relatedness. By contrast, we investigate the syntagmatic relation between words in adjective-noun and noun-noun combinations and present a new experimental design for measuring the strength of human associations. Of particular importance for syntagmatic relations are asymmetric associations, whose associational strength is much larger in one direction (e.g., from Pyrrhic to victory) than in the other (e.g., from victory to Pyrrhic). We develop a number of corpus-derived measures of asymmetric association and show that they predict the directedness of human associations with high accuracy.*

*Keywords* association measures; asymmetry; paradigmatic and syntagmatic associations; corpus-based; elicitation experiment; association norms

## 1 Introduction

For many tasks in natural language processing (NLP), it is useful to have available an automatic assessment of how closely two words are related to each other or how strongly they are associated with each other. Notions of relatedness and association are often defined as catch-all categories that lump together many different ways in which two words can be related. Corpus-based measures of relatedness and association are typically evaluated against data produced by human subjects in elicitation and rating experiments, again often conflating different types of lexical relations. Rather than taking association as an atomic notion, we argue in this article that there are four different types of lexical association that can be classified along the dimensions syntagmatic–paradigmatic and symmetric–asymmetric as shown in Table 1; and that corpus-based measures as well as gold standard data should be designed for the type of relatedness that is relevant in a particular context.

The ideas behind syntagmatic and paradigmatic relations between words have their origin in the work of de Saussure (1966). Traditionally, the relationship between two words is called syntagmatic if they occur in sequence:

Combinations based on sequentiality may be called *syntagmas*. The syntagma invariably comprises two or more consecutive units [...]. In its place in a syntagma, any unit acquires its value simply in opposition to what precedes, or to what follows, or to both.

(de Saussure 1966: 121)

In this paper, we use the term syntagmatic in the sense of morphosyntactic relations, specifically noun-noun compounds and prenominal adjectives.

In contrast, paradigmatic relations are orthogonal to the sequential syntagmatic axis. Two words are said to be paradigmatically related if they can be substituted for each other. Such words usually have the same part of speech.

Many prototypical paradigmatic association pairs like *girl–boy* and *good–bad* are *symmetric*, by which we mean that they prime each other with similar strength in free association experiments. As we will show in this article, syntagmatic associations are frequently

*asymmetric*; they consist of two elements where one strongly predicts the other, but not vice versa.

We give examples for each of the four possible types of association in Table 1. The pairs *bad-good* and *bird-canary* are paradigmatic. The pairs *epileptic-seizure* and *Christmas-decorations* are syntagmatic. The pairs *bad-good* and *epileptic-seizure* are symmetric: their elements prime each other with about the same strength. For example, in the USF word association norms (see Section 2.2), 75% of subjects give *good* as a response for *bad*, and 76% give *bad* as a response for *good*. In our elicitation experiment, we found that *epileptic-seizure* received a forward score of .541 and a backward score of .462 which supports our symmetry claim (see Section 5.1 for a definition of these scores).

The pairs *bird-canary* and *Christmas-decorations* are asymmetric. The second element strongly primes the first, but the first only induces a weak tendency for subjects to name the second. In the USF data set, 69% of subjects give *bird* as a response for *canary*, but only 6% give *canary* as a response for *bird*. Similarly, *Christmas-decorations* receives a forward score of 0.266 and a backward score of 0.82, backing our asymmetry claim.

In this article, we demonstrate that, for adjective-noun and noun-noun combinations, these asymmetry effects are characteristic of human linguistic performance and can be accurately predicted from corpus data. Since standard evaluation sets such as USF or EAT and related pair data (see Section 2.2) are not suitable for our purposes – because they lump together different types of associations and do not measure asymmetry – we designed and conducted a novel experiment in order to elicit human syntagmatic associations.

The article is structured as follows. In Section 2, we describe our motivation and the background to our study. In Section 3, we describe asymmetry effects in corpus data and develop suitable statistical association measures. In Section 4, we present our experimental design for measuring human syntagmatic associations. Section 5 analyzes the results obtained for a sample of adjective-noun and noun-noun combinations, and shows that the directedness of human association is accurately predicted by the corpus data. Section 6 presents our conclusions.

## 2 Background

### 2.1 Motivation for studying asymmetric measures

Tversky (1977) argued that similarity is an asymmetric relation, criticizing the inherently symmetric aspect of metric-based models of similarity. He backed his view with a number of rating experiments in which subjects had to assess the similarity between different kinds of objects, for example figures, letters and countries. *North Korea*, for example, is judged more similar to *China* than vice versa. According to Tversky, the reason for this lies in the subjects’ feature representation of the two words. A large number of features are used to represent the concept *China*, only some of which are also included in the representation of *North Korea*. Conversely, a small number of features are used for *North Korea*, many of which are part of *China*’s representation.

Tversky showed that asymmetry in similarity is a cognitive phenomenon; but it can also be measured in corpus data. In the context of estimating co-occurrence probabilities for unseen events in language models, several measures of distributional similarity were discussed (Dagan et al. 1999). While most of the studied measures are symmetric, one asymmetric measure has received further attention: the alpha skew divergence  $s_\alpha$  (Lee 1999; 2001). It is a weighted version of the asymmetric Kullback-Leibler divergence (Kullback and Leibler 1951). Lee (1999) mentions the subject of asymmetry in similarity, but does not investigate it further.

Weeds (2002) emphasizes the asymmetric aspect of the skew divergence and its potential usefulness in capturing asymmetry in similarity. She links asymmetric substitutability to the hypernymy relation and proposes that *fruit* and *apple* are similar to each other but

Table 1: Relation matrix for stimulus-response pairs

	paradigmatic	syntagmatic
symmetric	<i>bad – good</i>	<i>epileptic seizure</i>
asymmetric	<i>bird – canary</i>	<i>Christmas decorations</i>

*fruit* is more similar to *apple* than *apple* is to *fruit*. Applied to hypernymy, this would be reflected in  $s_\alpha(\text{hyper}(x), x)$  being lower than  $s_\alpha(x, \text{hyper}(x))$  (a lower score means higher similarity). In an initial experiment, Weeds was able to predict hypernyms and hyponyms in 156 pre-selected word pairs in over 90% of the cases using the above formula.

In a recent study, Michelbacher et al. (2007) have examined asymmetry in paradigmatic associations. To capture the kind of asymmetric *apple-fruit* relation, they define asymmetric rank measures based on Pearson’s  $\chi^2$  test and conditional probabilities. They gathered asymmetric association data from the British National Corpus (BNC) and evaluated the results against data computed from the USF Free Association Norms (see Section 2.2). The measures were able to predict asymmetry in associations but with a relatively high error rate.

More recent work in cognitive science has looked at syntagmatic and paradigmatic associations as inspiration for or tests of computational models (Griffiths et al. 2007; Dennis 2004; Jones and Mewhort 2007; Schütze and Walsh 2008).

## 2.2 Free association experiments

This section contains a brief description of the free association experiments and the corresponding norms that are related to the research presented in this paper, namely the so-called Minnesota norms collected by Russell and Jenkins (Jenkins 1970), the Palermo and Jenkins set (Palermo and Jenkins 1964), the University of South Florida norms (USF, Nelson et al. 1998) and the Edinburgh Word Association Thesaurus (EAT, Kiss et al. 1973).

**The Minnesota norms, Palermo and Jenkins** Both norms are closely related. The Minnesota norms were collected by presenting 100 stimulus words<sup>1</sup> to 1,008 college students of introductory psychology classes in 1952 (Wettler and Rapp 1993). The well-known Palermo and Jenkins data set was presented in *Word association norms: Grade school through college* (Palermo and Jenkins 1964). It is an extension of the previous experiment including students of different age groups. In addition to the 100 original words, another 100 words more suitable for young speakers were added. A variety of parts-of-speech including nouns, adjectives, verbs, adverbs and prepositions were used. In both studies, each stimulus word was presented with a blank line to the right of it and subjects were asked to write what first came to their mind on the line. 1,000 subjects ranging from 4th graders to undergraduate students took part in the study.

**USF** The *University of South Florida Word Association Rhyme and Word Fragment Norms* is a collection of word associations compiled by Nelson et al. at the University of South Florida. Data collection started in 1973 and went on for two decades. More stimulus words were added over the course of time. The finished data set was published in 1998. On average, each stimulus word was presented to around 150 subjects and each subject had to complete a booklet of 100 to 200 words. In total, the database contains 5,019 stimulus

words. The elicitation procedure was almost identical to the one used by Palermo and Jenkins (1964). More than 6,000 participants produced nearly 750,000 responses. The full database with detailed information about every stimulus-response pair is available for download at <http://web.usf.edu/FreeAssociation/>.

**EAT** The *Edinburgh Word Association Thesaurus* was created by Kiss et al. (1973). It contains 8,400 stimulus words including the stimuli used by Palermo and Jenkins (1964). Each stimulus was presented to 100 different subjects. The elicitation procedure was, again, very similar to Palermo and Jenkins, namely, that subjects were presented a list of stimuli without context and were asked to write down the first word they could think of. Subjects were urged to complete the task as quickly as possible. An interactive version of the data set is available online at <http://www.eat.rl.ac.uk/>.

## 2.3 Research using elicited data

In psycholinguistics, researchers have been studying association norms for over a century to explore the organization of the mental lexicon and how information is retrieved from it during language production and comprehension (e.g. Clark 1971). We refer the reader to Mollin (2009) for a more detailed discussion of word association norms in psycholinguistics. Association norms have also been used as benchmarks for models of human semantic knowledge (Griffiths et al. 2007).

