Computation of Word Associations Based on the Co-Occurences of Words in Large Corpora[1]

Manfred Wettler & Reinhard Rapp
University of Paderborn, Cognitive Psychology
Postfach 1621, D-4790 Paderborn, Germany

Abstract

A statistical model is presented which predicts the strengths of word-associations from the relative frequencies of the common occurrences of words in large bodies of text. These predictions are compared with the Minnesota association norms for 100 stimulus words. The average agreement between the predicted and the observed responses is only slightly weaker than the agreement between the responses of an arbitrary subject and the responses of the other subjects. It is shown that the approach leads to equally good results for both English and German.

1 Introduction

In the association experiment first used by Galton (1880) subjects are asked to respond to a stimulus word with the first word that comes to their mind. These associative responses have been explained in psychology by the principle of learning by contiguity: "Objects once experienced together tend to become associated in the imagination, so that when any one of them is thought of, the others are likely to be thought of also, in the same order of sequence or coexistence as before. This statement we may name the law of mental association by contiguity." (William James, 1890, p. 561).

When the association experiment is conducted with many subjects, tables are obtained which list the frequencies of particular responses to the stimulus words. These tables are called association norms. Many studies in psychology give evidence that there is a relation between the perception, learning and forgetting of verbal material and the associations between words.

If we assume that word-associations determine language production, then it should be possible to estimate the strength of an associative relation between two words on the basis of the relative frequencies that these words co-occur in texts. Church et al. (1989), Wettler & Rapp (1989) and Church & Hanks (1990) describe algorithms which do this. However, the validity of these algorithms has not been tested by systematic comparisons with associations of human subjects. This paper describes such a comparison and shows that corpus-based computations of word associations are similar to association norms collected from human subjects.

According to the law of association by contiguity, the association strength between two words should be a function of the relative frequency of the two words being perceived together, i.e. the relative frequency of the two words occuring together. Further more, the association strength between words should determine word selection during language or speech production: Only those words can be uttered or written down which associatively come to mind. If this assumption holds, then it should be possible to predict word associations from the common occurences of words in texts.

[1] This research was supported by the Heinz-Nixdorf-Institute (project 536)

2 Model

According to the law of association by contiguity the learning of associations can be described as follows: If two words i and j occur together, the association strength $a_{i,j}(t)$ between i and j is increased by a constant fraction of the difference between the maximum and the actual association strength. This leads for association strengths between 0 and 1 to the following formula:

$$a_{i,j}(t+1) = a_{i,j}(t) + (1 - a_{i,j}(t)) \cdot \theta_1 \qquad \text{if} \quad (i\&j) \tag{1}$$

If word i occurs in another context, i. e. not in proximity to word j, the association strength between i and j is decreased by a constant fraction:

$$a_{i,j}(t+1) = a_{i,j}(t) \cdot (1 - \theta_2) \qquad \text{if} \quad (i\&\neg j) \tag{2}$$

Under the assumption that the learning rate θ_1 and the inhibition rate θ_2 are of identical size, the expected value $a_{i,j}$ of the association strength $a_{i,j}(t)$ from i to j for $t \rightarrow \infty$ is equal to the conditional probability of j given i (compare Foppa, 1965):

$$a_{i,j} = p(j|i) \tag{3}$$

From these assumptions it could be expected that a stimulus word i leads to those response j, for which the value of equation 3 is at a maximum.

Rapp & Wettler (1991) compared this with other predictions, where additional assumptions on learning and reproduction were taken into account. With equation 3, mainly words with high corpus frequencies, e. g. function words, were predicted as associative responses. The predictions were improved when the following formula was used with an exponent of $\alpha = 0.66$, and the word with the highest $r_{i,j}$ was considered to be the associative response.

$$r_{i,j} = \frac{p(j|i)}{p(j)^{\alpha}} \tag{4}$$

The introduction of the denominator indicates that in the association experiment less frequent words are used than during language production. This inhibition of frequent words can be explained by the experimental situation, which furthers responses that are specific to the stimulus word. The exponential function can be interpreted as the tendency that subjective estimations are often found to be exponential functions of the quantities to be estimated.

