Identifying Predictive Collocations

Silas Weinbach (silasw@coli.uni-saarland.de)
Department of Computational Linguistics
Campus C7.2, 66123 Saarbrücken, Germany

Vera Demberg (vera@coli.uni-saarland.de)
Cluster of Excellence, Saarland University,
Campus C7.4, 66123 Saarbrücken, Germany

Abstract

Idioms and common multi-word expressions are often argued to be stored as chunks of words or fixed configurations in the mind, and to therefore be accessed faster and interpreted more easily than fully compositional word combinations. Experimental research has furthermore shown that a specific "recognition point" can be identified in such expressions, at which enough information is present to access the meaning of the whole expression and predict the remaining words of the collocation.

In this paper, we suggest measures for automatically identifying those multi-word expressions where the first part is particularly predictive of the rest, and evaluate our measures against human association data collected in a cloze test.

Keywords: Predictivity; Multi-Word Expressions; Collocations; Entropy

Introduction

"When her boyfriend proposed to her, she was in seventh heaven." "After jogging, he quenched his thirst with some nice orange juice." The above sentences contain collocations where the first part of the collocation (e.g., "in seventh", "quench") is very predictive of the second part ("heaven" and "thirst", respectively). Such predictive collocations can be idiomatic (as in the first example), or literal, fully compositional configurations. Previous studies observing human processing of idioms have argued that there exists a "recognition point", at which comprehenders have identified the idiom and can predict the rest. Some also argue that not only idioms, but also frequent collocations, may be stored in the lexicon.

However, by far, not all collocations are predictive, consider for example light verb constructions where a very unpredictive verb is combined with a sense-carrying noun. Being able to pick out predictive collocations among the set of all collocations, and automatically identifying the recognition point in idioms could be very useful for psycholinguistic models of language processing: Processes of predicting specific upcoming words, and accessing idiomatic meaning could then potentially be captured in a broad-coverage model.

This paper takes a first step in this direction by proposing a number of alternative statistical methods for identifying predictive collocations and evaluating them with respect to a cloze task where people were asked to complete verbs with the argument they associated most strongly. This evaluation captures the predictive strength of a verb in the absence of further predictive context, and is supposed to compare which of the measures works best at identifying good candidates for predictive collocations.

Background and Related Work

Collocations are commonly used phrasal expressions which have become characteristic for a language or jargon (Smadja, 1993). They are idiosyncratic because there is no rule which can tell us why a some specific lexemes (e.g., "strong tea" instead of "powerful tea") are combined to express a particular concept (McKeown & Radev, 2000).

Representation of Collocations in Humans

Idioms are a special type of collocations whose semantic meaning is not compositional of the meaning of the words it contains, but are more idiosyncratic such as "give a whirl" (meaning to try) or "spill the beans". The status of these expressions in the lexicon is still under debate. It has often been argued (Swinney & Cutler, 1979) that these idiomatic expressions should be part of the lexicon. Some have even argued that non-idiomatic collocations may likewise be stored as chunks in longterm memory (Ellis, 2001; Ellis, Frey, & Jalkanen, 2009).

An alternative model was proposed by Cacciari and Tabossi (1988) and holds that both decomposable and idiomatic expressions are represented in the lexicon "as configurations" and that these configurations can get activated during processing as soon as enough information has been perceived to render the collocation recognizable. Tabossi, Fanari, and Wolf (2009) present evidence that both idiomatic and literal collocations may be stored in memory as such configurations.

On the other hand, Vespignani, Canal, Molinaro, Fonda, and Cacciari (2010) find in an ERP experiment which compares the processing of idiomatic expressions with literal phrases that language comprehenders have categorial templates for idioms in their lexicon, and that these can be activated at a specific recognition point after which a prediction process is initiated. Their results suggest that this prediction process can be distinguished from non-idiomatic predictive mechanisms. If such effects are to be modelled in a computational model, it is necessary for the model to have access to a set of idiomatic expressions and their recognition points.

The goal of this paper is not to answer the question concerning which types of collocations may be stored in memory and which ones may be processed compositionally. Instead, we evaluate statistical measures for automatically identifying predictive collocations. The methods and measures are generally applicable and may later be used in combination with a filter for identifying idiomatic expressions.