Church and Hanks (1990) were among the first to observe that measures of relatedness derived from machine-readable corpora correlate with the human responses given in free association and rating tasks. A number of studies have confirmed that human associations can be predicted with the aid of corpus-based association measures (e.g. Spence and Owens 1990; Rapp 2002; Sahlgren 2006; Michelbacher et al. 2007). However, the data sets utilized in these studies (for example, USF or EAT – see Section 2.2) do not distinguish between different types of semantic relatedness or association and they only contain a small portion of syntagmatic combinations. Table 2 gives examples of the various relationships between stimulus and response that occur in these data sets. The table was compiled by Hutchison (2003) who classified each stimulus and response pair of Palermo and Jenkins’s norms. Almost all relations are paradigmatic, but three comprise syntagmatic pairs: on the one hand, the groups that Hutchison called *forward* and *backward phrasal associates* and that we refer to as syntagmatic combinations in our terminology. On the other hand, the group labeled *associated properties* can also be thought of as syntagmatic, for example in adjective coordinations (*a deep, dark hole*). In total, only 16.7% of the pairs were classified into these relations.<sup>2</sup> Washtell and Markert (2009) report higher numbers of syntagmatic relations in free associations. For two data sets, Kent and Rosanoff (1910) and Russell and Jenkins (Jenkins 1970), they found 27% and 39%, respectively. Apart from the fact that they used different data sets than Hutchison, a likely cause for the higher number lies in Washtell and Markert’s definition of syntagmatic. It is more lax than ours covering meronymy, holonymy and other “harder-to-classify topical or idiomatic relationships (*family–Christmas, rock–roll*)” (Washtell and Markert 2009: 1).

Since we focus on syntagmatic combinations, only a low number of stimulus-response pairs is suitable for our investigation. Hence, we think that it is justified to disregard free association data in favor of a novel experiment designed for syntagmatic combinations (see Section 4).

There has been a large body of work on evaluating corpus-derived measures of semantic relatedness (including Miller and Charles (1991); Resnik (1996); Finkelstein et al. (2002); Gurevych (2005); Budanitsky and Hirst (2006); Strube and Ponzetto (2006); Gabrilovich and Markovitch (2007); as well as Lapata et al. (2001) and Keller and Lapata (2003) for syntagmatic combinations). These studies often use the data set by Rubenstein and Goodenough (1965) or similar data. Rubenstein and Goodenough’s data contains word

Table 2: Common relationships between stimulus and response in Palermo and Jenkins (1964) association norms compiled by Hutchison (2003)

Association Type (and Example)	Percentage Rate
Synonyms ( <i>afraid-scared</i> )	14.1
Antonyms ( <i>day-night</i> )	24.3
Natural category ( <i>sheep-goat</i> )	9.1
Artificial category ( <i>table-chair</i> )	5.1
Perceptual only ( <i>pizza-saucer</i> )	0.0
Supraordinate ( <i>dog-animal</i> )	5.6
Perceptual property ( <i>canary-yellow</i> )	11.1
Functional property ( <i>broom-sweep</i> )	12.1
Script relation ( <i>orchard-apple</i> )	6.1
Instrument ( <i>broom-floor</i> )	6.1
Forward phrasal associate ( <i>baby-boy</i> )	11.6
Backward phrasal associate ( <i>boy-baby</i> )	4.1
Associated properties ( <i>deep-dark</i> )	1.0
Unclassified ( <i>mouse-cheese</i> )	5.1

pairs together with a numerical value indicating the relatedness between the components of the pair. This value was determined in experiments with humans wherein subjects had to rate the relatedness between given pairs on a fixed scale. This methodology obscures any possible asymmetry effect because both words are presented to the user simultaneously. One goal of our article is to draw attention to different types of association and relatedness and to the importance of stating clearly which type of relatedness is relevant in a particular scenario (symmetric or asymmetric, syntagmatic or paradigmatic) and evaluating related pairs accordingly.

In cognitive linguistics, there is a general consensus about the necessity to support hypotheses about linguistic phenomena and theories with usage-based evidence. It remains unclear, however, which methodological approach to corpus data is most suitable for obtaining such evidence, and whether different techniques are needed for different phenomena and hypotheses. Some recent studies use data elicited in psycholinguistic experiments in order to evaluate different methods of analyzing corpus data. It has been found, for example, that aspects of human language processing can be modeled with association measures and that different association measures vary in their ability to predict human intuitions (e.g. Wiechmann 2008; Gries et al. 2005).

We see our study as a further step in this direction. In accordance with the approaches sketched above, we employ a number of statistical measures and compare their predictions with data obtained from human subjects. However, we move our focus to a phenomenon that has not been considered in previous studies, namely the asymmetry of word associations.

## 2.4 Right- and left-predictive combinations

Symmetry and asymmetry of syntagmatic relations have been investigated by Kjellmer (1991):

A large part of our mental lexicon consists of combinations of words that customarily co-occur. The occurrence of one of the words in such a combination can be said to predict the occurrence of the other(s). (Kjellmer 1991: 112)

These word combinations are either symmetric or asymmetric. In *right-predictive* asymmetric combinations such as *Pyrrhic victory*, *bonsai tree* or *wellington boots*, the first component suggests (or predicts) the second, but not the other way around. For *left-predictive* asymmetric combinations, the opposite is the case: the second components of *deadly nightshade*, *high fidelity*, and *arms akimbo* suggest the first components, but not vice versa. For our study, we focus on two-word adjective-noun and noun-noun combinations that occur within noun phrases. Many recurring word pairs of this type tend to appear in uninterrupted sequence, which makes them more suitable for elicitation experiments than, e.g., verb-object combinations that are often discontinuous.

Sinclair (1991) introduced the notion of *upward* and *downward collocation*. In his terminology, a collocation – “the occurrence of two or more words within a short space of each other in a text” (Sinclair 1991: 170) – consists of a *base* word and a *collocate*. In an upward collocation, the collocate is more frequent than the base; in a downward collocation, the collocate is less frequent than the base. Based on the assumption that *new* is more frequent than *tree* which is in turn more frequent than *bonsai*, *new tree* is an instance of upward collocation and *bonsai tree* is an instance of downward collocation with base *tree* and collocates *new* and *bonsai*, respectively. In relation to Kjellmer’s notions, we expect stronger predictiveness from collocate to base in the case of downward collocation, and vice versa for upward collocation.

### 3 Asymmetric association measures

#### 3.1 Corpus data

The corpus associations used in this work are based on data extracted from the XML edition of the BNC.<sup>3</sup> Following Evert and Kermes (2003), we implemented an *extraction pipeline* with the following three stages:

1. Add linguistic information to the corpus in the form of part-of-speech (POS) tags and lemmatization.
2. Extract a list of suitable word pairs (here, adjective-noun and noun-noun combinations) based on POS patterns and other morphosyntactic constraints. Optionally, the size of the word list can be reduced with a number of linguistic and heuristic filters.
3. Use statistical measures to compute the association strength of each word pair, based on co-occurrence frequency data in the form of a contingency table. For our purposes, special asymmetric association measures are required (see Section 3.3).

The first step of the pipeline uses the lemmatization and C5 part-of-speech tags provided as part of the BNC annotation. The C5 tagset consists of 61 different tags, which were automatically assigned by the CLAWS tagger (Leech et al. 1994). Our pipeline extracts lemma pairs rather than word pairs because preliminary experiments indicated that syntagmatic associations often hold between lemmas rather than particular word forms. Throughout this article, we refer to lemmas and lemma pairs simply as “words” and “word pairs”.

In the second step, we extracted adjective-noun and noun-noun combinations based on the automatic POS tagging. Proper nouns were allowed in conjunction with a preceding adjective or common noun. We did not extract pairs consisting of two proper nouns because they introduced too much noise. This approach results in the following POS patterns:

common noun followed by common noun:

[NN] [NN]

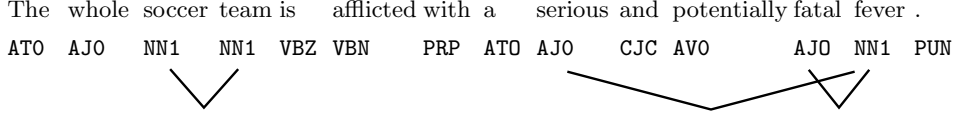


Figure 1: Three examples of noun-noun and adjective-noun pairs identified by the extraction pipeline.

Table 3: 2-by-2 contingency tables with observed and expected frequencies

	$W_1 = w_1$	$W_1 \neq w_1$		$W_1 = w_1$	$W_1 \neq w_1$
$W_2 = w_2$	$O_{11}$	$O_{12}$	$W_2 = w_2$	$E_{11} = \frac{R_1 C_1}{N}$	$E_{12} = \frac{R_1 C_2}{N}$
$W_2 \neq w_2$	$O_{21}$	$O_{22}$	$W_2 \neq w_2$	$E_{21} = \frac{R_2 C_1}{N}$	$E_{22} = \frac{R_2 C_2}{N}$

adjective or common noun followed by a proper noun:

([ADJ] ([COMMA | CONJ | ADJ | ADV]\* [ADJ])? | [NN]) [NPO]

adjective or proper noun followed by a common noun:

([ADJ] ([COMMA | CONJ | ADJ | ADV]\* [ADJ])? | [NPO]) [NN]

The noun-noun pattern for compounds is straightforward. The adjective-noun patterns are slightly more complex because they are designed to match adjacent as well as more distant adjective-noun modification. In addition, they allow proper nouns as modifiers to account for combinations such as *wellington boots*.

Figure 1 shows a noun-noun pair (*soccer team*) and two adjective-noun pairs (*serious fever*, *fatal fever*) that were identified by the extraction pipeline.