3 Association norms used

For the comparison between the predicted and the associations of human subjects we have used the association norms collected by Russell & Jenkins (Jenkins, 1970). They have the advantage that translations of the stimulus words were also given to German subjects (Russell & Meseck, 1959, and Russell, 1970) so that our model could be tested for English as well as for German.

The Russell & Jenkins association norms, also referred to as the Minnesota word association norms, were collected in 1952. The 100 stimulus words from the Kent-Rosanoff word association test (Kent & Rosanoff, 1910) were presented to 1008 students of two large introductory psychology classes at the University of Minnesota. The subjects were instructed, to write after each word "the first word that it makes you think of". Seven years later, Russell & Meseck (1959) repeated the same experiment in Germany with a carefully translated list of the stimulus words. The subjects were 331 students and pupils from the area near Würzburg. The quantitative results reported on later will be based on comparisons with these norms.

The American as well as the German association norms were collected more than 30 years ago. The texts which were used to simulate these associations are more recent. One might expect therefore that this discrepancy will impair the agreement between the observed and the predicted responses. Better predictions might be attained if the observed associations had been produced by the same subjects as the texts from which the predictions are computed. However, such a procedure is hardly realizable, and our results will show that despite these discrepancies associations to common words can be predicted successfully.

4 Text corpora

In order to get reliable estimates of the co-occurences of words, large text corpora have to be used. Since associations of the "average subject" are to be simulated, the texts should not be specific to a certain domain, but reflect the wide distribution of different types of texts and speech as perceived in every day life.

The following selection of some 33 million words of machine readable English texts used in this study is a modest attempt to achieve this goal:

- Brown corpus of present day American English (1 million words)
- LOB corpus of present day British English (1 million words)
- Belletristic literature from Project Gutenberg (1 million words)
- Articles from the New Scientist from Oxford Text Archive (1 million words)
- Wall Street Journal from the ACL/DCI (selection of 6 million words)
- Hansard Corpus. Proceedings of the Canadian Parliament (selection of 5 million words from the ACL/DCI-corpus)
- Grolier's Electronic Encyclopedia (8 million words)
- Psychological Abstracts from PsycLIT (selection of 3.5 million words)
- Agricultural abstracts from the Agricola database (3.5 million words)
- DOE scientific abstracts from the ACL/DCI (selection of 3 million words)

To compute associations for German the following corpora comprising about 21 million words were used:

- LIMAS corpus of present-day written German (1.1 million words)
- Freiburger Korpus from the Institute for German Language (IDS), Mannheim (0.5 million words of spoken German)
- Mannheimer Korpus 1 from the IDS (2.2 million words of present-day written German from books and periodicals)
- Handbuchkorpora 85, 86 and 87 from the IDS (9.3 million words of newspaper texts)
- German abstracts from the psychological database PSYNDEX (8 million words)

For technical reasons, not all words occuring in the corpora have been used in the simulation. The vocabulary used consists of all words which appear more than ten times in the English or German corpus. It also includes all 100 stimulus words and all responses in the English or German association norms. This leads to an English vocabulary of about 72000 and a German vocabulary of 65000 words. Hereby, a word is defined as a string of alpha characters separated by non-alpha characters. Punctuation marks and special characters are treated as words.

5 Computation of the association strengths

The text corpora were read in word by word. Whenever one of the 100 stimulus words occured, it was determined which other words occured within a distance of twelve words to the left or to the right of the stimulus word, and for every pair a counter was updated. The so defined frequencies of co-occurence $H(i\&j)$, the frequencies of the single words $H(i)$ and the total number of words in the corpus Q were stored in tables. Using these tables, the probabilities in formula (4) can be replaced by relative frequencies:

$$\frac{H(i\&j)}{H(i)} \Big/ \frac{H(j)^{\alpha}}{Q^{\alpha}} = \frac{Q^{\alpha}}{H(i)} * \frac{H(i\&j)}{H(j)^{\alpha}} \tag{5}$$