An important point for our study however is the relevance of a recognition point and the notion of predictability of a multi-word expression. Tabossi, Fanari, and Wolf (2005) showed that only the meanings of *predictable* idioms, but not of all idioms, become available early on in idiom processing. Such a prediction process may be beneficial to language understanding because, as Tabossi et al. (2005) finds, recog-

nizing an initial fragment of a predictable idiom inhibits the recognition of the literal meaning of the rest of the expression and hence facilitates comprehension by reducing ambiguity.

Automatically Identifying Collocations

The basic idea in automatically identifying collocations (for a good overview, see (Manning & Schütze, 1999)) is to count how often a set of words occur together within a specific distance of one another (e.g., always adjacent) or within a syntactic relationship (e.g., verb-argument). Many word-pairs or multi-word expressions with frequent co-occurrence however aren't interesting collocations (like "in the") because the reason for their high co-occurrence frequency is the high frequency of each of the words and the syntactic constraints with which they occur.

Two strategies are commonly used to ignore such cases: the first one is to use statistical tests that indicate whether two words were observed together more often than would be expected otherwise. Collocation candidates are then ranked with respect to significance scores. Note though that only the ranking, but not the exact significance level, is usually considered interesting, as most co-occurrences are significant simply due to the fact that language has some regular patterns due to syntactic rules (Manning & Schütze, 1999). Another common approach is to calculate the pointwise mutual information (PMI) between two words.

The second strategy is to specify what types of collocations should be found by specifying POS tag patterns or dependency relations between words (e.g., only considering adjective-noun pairs or only considering modifiers of nouns).

Finally, automatic methods developed for detecting idiomatic collocations often also use semantics to identify these expressions: in non-compositional expressions, the meaning of the words in the idiom are less likely to be semantically related to the rest of the context (Katz & Giesbrecht, 2006).

The following paragraphs are going to explain the most commonly used measures for detecting collocations, as well as the word patterns used in this work.

Association Measures for Identifying Collocations To assess whether a pair of words w_1w_2 is a collocation, we can count how often these words can be observed together $O_{1,1}$, and calculate how often we would expect to see them together given their unigram frequencies and the size N of our data set: $E_{1,1} = \frac{freq(w_1)}{N} \times \frac{freq(w_2)}{N} \times N$. If we observe them together much more often than would be expected given their unigram frequencies, we conclude that they are strongly associated and represent a collocation.

The most commonly used association measures (AMs) are the following: In a t-test (see for example (Manning & Schütze, 1999)), the higher the t-value, the more likely that the observed co-occurrence of the words w_1 and w_2 would not have happened by chance.

$$\text{t}-\text{Test}: \quad t = \frac{\frac{O_{1,1}-E_{1,1}}{N}}{\sqrt{O_{1,1}}}$$

An alternative is the z-score (variant suggested by Evert (2008)). The formula below estimates the mean of the distribution as $E_{1,1}$ and its standard deviation as $\sqrt{E(1,1)}$

$$\text{z}-\text{score}: \quad z = \frac{O_{1,1}-E_{1,1}}{\sqrt{E_{1,1}}}$$

Pearson's χ^2 test (for a more detailed description, see (Manning & Schütze, 1999)) is very similar to the z-score, except it uses the square of the z-values and takes into account not only the probability of the words occurring together ($O_{1,1}$), but compares also the estimated and observed frequencies of a w_1 not occurring with w_2, w_2 not occurring with w_1 and the co-occurrence of words different from both w_1 and w_2.

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

Finally, the log likelihood ratio λ, similarly to χ^2, uses weighted on the similarity of the words w_1 and w_2 occurring together or with different words.

$$\lambda = 2\sum_{i,j} O_{i,j} \log \frac{O_{i,j}}{E_{i,j}}$$

Pointwise Mutual Information (PMI; Church and Hanks (1989)) is and information-theoretic concept and measures how much information is shared between words w_1 and w_2 – it is a symmetric measure. If there are two words with only occur in the context of each other, then one of the words conveys all the information that the two of them convey and their mutual information is maximal.

$$PMI = \log_2 \frac{O_{1,1}}{E_{1,1}}$$

Filters Previous work on collocation extraction has shown (Seretan & Wehrli, 2009; Fazly, Cook, & Stevenson, 2009; Lin, 1998) that result quality depends also on choosing good patterns in which to observe collocation candidates. These have been defined via windows of observation, via fixed POS tag sequences or via syntactic dependencies, as for example from a dependency parser. The present study focusses on verb-argument pairs as extracted from a large text resource using a dependency parser.