In order to access the corpus data efficiently, the BNC was indexed with the *IMS Open Corpus Workbench*<sup>4</sup>, preserving POS and lemma information. The *Corpus Query Processor* (CQP) was used to match POS patterns and extract word pairs. In the third step of the pipeline, contingency tables and association scores were computed with the help of the UCS toolkit<sup>5</sup> (Evert 2004). From the 112,102,325 tokens that were indexed, we extracted 2,014,116 pair tokens, which contained 391,454 different pair types. In order to remove noise, we applied a frequency filter of  $f \geq 3$  before the calculation of association scores.

### 3.2 Standard association measures

For each word pair  $(w_1, w_2)$ , co-occurrence frequency data from the BNC are collected in two *contingency tables*, as shown in Table 3. The left table contains the *observed frequencies*:  $O_{11}$  is the number of co-occurrences of  $w_1$  and  $w_2$  in the corpus,  $O_{21}$  is the total number of co-occurrences of  $w_1$  with a different word than  $w_2$ , etc. The right table contains *expected frequencies*  $E_{ij}$  under the assumption of statistical independence, which can be computed from the row sums  $R_i$  and column sums  $C_j$  of the observed frequencies. We use Evert’s (2004) notation and terminology.

Based on the observed and expected frequencies, a number of *association measures* (AMs) can be computed to quantify the strength of attraction between  $w_1$  and  $w_2$ . In this

Table 4: Standard association measures expressed as functions of observed and expected frequencies

association measure	formula
frequency	$f = O_{11}$
log-likelihood	$G^2 = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$
t-score	$t = \frac{O_{11} - E_{11}}{\sqrt{O_{11}}}$
chi-square	$X^2 = \frac{N( O_{11}O_{22} - O_{12}O_{21}  - N/2)^2}{R_1R_2C_1C_2}$

work, we focus on three well-known measures derived from statistical hypothesis tests: *log-likelihood* ( $G^2$ ), *t-score* ( $t$ ) and *chi-square* ( $\chi^2$ ) with Yates’ continuity correction applied (see e.g. Agresti 2002: Ch. 1; Manning and Schütze 1999: Ch. 5).

In addition to these statistical measures, we also consider plain co-occurrence *frequency* ( $f$ ), based on the assumption that a more frequently occurring word pair is more likely to be interesting. Previous work has shown that co-occurrence frequency performs surprisingly well in multiword and terminology extraction tasks (e.g. Daille 1996; Krenn and Evert 2001; Wermter and Hahn 2006). Table 4 lists definitions of the four measures in terms of observed and expected frequencies (Evert 2004: Ch. 3).

### 3.3 Rank measures

The standard AMs defined in Table 4 are symmetric in the sense that they do not capture the left-predictiveness or right-predictiveness that Kjellmer observed in many word combinations. All four measures are invariant under transposition of the contingency table, i.e. the association score remains the same if the rows and columns are exchanged. Michelbacher et al. (2007) introduced a *rank measure* based on the chi-square test to capture the asymmetry of paradigmatic associations.

In this article, we generalize the notion of a rank measure to arbitrary symmetric AMs and evaluate the ability of these rank measures to capture the asymmetry of syntagmatic associations.

In order to transform a standard symmetric AM into a rank measure that computes separate scores for the left- and right-predictiveness of a word pair, we implement the following procedure. For a left-to-right rank measure based on t-score ( $t$ ):

1. Compute symmetric association scores  $t$  for all word pairs  $(w_1, w_2)$ .
2. For each word  $w_1$ , create an *association list* of all components  $w_2$  that co-occur with  $w_1$  in the corpus and sort the list by association strength  $t$  in descending order.
3. Starting at the top, replace the association scores by ranks  $1, 2, 3, \dots$ <sup>6</sup>

Right-to-left rank measures are computed accordingly, exchanging  $w_1$  and  $w_2$  in the ranking procedure.

Table 5 shows the ten nouns ( $w_2$ ) that are most strongly associated with the adjective *rich* ( $w_1$ ), together with the association scores computed by the t-score measure ( $t$ ). We write the left-to-right rank measure based on  $t$  as  $R_t^\rightarrow(w_1, w_2)$  and call it the *forward rank* of  $(w_1, w_2)$ . Note that a *small* forward rank indicates a *high* degree of right-predictiveness.



Table 5: Determining the forward rank of *rich man*

$w_1$	$w_2$	$t$		$w_1$	$w_2$	$R_t^{\rightarrow}(w_1, w_2)$
<b>rich</b>	<b>man</b>	<b>16.563</b>		<b>rich</b>	<b>man</b>	<b>1</b>
rich	peasant	12.919		rich	peasant	2
rich	country	12.756		rich	country	3
rich	people	8.386		rich	people	4
rich	variety	7.423	$\rightarrow$	rich	variety	5
rich	source	7.861		rich	source	6
rich	color	7.568		rich	color	7
rich	soil	6.018		rich	soil	8
rich	nation	6.766		rich	nation	9
rich	world	5.714		rich	world	10

Table 6: Determining the backward rank of *rich man*

$w_1$	$w_2$	$t$		$w_1$	$w_2$	$R_t^{\leftarrow}(w_1, w_2)$
young	man	62.492		young	man	1
old	man	51.602		old	man	2
tall	man	19.270		tall	man	3
dead	man	18.661		dead	man	4
<b>rich</b>	<b>man</b>	<b>16.563</b>	$\rightarrow$	<b>rich</b>	<b>man</b>	<b>5</b>
poor	man	15.986		poor	man	6
white	man	14.279		white	man	7
married	man	14.620		married	man	8
gay	man	14.487		gay	man	9
big	man	14.456		big	man	10

For example,  $R_t^{\rightarrow}(\text{rich}, \text{man}) = 1$  means that *man* is the noun most strongly predicted by the adjective *rich* according to the t-score measure.

In a similar manner, we denote the *backward rank* of a word pair  $(w_1, w_2)$  according to  $t$  by  $R_t^{\leftarrow}(w_1, w_2)$ . As can be seen from Table 6, the backward rank of  $(\text{rich}, \text{man})$  is  $R_t^{\leftarrow}(\text{rich}, \text{man}) = 5$ . In this case, the forward rank (1) is lower than the backward rank (5), indicating higher right-predictiveness than left-predictiveness.

Note that the association score of the pair  $(\text{rich}, \text{man})$  is  $t = 16.536$  in both association lists. This score was computed from a single contingency table of observed frequencies, which is all the information that a standard AM has access to. By contrast, the corresponding left-to-right rank measure  $R_t^{\rightarrow}$  looks at the distribution of the association scores for all word pairs  $(w_1, \cdot)$ ; and the right-to-left measure  $R_t^{\leftarrow}$  looks at the distribution for all word pairs  $(\cdot, w_2)$ . In this way, different degrees of right- and left-predictiveness can be calculated.

Rank measures are a general and flexible tool for capturing asymmetry effects in word combinations. They can be applied to any symmetric AM and transform this AM into an asymmetric measure of right- and left-predictiveness. Each AM gives rise to a different asymmetric rank measure. Table 7 illustrates this point by showing left-to-right and right-to-left rank scores for the word pairs *heavy smoker* and *bonsai tree*, according to four different rank measures based on the standard AMs introduced in Section 3.2:

Table 7: Comparing rank measures based on frequency ( $f$ ), log-likelihood ( $G^2$ ), t-score ( $t$ ) and chi-square ( $\chi^2$ )

$w_1$	$w_2$	$R_f^{\rightarrow}$	$R_f^{\leftarrow}$	$R_{G^2}^{\rightarrow}$	$R_{G^2}^{\leftarrow}$	$R_t^{\rightarrow}$	$R_t^{\leftarrow}$	$R_{\chi^2}^{\rightarrow}$	$R_{\chi^2}^{\leftarrow}$
<i>heavy</i>	<i>smoker</i>	17	1	9	1	15	1	5	4
<i>bonsai</i>	<i>tree</i>	1	64	1	37	1	53	1	25

$R_f$  based on frequency  $f$   
 $R_{G^2}$  based on log-likelihood  $G^2$   
 $R_t$  based on t-score  $t$   
 $R_{\chi^2}$  based on the  $X^2$  test statistic

According to the first three rank measures, *heavy smoker* is a strongly left-predictive combination. The backward rank is 1 in all three cases whereas the forward rank is considerably higher. The rank measure based on  $\chi^2$  does not agree with the other measures, suggesting an almost symmetric pair with equal right- and left-predictiveness (although the backward rank is still slightly lower). The pair *bonsai tree* is strongly right-predictive according to all four measures. The forward and backward ranks are in accordance with the assessment of Kjellmer who used *bonsai tree* as an example for a clearly right-predictive combination.

The ranks do not take the frequency of the words into account and are therefore independent of the magnitude of association strength. For our purpose, this is not a problem. First, we want to examine asymmetry for each word pair individually without comparing ranks of different pairs. Second, in an elicitation experiment, low-frequency words will still trigger responses – and the best responses will receive low ranks. The ranks tell us how good the associations are *relative to the stimulus*.

In accordance with Michelbacher et al. (2007), we also measure right- and left-predictiveness with conditional probabilities.

$$P^{\rightarrow}(w_2|w_1) = \frac{P(w_1, w_2)}{P(w_1)} \quad P^{\leftarrow}(w_1|w_2) = \frac{P(w_1, w_2)}{P(w_2)}$$

We added arrows to emphasize right-predictiveness ( $P^{\rightarrow}$ ) and left-predictiveness ( $P^{\leftarrow}$ ). For example,  $P^{\leftarrow}(w_1|w_2)$  denotes the probability that  $w_1$  appears as the first component in a pair when  $[-w_2]$  is already given. The probabilities are maximum-likelihood estimates.

Note that because of

$$\frac{P(w_2, w_1)}{P(w_1)} = \frac{\frac{O_{11}}{N}}{\frac{O_{11}+O_{12}}{N}} = \frac{O_{11}}{O_{11} + O_{12}}$$

the rank measure based on conditional probabilities is identical to  $R_f$ . It is therefore not included separately in our evaluation.

### 3.4 Analysis of the distribution of ranks

Corpus-based measures of asymmetry are only interesting if such asymmetry is a frequent phenomenon. As we have argued earlier in this paper, we expect that syntagmatic associations are often asymmetric and can only be characterized adequately by a measure that allows for large differences in ranks. In order to explore this property of rank measures, we cross-tabulated the forward and backward ranks for the 391,454 word pairs with  $f \geq 3$  extracted from the BNC (see Section 3.1). Rank values were collected into logarithmically scaled bins (ranks 1–2, 3–5, 6–10, 11–20, 21–35, 36–60, 61–100, 101–160, 161–250, 251–500, 501+), such that all bins contain a similar number of items.

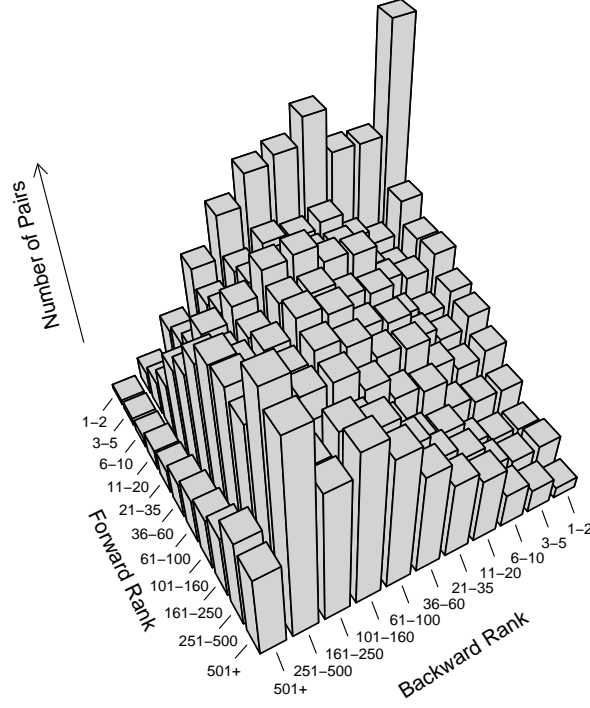


Figure 2: Cross-tabulation of forward and backward ranks for the log-likelihood measure.

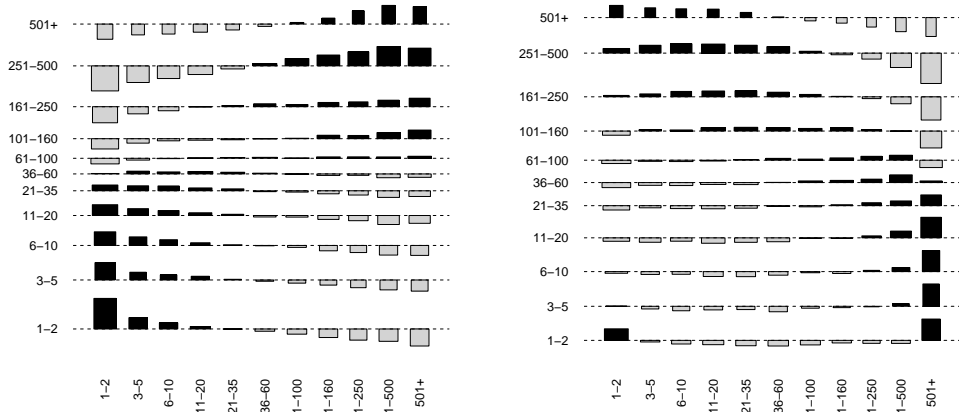


Figure 3: Association plots of forward and backward ranks for the log-likelihood-based rank measure  $R_{G^2}$  (left panel) and the frequency-based rank measure  $R_f$  (right panel).

Figure 2 shows a bar plot of the cross-tabulation of forward and backward ranks obtained for  $R_{G^2}$ , the rank measure based on log-likelihood. Bars along the main diagonal of the histogram – running from bottom to top in the printout – correspond to *symmetric* word pairs with nearly equal forward and backward ranks. The greater the distance of a bar from this main diagonal, the more asymmetric the corresponding word pairs are.

It is obvious that there is a considerable number of asymmetric word pairs with low forward and high backward rank (bars along the back left of the plot), and also vice versa (bars along the back right). On the other hand, very low forward ranks (ranks 1–2) correlate strongly with very low backward ranks, and very high forward ranks correlate with very high backward ranks (tall bars at both ends of the main diagonal). This is hardly surprising, since forward and backward ranks are based on the same symmetric association score: a highly associated word pair is more likely to achieve a low rank both in the “forward” and the “backward” list. Likewise, a word pair where  $w_1$  and  $w_2$  are close to statistical independence is more likely to be assigned high ranks in both lists.

Interestingly, the plot shows many word pairs with forward ranks 1 or 2, but much higher backward rank (roughly between 10 and 100, along the back left side of the histogram). According to  $R_{G^2}$ , these word pairs are strongly right-predictive. In comparison, the number of strongly left-predictive pairs is much smaller – there are no equally high bars along the back right side of the chart (corresponding to backward ranks of 1 or 2 and forward ranks between 10 and 100). This suggests that right-predictiveness is more common in English than left-predictiveness, at least for adjective-noun and noun-noun combinations. This observation is supported by our elicitation experiments, in which more word pairs were found to be right-predictive than left-predictive (see Section 5). The prevalence of right-predictive combinations is probably related to the fact that the preceding word is an important factor when deciding which word to produce next. This causal relationship along the time axis promotes the formation of right-predictive combinations. There exists no equally strong mechanism for producing left-predictive combinations.

The association plot in the left panel of Figure 3 shows more clearly to what extent forward and backward ranks are correlated. Black bars above the midlines indicate that a given combination of forward and backward rank appears for more word pairs than expected if the rankings were independent (i.e., a positive correlation between forward and backward rank). Grey bars below the lines indicate a smaller number of word pairs than expected (i.e., a negative correlation). It is obvious from the plot that very low forward ranks correlate strongly with very low backward ranks, and similarly for very high ranks. Again, this shows that very strongly and very weakly associated word pairs tend to be symmetric according to the rank measure. By contrast, the almost vanishing bars near the center of the plot show that forward and backward ranks are practically independent in a middle range (roughly ranks 10–100). Here, the rank measure is able to make a distinction between symmetric and asymmetric pairs.

A second important question is whether different symmetric AMs lead to different rank distributions. The right panel of Figure 3 shows an association plot for  $R_f$  (the rank measure based on co-occurrence frequency). The distribution of ranks is strikingly different from that of  $R_{G^2}$ , with forward and backward ranks almost independent for ranks below about 250. There is a considerable number of highly asymmetric word pairs, characterized by a very high rank (above 500) in one direction and a low rank (below 35) in the other direction (black bars along the top and right edges of the plot). This observation may be surprising at first, but it is easily explained for the left-predictive case by combinations of a high-frequency word  $w_2$  (e.g., *disease*) with a low-frequency word  $w_1$  (e.g., *adiposogenital*) that almost always occurs with  $w_2$  (and analogously for the right-predictive case).

Association plots for the other two measures are qualitatively similar to the log-likelihood ( $G^2$ ) pattern, with a somewhat stronger correlation for chi-square ( $\chi^2$ ) and a slightly larger region of near-independence for t-score ( $t$ ). This is perhaps not surprising since all three measures are based on statistical hypothesis tests. The observed differences

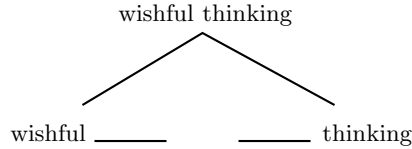


Figure 4: The word pair *wishful thinking* split into a forward and a backward stimulus

between the rank distributions agree with the known tendencies of  $\chi^2$  to overestimate and of  $t$  to underestimate the significance of association (Evert 2004: 111).

## 4 Elicitation experiment

Free association experiments have frequently been conducted to gather data about spontaneous human associations. In these experiments, a stimulus is presented and the subject is asked to produce one or more related words, e.g., those words that first come to mind when thinking about the stimulus.

In this type of experiment (described in Section 2.2), there are no restrictions on what type of response the subject can give (cf. Table 2). When humans associate freely, they produce mostly paradigmatic combinations. While there are some syntagmatic associations in the norms produced from such experiments – e.g., *blue*  $\rightarrow$  *sky* or *big*  $\rightarrow$  *deal* – they are always right-predictive, making these norms unsuitable for our purpose.

Nevertheless, as noted in Section 2.3, word association norms do contain a portion of syntagmatic responses. Furthermore, it has been shown that grammatical stimulus-response pairs can be collected systematically in elicitation experiments when subjects are explicitly asked to produce them (McGee 2009). With these findings in mind, we decided to base our experimental design on classical free associations experiments but with a restriction to syntagmatic responses.

We instructed subjects to produce responses that result in a well-formed phrase when combined with the stimulus. The key problem is how to present stimuli in a way that elicits the desired data without biasing the subjects’ responses.

The experimental design we decided on splits each word pair  $(w_1, w_2)$  into two separate stimuli: a *forward stimulus*  $[w_1 \_]$  and a *backward stimulus*  $[\_ w_2]$ . That is, either the first or the second component of the pair is replaced by the blank  $\_$  to indicate to participants that a word has been removed and needs to be provided. This design allows for testing both directions of association, from  $w_1$  to  $w_2$  and from  $w_2$  to  $w_1$ . An example is shown in Figure 4.