Asymmetric Association Measures While there is a large body of literature on the topic of automatic recognition of multi-word expressions and idioms, there is almost no work on asymmetric association measures. An exception is Michelbacher, Evert, and Schütze (2007, 2011), who use conditional probability (see below), as well as a number of rank measure which are based on the traditional association measures explained above. As we found out after first submitting this paper, a related proposal for developing directional association measures has been made by Gries (to appear). A comparison between our measures and the associative-learning based approach should be addressed in future work.

Proposed Predictive Measures

One way of capturing how predictive one word is of another word is to calculate the conditional probability (CP; also suggested by Michelbacher et al., 2007; 2011) of the second word given the first word. High CP indicates that the first word is highly predictive of the second word.

$$CP(w_1, w_2) = P(w_2|w_1)$$

A straightforward approach to predictive collocations is to use conditional probability as an association measure, or to combine existing measures for association between two words with the conditional probability of the second word given the first word. Different ways of combining the measures are possible, such as for example weighted additive combination ($a \times CP + b \times AM$), or multiplicative combination ($CP \times AM$).

In our preliminary experiments, it turned out that the additive models (which essentially represent a form of averaging between the measures) do not perform well. While they boost the score of collocation candidates which are both strongly associated and predictive, they do usually not boost it enough to achieve rankings higher than those of candidate collocations which are extremely good on just one of the measures, such that the resulting highest ranked candidates still contain a lot of highly associated but non-predictive word pairs.

Multiplicative combination, on the other hand, can be thought of as a filter that ranks down any collocation candidates which are highly associated but not predictive, and boost highly predictive candidates, resulting in a cleaner list of predictive collocations. Based on this observation, we propose the following new measures: CP, CP$\times\chi^2$, CP$\times$PMI and CP$\times\lambda$, which we will evaluate in the remainder of this paper.

Comparison of Association Measures

It is instructive to inspect how similar the alternative association measures are to one another. To this end, we sorted 3.6 million adjective-noun pairs from the ukwac corpus (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008) according to each of our association measures and calculated the correlations between these sorted lists. Table 1 shows that four of our measures, χ^2, Z, λ and PMI actually result in very similar rankings, with correlations $\rho > .9$. Only rankings by t-value look a bit more dissimilar, and relatively more similar than other measures to the overall frequency of word pairs (indicated as FRQ in Table 1). It is also important to observe that conditional probability (CP) leads to a very different ranking and is only correlated at $0.28 < \rho < 0.4$ with the other measures.

Identification of Predictive Collocations

We dependency-parsed the Gigaword Corpus[1] using the Stanford parser (Marneffe, MacCartney, & Manning, 2006). From the Gigaword's 1.7 billion tokens, we extracted all dependency triples of the type "VB*:dobj:NN*" (i.e., verbs and their direct arguments), for which the verb occurred to the left of the argument in the text. Verb-argument pairs which occurred less than 16 times in the corpus were excluded from the analysis, as some of the association measures are not applicable when counts are too low. Furthermore, we removed all verb-argument pairs containing words which were not in WordNet under the correct POS tag. This later step filtered out POS-tagging errors like "unsalted butter" or "quantum mechanic" where "unsalted" and "quantum" were tagged as verbs, or "smile slyly" where "slyly" was tagged as a noun, as well as foreign language material.

[1] http://www.ldc.upenn.edu

Table 1: Correlation (Spearman's rho) between different association measures for top 500 ranks.