Subjects were instructed to fill the blank in a way that created a well-formed phrase. We imposed no other restrictions on admissible responses to avoid any type of bias. In particular, no context was provided that might have disambiguated ambiguous stimuli or suggested a response from a particular domain.

Because of this unrestricted nature of the experiment, subjects often produced part-of-speech combinations that were not compatible with the data extracted from the BNC. Such unusable responses included determiners, pronouns and cases where subjects interpreted a stimulus word as a verb or adverb rather than as an adjective or noun. For example,  $[cut \_]$  was often extended to *cut down* or *cut off* instead of a noun-noun or adjective-noun phrase such as *cut glass*. We discarded part-of-speech mismatches in order to be able to perform a clean analysis of adjective-noun and noun-noun associations.

### 4.1 Pair selection

We used a hybrid selection method to sample stimuli, adapting the methodology of Krenn and Evert (2005). We started with a pool  $P$  of candidates and took a random sample  $M$

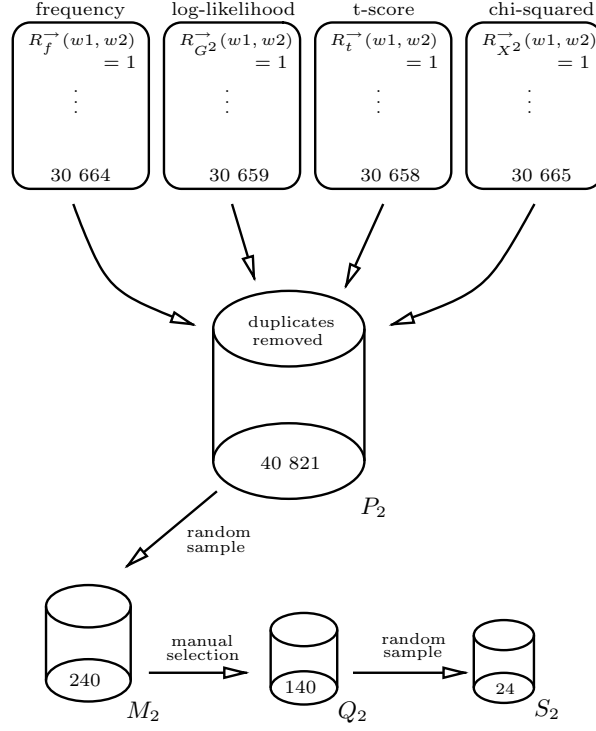


Figure 5: Sampling procedure for stimuli (illustrated for a sample  $S_2$  of right-predictive word pairs from the pool  $P_2$ )

from  $P$ . We then created a subset  $Q$  from  $M$  by removing extraction noise and technical terms. Finally, we took a further random sample  $S$  from  $Q$  to get the desired number of stimuli.

This procedure was applied to three different pools:

- $P_1$ : all pair types  $(w_1, w_2)$
- $P_2$ : pair types with strong right-predictiveness (according to at least one of the association measures)
- $P_3$ : pair types with strong left-predictiveness (according to at least one of the association measures)

The process is illustrated in Figure 5 for strongly right-predictive word pairs (i.e., candidate sets  $P_2/M_2/Q_2/S_2$ ).

The first pool,  $P_1$ , contains all 2,014,116 pair types that we extracted from the BNC with the procedure described in Section 3.1.<sup>7</sup> The pools  $P_2$  and  $P_3$  are motivated by two constraints that any experiment designed to elicit syntagmatic responses must satisfy. First, we can only present a limited number of stimuli to each subject. This means the overall number of stimuli must be relatively small. Second, we must ensure that the elicited data are useful for our evaluation. Since a random sample would mostly contain weakly associated pairs, it was necessary to bias the selection of stimuli.

To this end, we created a set  $P_2$  of strongly right-predictive pairs and a set  $P_3$  of strongly left-predictive pairs, where strong predictiveness was defined as  $R^\rightarrow(w_1, w_2) = 1$  (for  $P_2$ ) and  $R^\leftarrow(w_1, w_2) = 1$  (for  $P_3$ ). We chose this criterion to obtain good candidates for asymmetric word combinations.

To create  $P_2$ , the rank criterion  $R^\rightarrow(w_1, w_2) = 1$  was applied to all pairs and for all four association measures. The four resulting sets (represented by rectangles in Figure 5) were merged. Multiple occurrences of the same pair were removed, resulting in a pool

$P_2$  of 40,821 candidates. The same procedure was carried out to obtain  $P_3$  with a size of 26,600 candidates.

In the next step, random samples of fixed sizes were drawn from each of the  $P_i$ . The three resulting samples were  $Q_1$  (336 pairs),  $Q_2$  (240 pairs) and  $Q_3$  (240 pairs). The sample sizes were chosen to be low enough to allow for manual review of all pairs. Pairs with frequency  $f \leq 5$  were removed. We then reviewed each of the remaining pairs and removed extraction noise, rare technical terms and rare proper nouns. Specifically, technical terms from specialized fields like mathematics, biology, computing, and medicine were removed. Examples include *ileocolonic resection*, *configurational entropy*, *Unix file* and *non-zero element* – terms which are unlikely to be familiar to the general population.

The resulting pools  $Q_1$ ,  $Q_2$  and  $Q_3$  were the basis for three final random samples:  $S_1$  (16 pairs from  $Q_1$ ),  $S_2$  (24 pairs from  $Q_2$ ) and  $S_3$  (24 pairs from  $Q_3$ ). These 64 pairs were then replaced by their most frequent surface realizations in the BNC, in order to ensure that subjects would not be distracted by the use of uncommon base forms from the automatic lemmatization. For example, *wellington boot* was turned into *wellington boots* and *christmas decoration* into *Christmas decorations*.

## 4.2 Conducting the experiment

Test subjects were randomly split into two groups, group I and group II. When group I was presented a pair with the first component missing, group II saw the same pair with the second component missing and vice versa. This procedure ensured that subjects were not biased by a previous stimulus (e.g., seeing  $[- tree]$  after  $[bonsai -]$ ). Stimuli of types  $[w_1 -]$  and  $[- w_2]$  were split equally between the two groups.

Subjects were given detailed instructions to ensure they would not mistake the experiment for a free association task. They were encouraged to take some time to think of the stimulus word in different contexts and scenarios. They were also permitted to give multiple answers, or no answer at all.

The experiment was carried out online at the Portal for Psychological Experiments on Language.<sup>8</sup> The subjects were informed that only native speakers of English were allowed to participate. The full instructions as well as a complete list of stimuli and responses are available in the Web appendix<sup>9</sup>. 168 subjects took part in the experiment, 74 for group I and 94 for group II. The discrepancy between the two groups is due to the fact that some subjects did not complete the experiment. We collected a total of 43,101 responses. This means that on average, a subject provided 4 responses per stimulus.

We only included data from completed experiments in our analysis. We removed 3 pairs – *common destiny*, *independent charts*, and *old self* – because they were never successfully elicited in either direction. For example, *common* was never elicited as  $w_1$  for the stimulus  $[- destiny]$  and *destiny* was never elicited as  $w_2$  for the stimulus  $[common -]$ . The analysis described in the next section was performed for the remaining 61 pairs.

We lemmatized the subjects' input for our analysis. Spelling variants were unified to British English to facilitate the comparison with the corpus data. Manual spelling correction and normalization was applied when necessary, e.g., *Xmas* was normalized to *Christmas*.

For each subject and stimulus, we kept track of the order in which responses were given. We assume that the order of elicitation directly corresponds to association strength in that the first answer given has the highest association to the stimulus word and so on.

## 5 Experimental results and analysis

In this section, we define the direction scores used to evaluate subjects' responses and perform both a qualitative and a quantitative evaluation of the experimental results.

Table 8: Comparison of our results with free association norms

stimulus <i>No.</i>	syntagmatic [ <i>white</i> _]	paradigmatic <i>white</i>		syntagmatic [ <i>water</i> _]	paradigmatic <i>water</i>	
		EAT	USF		EAT	USF
1	wash	black	black	bottle	wet	drink
2	Christmas	red	pure	works	drink	cool
3	house	snow	clean	fall	tap	wet
4	out	sheet	snow	fountain	sea	swim
5	board	ice	light	slide	cold	thirsty
7	wedding	beach	color	cooler	h2o	faucet
6	water	nothing	paper	jug	hot	pool
8	dress	blank	red	pipe	rain	thirst
9	man	block	—	park	river	ice
10	noise	blue	—	balloon	thirst	cold

### 5.1 Direction scores

We scored the subjects’ responses using a mean reciprocal rank measure (cf. Voorhees 1999). Two *direction scores* were defined – a *forward score*  $f(w_1, w_2)$  and a *backward score*  $b(w_1, w_2)$ , as given by the following equations:

$$f(w_1, w_2) = \frac{1}{C([w_1 \_])} \sum_{i=1}^{C([w_1 \_])} \frac{1}{r_i(w_2)}$$

$$b(w_1, w_2) = \frac{1}{C([\_ w_2])} \sum_{i=1}^{C([\_ w_2])} \frac{1}{r_i(w_1)}$$

Here,  $C([w_1 \_])$  is the total number of subjects that were presented with stimulus  $[w_1 \_]$  and  $r_i(w_2)$  is the rank of  $w_2$  in the list of responses to  $[w_1 \_]$  given by subject  $i$ ;  $C([\_ w_2])$  is the number of subjects presented with stimulus  $[\_ w_2]$  and  $r_i(w_1)$  the rank of  $w_1$  in the list of responses to  $[\_ w_2]$  by subject  $i$ . If a subject did not produce the response in question, we assigned rank  $r = 1000$ . The highest possible direction score in this scheme is 1.0.

### 5.2 Qualitative evaluation

The composition of responses given in our study differs considerably from previous association norms. As an example, Table 8 shows the 10 highest-scoring responses for *white* and *water* in our syntagmatic experiment and in two free association experiments.<sup>10</sup> For this comparison we did not filter out responses like *white out* that do not constitute adjective-noun or noun-noun combinations. Most responses in the free association experiments are of a paradigmatic nature except for *sheet*, *beach*, *block*, *snow* and *light* for the stimulus *white*.

The responses in the new experiment, however, exclusively consist of syntagmatic associations, that is, they all produce well-formed phrases when the response is inserted into the empty slot of the stimulus.

A qualitative analysis of the 61 pairs revealed four major groups. Group A contains all pairs where the rank measures conform with human responses in that they agree on which direction of association is stronger. The bulk of the pairs (48) belong to group A. Group B is a small group consisting of 4 cases where corpus data and human elicitation contradict each other. We also found borderline cases where the rank measures provide evidence for both right-predictiveness and left-predictiveness, but could not be aligned



with human elicitations (group C, 4 elements). We regard pairs where the rank measures suggest very strong association (rank  $\leq 2$ ) in both directions as a special case and put these pairs in a separate group of high mutual predictiveness (group D, 5 elements).

Table 9 shows word pairs from the four groups with detailed information on corpus ranks and scores from the elicitation experiment. For reasons of space, we only give a subset of the 48 pairs in group A.

For most pairs in groups A and B, the four measures agree on the direction of predictiveness. There were 11 pairs where the four measures did not agree, marked with ‘\*’ in Table 9. In most cases, it is the  $\chi^2$  measure that disagrees with the other measures.

For about 80% of the pairs – those in Group A – the statistical measures indicate the correct direction of association. This demonstrates that the rank measures are able to model human behavior in the elicitation experiment for most pairs. Group B, which contains pairs where the rank measures failed to make correct predictions, is reassuringly small with only 4 pairs. There are various possible explanations for the failure of the rank measures in these cases. For example, the pair *missile silos* exhibits almost equally strong predictiveness in both directions according to the subjects. This view is only partly reflected by the rank measures. The  $\chi^2$  measure comes close with a low forward and backward rank. However, the other measures only have a low backward rank, but not a low forward rank for this pair. They rank other words (e.g., *crisis*, *launcher* or *technology*) more highly. This discrepancy between human judgments and corpus data could be due to the fact that missiles were a dominant topic during the cold war – at the time when the BNC data were collected – and that subjects today are less familiar with them.

Group D contains word pairs where the rank measures indicate the strongest possible predictiveness in both directions, regardless of which status the human data suggest. The five phrases in this category are collocations – they are fixed, recurring expressions rather than free combinations.

A detailed study of the five pairs in D and the four pairs in B is beyond the scope of this article, but we suspect that corpus data fail to provide a good prediction of human behavior in these cases because of differences between spoken and written language. For example, *wishful thinking* is not very left-predictive according to the human subjects – subjects gave responses like *quick* ( $b = 0.1754$ ), *good* (0.1651), *clear* (0.1148) and *critical* (0.1082) more often than *wishful* (0.0068). But in the BNC, *wishful* is by far the most common adjective preceding *thinking* (147 instances vs. 65 for *new thinking* and 43 for *critical thinking*). Other reasons for the discrepancy between corpus-derived and human associations for groups B and D could be part-of-speech ambiguity (*thinking* is predominantly a present participle, not a noun) and dialectal differences – *bloody hell* and *unleaded petrol* are British English expressions that American English speakers may not be familiar with.

One important difference between the rank measures and raw conditional probabilities can be found in all pairs of group D except for *bloody hell*. We will illustrate the phenomenon for *wishful thinking*. Human judgement for this pair is overwhelmingly right-predictive ( $f = .9521$ ,  $b = .0068$ ). The word *wishful* only occurs with two different nouns in the corpus and almost all its occurrences are with *thinking* which in turn occurs with about 100 other adjectives. This is naturally reflected in the conditional probabilities:  $P^{\rightarrow}(\text{thinking}|\text{wishful}) = .924$  and  $P^{\leftarrow}(\text{wishful}|\text{thinking}) = .0899$ . However, the association score of the two words in the combination *wishful thinking* is high enough to outrank all other adjectives that appear with *thinking* resulting in rank 1 in both directions. This can simply be interpreted as the rank measures suggesting likely completions to a stimulus (based on the distribution in the corpus) whereas conditional probabilities are suited to measure absolute association strength. The other two pairs where conditional probabilities perform better than the rank measures are *South East* and *laboratory experiments* from group C. Here, the rank measures are ambivalent but the conditional probabilities capture the correct direction of predictiveness.

For the pair *aching void*, conditional probability makes the wrong prediction and the

rank measures are correct. Conditional probability suggests near-symmetry ( $P^{\rightarrow} = .0389$ ,  $P^{\leftarrow} = .0372$ ) whereas the ranks (except for  $R_{\chi^2}$ ) conform with the subjects' judgement of left-predictiveness.

### 5.3 Quantitative evaluation

We have introduced three different approaches to predictiveness and asymmetric association: (i) direction scores  $f$  and  $b$  computed from the elicitation experiment; (ii) rank measures  $R^{\rightarrow}$  and  $R^{\leftarrow}$  and (iii) conditional probabilities  $P^{\rightarrow}$  and  $P^{\leftarrow}$ . The latter two are based on corpus data. Scores, ranks and conditional probabilities are capable of capturing asymmetries between the two components of a word pair. In this section, we perform a quantitative evaluation of how well the predictions made by the corpus-based measures agree with the human scores.

Our test case is the distinction between right-predictive and left-predictive word pairs. We coded the human scores as a binary *response* variable  $Y$ :  $Y = 1$  indicates a right-predictive (RP) and  $Y = 0$  a left-predictive (LP) pair:<sup>11</sup>

$$Y = \begin{cases} 1 & \text{if } f \geq b \text{ (right-predictive)} \\ 0 & \text{if } f < b \text{ (left-predictive)} \end{cases}$$

Analogously, we transformed the conditional probabilities into a corresponding *predictor* variable:<sup>12</sup>

$$X = \begin{cases} 1 & \text{if } P^{\rightarrow} \geq P^{\leftarrow} \text{ (right-predictive)} \\ 0 & \text{if } P^{\rightarrow} < P^{\leftarrow} \text{ (left-predictive)} \end{cases}$$

We applied a similar procedure to each corpus-based rank measure. For instance, the predictor variable for the t-score measure  $t$  is given by:

$$X_t = \begin{cases} 1 & \text{if } R_t^{\rightarrow} \leq R_t^{\leftarrow} \text{ (right-predictive)} \\ 0 & \text{if } R_t^{\rightarrow} > R_t^{\leftarrow} \text{ (left-predictive)} \end{cases}$$

Recall that a lower rank indicates higher association. Therefore, a pair with  $R^{\rightarrow} > R^{\leftarrow}$  is *left*-predictive and is assigned a predictor value of  $X = 0$ . In the case of equal ranks, we also assigned  $X = 1$  (right-predictive) because the human subjects – as well as our corpus data, see Section 3.4 – showed a preference for right-predictiveness – the elicitation experiment yielded 34 word pairs with  $f > b$ , compared to 27 with  $f < b$ .

The first four rows of Table 10 show the accuracy of predictions made by the four rank measures. In order to combine information from all measures, we trained a linear model on the individual predictors, using a 6-fold cross-validation scheme (5 folds with 10 items each, and one fold with 11 items).

The data set contains 34 RP and 27 LP pairs. Therefore, a baseline classifier that assigns every pair to category RP (i.e.,  $X = 1$ ) achieves an accuracy of 55.7%. In our evaluation, we use a more optimistic cross-validation baseline where the most frequent category is chosen separately for each of the six data folds.<sup>13</sup> The resulting baseline accuracy of 62.3%, calculated over all 61 items, is reported in the last row of Table 10.

Because of the small sample size used for the evaluation, statistical significance testing is essential. As an indication of the amount of random variation, we calculated binomial 95% confidence intervals for the proportion of correct predictions, shown in the rightmost column of Table 10. For the combined model, this means that we ignore the additional random variation caused by the different training sets used in the cross-validation procedure. We feel that this approach is justified since there is considerable overlap between the training sets used in different steps of the cross-validation (viz., any two training sets share 4 of their 5 data folds). If our assumptions are tenable, then the evaluation results for individual folds can be treated as random samples of size 10 (or 11 for the last fold)

Table 9: Forward and backward scores and rank measures for a subset of the word pairs used in the elicitation experiment; '\*' indicates disagreement of rank-measures on direction of predictiveness.