ρ	FRQ	T	Z	χ^2	λ	PMI	CP
FREQ	1	.62	.28	.29	.46	.06	.2
T	.62	1	.86	.83	.88	.72	.28
Z	.28	.86	1	.97	.91	.96	.38
χ^2	.29	.83	.97	1	.97	.93	.4
λ	.46	.88	.91	.97	1	.82	.4
PMI	.06	.72	.96	.93	.82	1	.33
CP	.2	.28	.38	.4	.4	.33	1

Cloze Task

The goal of our experiment is to evaluate whether combining one of the established measures for collocation extraction with conditional probabilities will lead to a good measure for identifying predictive collocations, and which of the proposed measures works best. For the evaluation, we use a simple task which is independent of any sentential context: we ask human participants to complete a list of verbs with a noun they associate first, and then compare which of our measures predicts best the cloze probabilities of each verb. A reason for evaluating with a completion experiment instead of simply comparing to a verb's entropy or conditional probability on the corpus itself is that many of the highly ranked collocations in our measures are in fact not necessarily generally valid predictive collocations – some are very domain-specific, such as *rise percent* (from "rise 20 percent") and *tell reporter*.

Experimental Materials

For evaluating predictive collocations, we were looking for a set of verbs which contains a good portion of potentially predictive verbs. We therefore selected verbs for our completion experiment by first calculating ranked lists of some of our target measures that we want to compare: CP$\times\chi^2$, χ^2 and CP$\times\lambda$, and randomly chose 50 verbs out of the 200 best-ranked verb-object pairs of each measure. This procedure left us with a set of 118 verbs for our completion experiment. The rationale behind choosing verbs this way instead of just selecting a random set of verbs is that we wanted to avoid ending up with only a very small number of predictive verbs.

Table 2: Arguments filled in for the verb "heal" during our completion experiment. We also collected completion times for each response.

Answer.*w2*	Seconds	Answer.*w2*	Seconds
the sick	7	wounds	55
a wound	19	bodies	8
the sick	5	a wound	8
the wound	37	yourself	15
a wound	13	a sore	9
a wound	18	the wound	10
sores	4	wounds	4
a wound	7	the wound	7

Procedure

We ran our experiment via Amazon Mechanical Turk (Paolacci, Chandler, & Ipeirotis, 2010). In order to explain the task to our subjects, we gave them three examples of completed verb-argument pairs, using verbs which were not part of the 118 verbs that we wanted to collect completion data for: "to quench thirst", "to rob a bank" and "to feed the dog". We restricted our subjects to people living in the U.S. and instructed them to only take part in the experiment if they qualified as native speakers of English. Furthermore, we also restricted our pool of workers to ones that had in the past gotten $> 95\%$ of their HITs[2] approved and had successfully completed at least 1000 HITs. We collected a total of 1888 verb-argument associations (i.e., 16 associations for each verb). Each worker was allowed to complete as many verbs as they wanted (but, of course, each verb only once). The 1888 associations were completed by 40 separate workers.

Collected Data

For each verb, we collected 16 argument-associations. For example, see completions for the verb "heal" in Table 2. We lemmatized all answers, and dealt with typos (e.g., *havok* instead of *havoc*), orthographic variants (e.g., *judgment* vs. *judgement*) using minimum edit distance.

To assess the predictive strength of a verb, we calculated the entropy of each verb given the types of responses (after clustering them by lemma and dealing with typos etc, as described above). For example, the entropy of "heal" would be 1.53. As we collected at most 16 associations per verb, entropy ranges between 0 and 4 for our data set. We can then use the entropy to classify our verbs into highly selective verbs (such as "grit", "honk", "flex", "sing", "twiddle"), less selective verbs (e.g., "pay", "fire", "attend") and non-selective ones (e.g., "quote", "shout", "request"). In a linear mixed effects regression analysis with random intercept and random slope for verb entropy under subject, we found that verb entropy is a significant positive predictor of completion times ($p < 0.01$), i.e., when an argument of a verb is less predictable, people take longer to fill in the slot.

[2]HIT stands for "Human Intelligence Task" and is used as the official term for tasks in Amazon Mechanical Turk.

Evaluation

We evaluate our measures of predictive collocations in two ways. A good measure should rank highly those collocations where the first part is highly predictive of the second part.

Identifying Predictive Collocations We select a group of highly predictive verbs (determined by their entropy in the experiment) and generate verb-noun pairs by selecting the most common completion for those verbs in the experiment. This results in a list of verb-noun collocations where the verb is highly predictive of the noun. Next, we calculate the average rank of these verb-noun pairs for each of our measures, see table in Figure 1.