$f$	$b$	$(w_1, w_2)$	$R_f^{\rightarrow}$	$R_f^{\leftarrow}$	$R_{G^2}^{\rightarrow}$	$R_{G^2}^{\leftarrow}$	$R_t^{\rightarrow}$	$R_t^{\leftarrow}$	$R_{\chi^2}^{\rightarrow}$	$R_{\chi^2}^{\leftarrow}$
<i>group A: rank measures and direction scores conform</i>										
.5891	.2545	Academy Award	1	9	1	2	1	7	1	2
.3328	.0010	ancestral home	1	25	1	13	1	19	1	11
.5551	.1609	cable television	2	7	1	4	2	5	1	2
.0127	.0087	cut glass	1	75	1	46	1	58	1	40
.6760	.0010	felled tree	1	54	1	33	1	45	1	17
.0683	.0021	hunched shoulders	1	16	1	7	1	14	1	2
.0875	.0010	old-fashioned way	1	98	1	60	1	62	4	70
.1667	.0063	rightful place	1	26	1	6	1	15	2	4
.1500	.0496	rope ladder	1	4	1	4	1	4	2	4
.0241	.0010	shrewd idea	3	109	6	49	3	68	10	41
.1719	.0010	thick-set man	1	519	1	169	1	318	1	86
.0641	.0068	well-worn path	1	71	1	34	1	58	1	22
.0127	.0125	*impending retirement	9	18	8	14	9	18	9	9
.0606	.0563	*speech recognition	1	2	1	1	1	2	2	4
.0010	.0099	annual rent	29	2	20	1	28	2	15	7
.0266	.8208	Christmas decorations	11	1	8	1	11	1	9	3
.0010	.0101	female preferences	63	34	92	44	60	34	95	48
.0010	.0312	legal wrangling	151	1	58	1	110	1	25	1
.0081	.1325	smoked mackerel	5	1	3	1	5	1	3	1
.0010	.0031	southern bypass	21	1	15	1	20	1	10	1
.0046	.0426	welcome diversion	17	3	15	1	16	3	8	3
.0032	.0425	*bond issuance	10	1	7	1	10	1	2	2
.0549	.4500	*deadly nightshade	7	1	3	1	7	1	1	1
.0068	.0100	*aching void	5	3	3	1	5	3	1	2
.0478	.3893	*white collar	11	1	6	1	9	1	5	6
<i>group B: rank measures and direction scores do not conform</i>										
.0955	.1160	healthy food	6	19	6	20	5	15	13	53
.1562	.1543	missile silos	16	1	8	1	16	1	2	1
.0010	.0063	seasoned campaigners	1	9	1	6	1	9	1	6
.2922	.5301	*precious metals	1	2	1	2	1	2	1	1
<i>group C: rank measures ambivalent</i>										
.5411	.4620	*epileptic seizure	2	3	2	1	2	3	2	1
.0761	.0335	*dedicated follower	7	3	2	3	4	3	3	6
.4340	.0237	*laboratory experiments	2	1	1	1	2	1	1	1
.0683	.1836	*South East	1	2	3	2	1	2	5	2
<i>group D: high mutual predictiveness</i>										
.2962	.1337	bloody hell	1	1	1	1	1	1	1	1
.1275	.2833	*special needs	1	1	1	1	1	1	1	2
.6810	.2793	toxic waste	1	1	1	1	1	1	1	1
.2613	.1583	unleaded petrol	1	1	1	1	1	1	1	1
.9521	.0068	wishful thinking	1	1	1	1	1	1	1	1

Table 10: Accuracy of predictions made by the corpus measures

rank measure	correct predictions	95% confidence interval
$R_f$	88.5%	77.8% ... 95.3%
$R_{G^2}$	<b>90.2%</b>	79.8% ... 96.3%
$R_t$	88.5%	77.8% ... 95.3%
$R_{\chi^2}$	82.0%	70.0% ... 90.6%
combined	83.6%	71.9% ... 91.8%
cond. prob.	<b>90.2%</b>	79.8% ... 96.3%
<i>baseline</i>	62.3%	49.0% ... 74.4%

from the same population. Additional support for our approach is provided by the observation that the empirical standard deviation across the 6 data folds is smaller than the theoretical standard deviation for binomial samples of the same size.

All rank measures perform well, even compared to the optimistic baseline. The best result is achieved by log-likelihood ( $G^2$ ) with an accuracy of 90.2%. The binomial confidence interval indicates that the  $G^2$  rank measure will achieve a prediction accuracy of at least 79.8% on larger data sets. Frequency ( $f$ ) and t-score ( $t$ ) are tied in second place, with a score of 88.5%. This is no coincidence: the two measures happen to make identical predictions for all items in our data set (i.e.,  $X_f = X_t$ ), although they are not equivalent in general.<sup>14</sup> Chi-square ( $\chi^2$ ) performs considerably worse than the other rank measures, but is still much better than the baseline, with a 95% confidence interval ranging from 70% to about 90% accuracy. Surprisingly, the combined model does not improve on individual rank measures and is only slightly better than  $\chi^2$  with an accuracy of 83.6%. Conditional probabilities perform as well as the best rank measure (but with different predictions).

We used an exact version of McNemar’s test (Hollander and Wolfe 1999: 468–470) to assess the significance of result differences. This test considers only items for which the two models to be compared made different predictions. Due to the small sample size, there are no significant differences between any of the models. In particular, we were not able to show that  $G^2$  is significantly better than  $\chi^2$  (exact McNemar,  $p = .063$ ), even though Table 10 shows a clear difference. However, all models except for  $\chi^2$  are significantly better than the optimistic baseline (with p-values ranging from  $p = .001$  for  $G^2$  to  $p = .029$  for the combined model).

Unexpectedly, combining the rank information of all measures did not lead to an improvement over the best single measure. This can be interpreted as a sign of overtraining, although the difference may well be due to chance (McNemar’s test yields  $p = .219$  for  $G^2$  against the combined model). A simple ranking by co-occurrence frequency ( $f$ ) once again performs astonishingly well, reaching the same accuracy as t-score ( $t$ ). Both measures only take the first cell of the contingency table into account, but t-score additionally considers the difference between observed and expected frequencies. It is interesting to note that  $G^2$  is the best of the five models and  $\chi^2$  the worst, even though they are both independence tests using information from the full contingency table. A possible explanation is the tendency of  $\chi^2$  to overestimate significance in highly skewed contingency tables (see Dunning 1993; Evert 2004).

## 5.4 Applications of asymmetric measures

Query expansion, a popular application of association measures in natural language processing, is an asymmetric task. It is appropriate to rewrite the query *fruit* as *fruit OR mango* since documents about mangos are necessarily about fruit, but it is not appropriate to rewrite the query *mango* as *mango OR fruit*. Clearly, corpus-based measures of association are only useful in this context if they take such asymmetry into account.

Being able to measure asymmetry in similarity has potential benefits for other applications, for example anaphora resolution (e.g. Mitkov et al. 2001). Think of a sentence from a recipe book: *Take a large apple<sub>i</sub> and cut the fruit<sub>i</sub> into four pieces.* Here, with knowledge about asymmetric substitutability, it would be possible to figure out that *fruit* refers back to *apple*. On the other hand, it would be wrong to co-index *apple* and *fruit* in the following example: *Take a large fruit<sub>i</sub> and cut the apple<sub>i</sub> into four pieces.* Asymmetric association measures could be used to predict the felicity of substitution.

In general, asymmetric measures of similarity are an important factor in all NLP tasks that benefit from better treatment of mutual substitutability, for example reducing data sparseness in language models (Dagan et al. 1999) or the automatic acquisition of selectional preferences (Resnik 1996).

## 6 Conclusion

In work on semantic relatedness and free association in computational linguistics and natural language processing, different types of relations between words are often lumped together. We have discussed two important distinctions in this article, the distinction between syntagmatic and paradigmatic relations and the distinction between symmetric and asymmetric relations. Asymmetry in paradigmatic relations (e.g. asymmetric similarity) has received attention in the past in psychological and corpus-based studies and it has been shown that asymmetric similarity measures can be of use for a number of applications.

Previous research was often based on free association norms which capture mostly paradigmatic relations. In this article, we have investigated asymmetry in syntagmatic relations. We designed a novel experiment setup to collect human data on syntagmatic combinations. In our study, we compared syntagmatic combinations in corpora and in human-subject experiments and demonstrated that corpus-derived rank measures and conditional probabilities can predict the asymmetry of human syntagmatic associations with high accuracy. We found that conditional probabilities are suited to measure absolute association strength whereas rank measures are a good indicator for which responses could be the best completion to a given stimulus. We showed that a large proportion of combinations are asymmetric and that right-predictive asymmetry is more prevalent than left-predictive asymmetry.

We view our contribution as a first step towards richer models of how corpus data can be used to predict human lexical knowledge and as a basis for defining appropriate types of relatedness between words for the needs of different NLP applications.

In the field of corpus linguistics, asymmetry is an important property of collocations (Kjellmer 1991; Sinclair 1991) but has long been neglected due to a lack of appropriate techniques for corpus data. Our rank-based asymmetric association measures provide, for the first time, a suitable empirical operationalisation of asymmetric collocations. In addition, future theoretical discussions can draw on the results of our syntagmatic association experiment as a complementary form of evidence.

All rank measures included in our study are based on association measures derived from statistical significance tests, which are known to correlate strongly with co-occurrence frequency. Rankings obtained from measures of effect size such as Mutual Information, on the other hand, may provide entirely new perspectives on the right- and left-predictiveness of syntagmatic combinations. We plan to extend our analysis to a range of well-known effect-size measures. As we have pointed out in Section 3.3, however, rankings obtained from conditional probabilities are identical to the frequency ranks  $R_f^{\rightarrow}$  and  $R_f^{\leftarrow}$ . This observation suggests that association measures based on conditional probabilities – including the Dice coefficient and minimum sensitivity (Pedersen and Bruce 1996) – will also lead to a strong correlation with frequency ranking (and hence with the significance measures in our study).

## Bionotes

Lukas Michelbacher received his MSc in Computational Linguistics (with minor in Mathematics) from University of Stuttgart, Germany, in 2008 where he is currently working towards his doctoral dissertation on a graph-based approach to the extraction of nominal multi-word units. His research interests include the acquisition of lexical phenomena from text corpora, multi-word units and bilingual lexicon extraction.

Stefan Evert received his PhD in Computational Linguistics from the University of Stuttgart in 2004. He is currently assistant professor (*Juniorprofessor*) for Computational Linguistics at the Institute of Cognitive Science, University of Osnabrück, Germany. His research interests center around the statistical analysis of corpus frequencies and other quantitative linguistic data. Specific topics he is currently working on include collocations and their automatic identification in corpus data, distributional semantics, non-randomness in corpus frequency data, as well as word frequency distributions and Zipf's law.