An important note to keep in mind when interpreting the average ranks in the table in Figure 1 is that the set of verbs was originally randomly chosen from among the top-ranked 200 verb noun pairs of the measures CP×CHI, CHI and CP×λ; note also that Z is almost identical to CHI in the ranking it generates – these measures are therefore marked in bold in the table.

The newly proposed measure CP×CHI clearly outperforms the other measues. It has the lowest average rank, meaning that the verb-noun pairs which we have identified as being particularly predictive are ranked highest in this measure. Note that 44 verb-noun pairs satisfied the criterion of the verb entropy in the experiment being under the threshold of 1.5. This gives us an average rank of 22.5 as the best possible ranking which could possibly be achieved. Of course, not all possible verbs were tested in our experiment, hence direct comparison to this value is not meaningful. More important is the comparison to the average ranks of other measures. Clearly, the combined measure CP×CHI is much better than either of its parts, and also clearly outperforms CP×λ.

It is also fair to compare the measures which were not part of constructing the evaluation verb set (not in bold) to one another. Clearly, combining CP with the association measures improves identification of predictive collocations, in particular there is an interesting boost in the performance of the t-test measure when combined with conditional probabilities. We also conclude that λ is not a useful measure for identifying predictive collocations.

Additional insight comes from plotting average ranks for all verb-noun pairs with identical cloze probability, see Fig. 1. For a measure which is good at identifying predictive collocations, we expect there to be a linear relationship between cloze probabilities and log rank (log rank makes sense because there are by definition more different noun pairs when cloze probability is lower). Monotonicity in the trend of the log rank indicates that the measure correctly distinguishes between different levels of cloze predictability. Furthermore, average log rank for verb noun pairs with cloze probability 1 should be close to 0. The plots show that CP×CHI comes closest to the described ideal correlation. The r squared measure given in the title of each plot in Figure 1 quantifies the fit between the plotted data points are from the regression line.

verb entropy threshold 1.5	
measure	rank
CPxCHI	87.275
CHI	152.864
Z	152.948
CPxPMI	237.692
CPxT	258.124
CP	291.342
CPxλ	404.471
CPxFRQ	709.715
PMI	1037.151
λ	2403.245
Z	6863.703
T	9378.727
ceiling	22.5

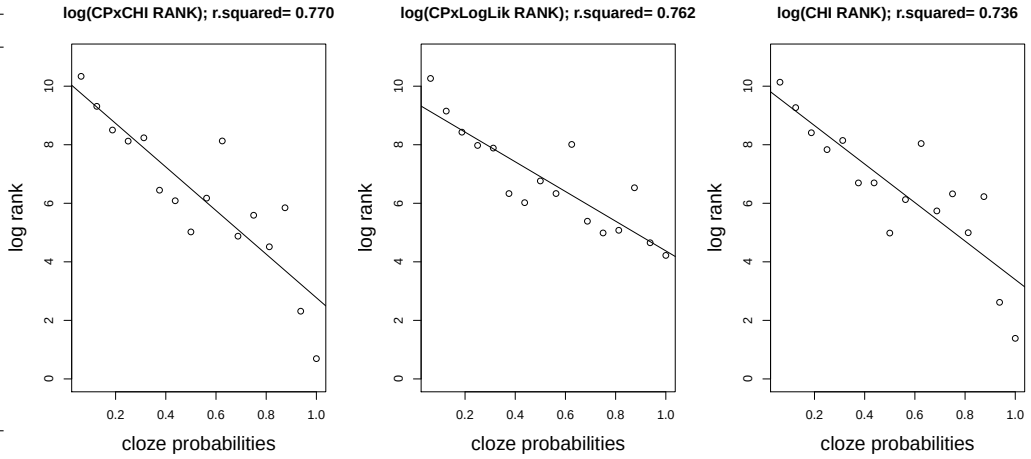

Figure 1: Table at left: Average rank in lists ranked according to association measures; set of predictive verb-noun pairs defined based on different thresholds for verb entropy in experiment.
Plots: Average rank for CPxCHI, CPxλ and CHI grouped by cloze probabilities as obtained from MTurk experiment.

While measure CP×λ also follows a clear linear relationship, it does not locate the items with high predictability in its lowest ranks, indicating that it might be a good measure for quantifying collocations in general but not for predictiveness given the first part.

Correlation with human associations Our second evaluation compares the association values from all measures to the cloze probabilities obtained in the experiment. We again evaluate on average association values for each set of verb-noun pairs with a given cloze probability, see Figures 2 and 3. A good measure should increase monotonically with increasing cloze probabilities.

Among previously existing measures, PMI values can explain the largest amount of the variance in terms of average PMI values compared to cloze probabilities from our experiments. It is clear from Figure 2 that the common frequency of the two words, as well as the log likelihood measure are very poor predictors of predictive collocations.

Among traditional measures combined with conditional probabilities, CP×FRQ, CP×λ and CP×T perform very poorly. The problem for these measures is that they reflect strongly the overall frequency of a word pair. On the other hand, average log CP×CHI values have the strongest linear relationship with cloze probabilities, with few atypical points, as also reflected in the high R^2. This result is thus consistent with the rank analysis in the first evaluation.

Conclusions and Outlook

Our experiments indicate that the combination of conditional probability and the χ^2 measure might work best for identifying collocations where the first word is highly predictive of the second one. While this paper went a first step in devoting some attention to the problem of identifying predictive collocations, suggesting possible measures and evaluating these measures on a cloze task of verb-argument associations, the next important step is to evaluate these methods on a more specific task such as automatically identifying recognition points of idioms. Furthermore, this paper has only dealt with one type of collocation (verb-argument pairs) and has focussed on collocations consisting of only two words.

In future work, we furthermore plan to evaluate the usefulness of predictive collocations by including them in a model of language processing in the form of lexical configurations.

Acknowledgments

We would like to thank Stefan Thater for providing the Stanford-parsed version of Gigaword.

References

Cacciari, C., & Tabossi, P. (1988). The comprehension of idioms. *Journal of memory and language*, *27*(6), 668–683.

Church, K. W., & Hanks, P. (1989). Word association norms, mutual information and lexicography. In *ACL* (Vol. 27, p. 76-83).

Ellis, N. (2001). Memory for language. *Cognition and second language instruction*, 33–68.

Ellis, N., Frey, E., & Jalkanen, I. (2009). The psycholinguistic reality of collocation and semantic prosody (1). *Studies in Corpus Linguistics (SCL)*, 89.

Evert, S. (2008). Corpora and collocations. *Corpus Linguistics. An International Handbook*, *2*.

Fazly, A., Cook, P., & Stevenson, S. (2009). Unsupervised type and token identification of idiomatic expressions. *Computational Linguistics*, *35*(1), 61–103.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. *In Proceedings of the WAC4 Workshop at LREC08*.

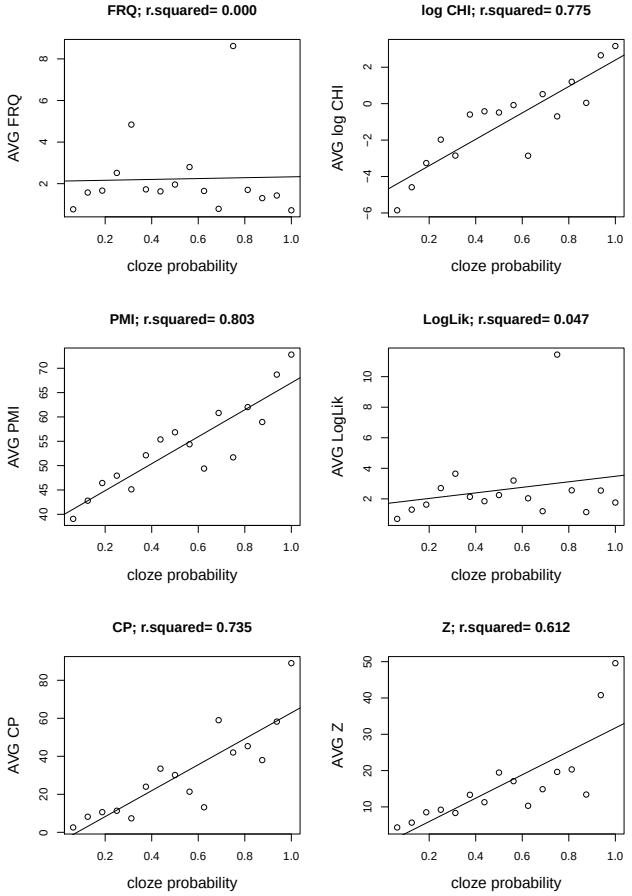

Figure 2: Average association values for each measure grouped by cloze probabilities from MTurk experiment.