Hinrich Schütze received his PhD in Computational Linguistics from Stanford University in 1995. He is a professor at the Institute for Natural Language Processing at the University of Stuttgart and heads the Statistical Natural Language Processing group, which conducts research on statistical models of language and applications of StatNLP like information retrieval and machine translation.

## Notes

<sup>1</sup>previously used by Kent and Rosanoff (1910)

<sup>2</sup>Since many pairs fall into several categories, the total percentage exceeds 100%.

<sup>3</sup><http://www.natcorp.ox.ac.uk/>

<sup>4</sup><http://cwb.sourceforge.net/>

<sup>5</sup><http://www.collocations.de/software.html>

<sup>6</sup>Ties are handled as in a typical “sports” ranking: if  $n$  consecutive items have the same score, they are all assigned the lowest free rank  $r$ ; the next item will be assigned rank  $r + n$ .

<sup>7</sup>Note that no frequency threshold is applied at this stage, resulting in a very large number of pair types.

<sup>8</sup><http://language-experiments.org/>.

<sup>9</sup><http://www.ims.uni-stuttgart.de/~michells/asymmetry>

<sup>10</sup>The USF data set lists only eight responses for the stimulus *white*.

<sup>11</sup>The case  $f = b$  did not occur in the human data.

<sup>12</sup>Again, the case  $P^+ = P^-$  did not occur.

<sup>13</sup>The baseline is optimistic because the most frequent category is determined from the test fold in each case, rather than from the training folds. For instance, if the first fold contained 7 RP pairs and 3 LP pairs, the optimistic baseline classifier would assign all pairs in this fold to category RP. If the third fold contained 4 RP and 6 LP pairs, the optimistic baseline would assign all pairs in this fold to category LP.

<sup>14</sup>Note that the difference between these measures and log-likelihood corresponds to a single word pair:  $G^2$  makes 55 correct predictions vs. 54 for  $f$  and  $t$ . Our experiment therefore provides no reliable evidence that any of the three measures is better than the other two.

## References

- Agresti, Alan. 2002. *Categorical Data Analysis*. Hoboken: Wiley. 2nd edn.
- Budanitsky, Alexander and Graeme Hirst. 2006. Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics* 32(1). 13–47.
- Church, Kenneth W. and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29.
- Clark, Herbert H. 1971. Word associations and linguistic theory. In John Lyons (ed.), *New Horizon in Linguistics*. 271–286. London: Penguin.
- Dagan, Ido, Lillian Lee and Fernando Pereira. 1999. Similarity-based models of word cooccurrence probabilities. *Machine Learning* 34(1). 43–69.

- Daille, Béatrice. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Judith L. Klavans and Philip Resnik (eds.), *The Balancing Act*. 49–66. Cambridge, MA: MIT Press.
- Dennis, Simon. 2004. An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences of the United States of America* 101(Suppl. 1). 5206–5213.
- Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics* 19(1). 61–74.
- Evert, Stefan. 2004. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. thesis. Institut für maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.
- Evert, Stefan and Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *10th Conference of The European Chapter of the Association for Computational Linguistics, EACL*. 83–86.
- Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems* 20(1). 116–131.
- Gabrilovich, Evgeniy and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of The Twentieth International Joint Conference for Artificial Intelligence, IJCAI*. 1606–1611.
- Gries, Stefan Th., Beate Hampe and Doris Schönefeld. 2005. Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. *Cognitive Linguistics* 16(4). 635–676.
- Griffiths, Thomas L., Mark Steyvers and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological Review* 114(2). 211–244.
- Gurevych, Iryna. 2005. Using the structure of a conceptual network in computing semantic relatedness. In *2nd International Joint Conference on Natural Language Processing, IJCNLP*. 767–778.
- Hollander, Myles and Douglas A. Wolfe. 1999. *Nonparametric Statistical Methods*. Hoboken: Wiley. 2nd edn.
- Hutchison, Keith A. 2003. Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin and Review* 10(4). 785–813.
- Jenkins, James J. 1970. The 1952 Minnesota word association norms. In Leo Postman and Geoffrey Keppel (eds.), *Norms of word association*. 1–38. New York: Academic Press.
- Jones, Michael N. and Douglas J.K. Mewhort. 2007. Representing word meaning and order information in a composite holographic lexicon. *Psychological review* 114(1). 1–37.
- Keller, Frank and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics* 29(3). 459–484.
- Kent, Grace H. and Aaron J. Rosanoff. 1910. A study of association in insanity. *American Journal of Insanity* 67(1). 317–390.
- Kiss, George R., Christine Armstrong, Robert Milroy and James Piper. 1973. An associative thesaurus of English and its computer analysis. In Adam J. Aitken, Richard W. Bailey and Neil Hamilton-Smith (eds.), *The Computer and Literary studies*. Edinburgh: University Press.
- Kjellmer, Göran. 1991. A mint of phrases. In Karin Aijmer and Bengt Altenberg (eds.), *English Corpus Linguistics*. London: Longman.
- Krenn, Brigitte and Stefan Evert. 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. In *ACL Workshop on Collocations*. 39–46.
- Krenn, Brigitte and Stefan Evert. 2005. Separating the wheat from the chaff: Corpus-driven evaluation of statistical association measures for collocation extraction. In Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder and Petra Wagner (eds.), *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen. Beiträge zur GLDV-Tagung 2005 in Bonn*. Vol. 8 of *Computer Studies in Language and Speech*. 104–117. Frankfurt am Main: Lang, Peter.

- Kullback, Solomon and Richard A. Leibler. 1951. On information and sufficiency. *The Annals of Mathematical Statistics* 22(1). 79–86.
- Lapata, Maria, Frank Keller and Scott McDonald. 2001. Evaluating smoothing algorithms against plausibility judgements. In *39th Annual Meeting of the Association for Computational Linguistics, ACL*. 346–353.
- Lee, Lillian. 1999. Measures of distributional similarity. In *37th Annual Meeting of the Association for Computational Linguistics, ACL*. 25–32.
- Lee, Lillian. 2001. On the effectiveness of the skew divergence for statistical language analysis. In *Artificial Intelligence and Statistics*. 65–72.
- Leech, Geoffrey, Roger Garside and Michael Bryant. 1994. CLAWS4: The tagging of the British National Corpus. In *15th International Conference on Computational Linguistics, COLING*. 622–628.
- Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.
- McGee, Iain. 2009. Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. *Corpus Linguistics and Linguistic Theory* 5(1). 79–103.
- Michelbacher, Lukas, Stefan Evert and Hinrich Schütze. 2007. Asymmetric association measures. In *International Conference on Recent Advances in Natural Language Processing, RANLP*.
- Miller, George A. and Walter G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes* 6(1). 1–28.
- Mitkov, Ruslan, Branimir Boguraev and Shalom Lappin. 2001. Introduction to the special issue on computational anaphora resolution. *Computational Linguistics* 27(4). 473–477.
- Mollin, Sandra. 2009. Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5(2). 175–200.
- Nelson, Douglas L., Cathy L. McEvoy and Thomas A. Schreiber. 1998. The University of South Florida word association, rhyme, and word fragment norms. <http://web.usf.edu/FreeAssociation/>.
- Palermo, David S. and James J. Jenkins. 1964. *Word association norms: Grade school through college*. Minneapolis: University of Minnesota Press.
- Pedersen, Ted and Rebecca Bruce. 1996. What to infer from a description. Tech. Rep. 96-CSE-04. Southern Methodist University. Dallas, TX.
- Rapp, Reinhard. 2002. The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *19th International Conference on Computational Linguistics, COLING*. Taipei, Taiwan.
- Resnik, Philip. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition* 61(1). 127–159.
- Rubenstein, Herbert and John B. Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM* 8(10). 627–633.
- Sahlgren, Magnus. 2006. *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. Ph.D. thesis. Department of Linguistics, Stockholm University.
- de Saussure, Ferdinand. 1966. *Course in general linguistics*. New York: McGraw-Hill.
- Schütze, Hinrich and Michael Walsh. 2008. A graph-theoretic model of lexical syntactic acquisition. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*. 917–926.
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Spence, Donald P. and Kimberly C. Owens. 1990. Lexical Co-Occurrence and Association Strength. *Journal of Psycholinguistic Research* 19(5). 317–330.
- Strube, Michael and Simone Ponzetto. 2006. Wikirelate! Computing semantic relatedness using wikipedia. In *Twenty-First National Conference on Artificial Intelligence, AAAI*. 1419–1424.
- Tversky, Amos N. 1977. Features of similarity. *Psychological review* 84(4). 327–352.



- Voorhees, Ellen M. 1999. The TREC-8 question answering track report. In *8th Text Retrieval Conference, TREC*. 77–82.
- Washtell, Justin and Katja Markert. 2009. A Comparison of windowless and window-based computational association measures as predictors of syntagmatic human associations. In *Conference on Empirical Methods in Natural Language Processing, EMNLP*. 628–637.
- Weeds, Julie. 2002. Asymmetry in similarity between words. In *Annual CLUK Colloquium*. 1–3. Leeds, UK.
- Wermter, Joachim and Udo Hahn. 2006. You can’t beat frequency (unless you use linguistic knowledge). In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, COLING/ACL*.
- Wettler, Manfred and Reinhard Rapp. 1993. Computation of word associations based on the co-occurrences of words in large corpora. In *1st Workshop on Very Large Corpora*. 84–93.
- Wiechmann, Daniel. 2008. On the computation of collocation strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4(2). 253–290.