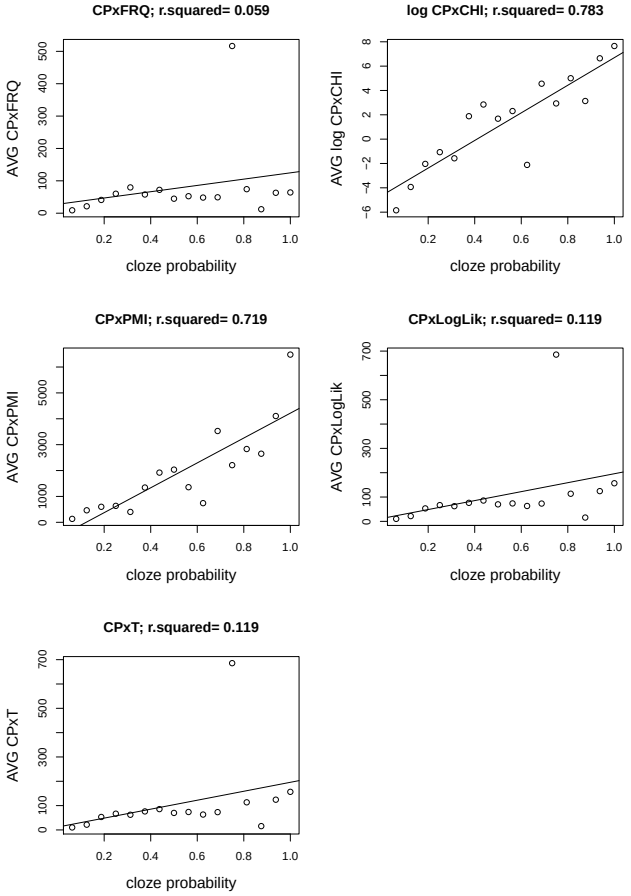

Figure 3: Average association values for each combined measure grouped by cloze probabilities from MTurk experiment.

Gries, S. T. (to appear). 50-something years of work on collocations: what is or should be next. *International Journal of Corpus Linguistics*, *18*(1).

Katz, G., & Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the workshop on multiword expressions* (pp. 12–19).

Lin, D. (1998). *Extracting collocations from text corpora.*

Manning, C. D., & Schütze, H. (1999). Foundations of statistical natural language processing. In (second ed., chap. 5). The MIT Press.

Marneffe, M.-C. de, MacCartney, B., & Manning, C. D. (2006). Generating typed dependency parses from phrase structure parses. *LREC*.

McKeown, K. R., & Radev, D. R. (2000). Collocations. In *A handbook of natural language processing* (pp. 507–523).

Michelbacher, L., Evert, S., & Schütze, H. (2007). Asymmetric association measures. *Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007)*.

Michelbacher, L., Evert, S., & Schütze, H. (2011). Asymmetry in corpus-derived and human word associations. *Corpus Linguistics and Linguistic Theory*, *7*(2), 245–276.

Paolacci, G., Chandler, J., & Ipeirotis, P. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*(5), 411–419.

Seretan, V., & Wehrli, E. (2009, March). Multilingual collocation extraction with a syntactic parser. *Language Resources and Evaluation*, *43*(1), 71–85.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Association for Computational Linguistics*.

Swinney, D., & Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*, *18*(5), 523–534.

Tabossi, P., Fanari, R., & Wolf, K. (2005). Spoken idiom recognition: Meaning retrieval and word expectancy. *Journal of psycholinguistic research*, *34*(5), 465–495.

Tabossi, P., Fanari, R., & Wolf, K. (2009). Why are idioms recognized fast? *Memory & cognition*, *37*(4), 529–540.

Vespignani, F., Canal, P., Molinaro, N., Fonda, S., & Cacciari, C. (2010). Predictive mechanisms in idiom comprehension. *Journal of cognitive neuroscience*, *22*(8), 1682–1700.