In this formula the first term on the right side does not depend on j and therefore has no effect on the prediction of the associative response. With $H(j)$ in the denominator of the second term, estimation errors have a strong impact on the association strengths for rare words. Therefore, by modifying formula (5), words with low corpus frequencies had to be weakened.

$$\tilde{r}_{i,j} = \begin{cases} H(i\&j)/H(j)^{\alpha} & \text{für} \quad H(j) > \beta \cdot Q \\ H(i\&j)/(\gamma \cdot Q) & \text{für} \quad H(j) \leq \beta \cdot Q \end{cases} \tag{6}$$

According to our model the word j with the highest associative strength $\tilde{r}_{i,j}$ to the stimulus word i should be the associative response. The best results were observed when parameter α was chosen to be 0.66. Parameters β and γ turned out to be relatively uncritical, and therefore to simplify parameter optimization were both set to the same value of 0.00002.

Ongoing research shows that formula (6) has a number of weaknesses, for example that it does not discriminate words with co-occurence-frequency zero, as discussed by Gale & Church (1990) in a comparable context. However, since the results reported on later are acceptable, it probably gets the major issues right. One is, that subjects usually respond with common, i.e. frequent words in the free association task. The other is, that estimations of co-occurence-frequencies for low-frequency-words are too poor to be useful.

6 Results

In table 1 a few sample association lists as predicted by our system are compared to the associative responses as given by the subjects in the Russell & Jenkins experiment. A complete list of the predicted and observed responses is given in table 2. It shows for all 100 stimulus words used in the association experiment conducted by Russell & Jenkins, a) their corpus frequency, b) the primary response, i.e. the most frequent response given by the subjects, c) the number of subjects who gave the primary response, d) the predicted response and e) the number of subjects who gave the predicted response.

The valuation of the predictions has to take into account that association norms are conglomerates of the answers of different subjects which differ considerably from each other. A satisfactory prediction would be proven if the difference between the predicted and the observed responses were about equal to the difference between an average subject and the rest of the subjects. The following interpretations look for such correspondences.

For 17 out of the 100 stimulus words the predicted response is equal to the observed primary response. This compares to an average of 37 primary responses given by a subject in the Russell & Jenkins experiment. A slightly better result is obtained for the correspondence between the predicted and the observed associations when it is considered, how many

Stimulus	Predicted Responses	$\bar{r}_{i,j}$	Observed Responses	No. Subj.
blue	green	2.144	sky	175
	red	1.128	red	160
	yellow	1.000	green	125
	white	0.732	color	66
	flowers	0.614	yellow	56
	sky	0.600	black	49
	colors	0.538	white	44
	eyes	0.471	water	36
	bright	0.457	grey	28
	color	0.413	boy	20
butter	bread	0.886	bread	637
	milk	0.256	yellow	81
	eggs	0.197	soft	30
	lb	0.179	fat	24
	sugar	0.157	food	22
	fat	0.147	knife	20
	peanut	0.145	eggs	16
	fats	0.138	cream	14
	flavor	0.130	milk	13
	wheat	0.128	cheese	9
baby	mother	0.618	boy	162
	foods	0.427	child	142
	breast	0.353	cry	113
	feeding	0.336	mother	71
	infant	0.249	girl	51
	birth	0.245	small	43
	born	0.242	infant	27
	milk	0.208	cute	21
	her	0.206	little	18
	nursing	0.202	blue	17
cold	hot	1.173	hot	348
	warm	1.164	snow	218
	weather	0.736	warm	168
	winter	0.603	winter	66
	climate	0.474	ice	29
	air	0.424	Minnesota	13
	war	0.342	wet	13
	wet	0.333	dark	10
	water	0.330	sick	9
	dry	0.315	heat	8

Table 1: Comparison between the ten strongest predicted and the ten most frequent observed responses for four stimulus words. $\bar{r}_{i,j}$ was computed according to formula 6.

subjects had given the predicted response: Averaged over all stimulus words and all subjects, a predicted response was given by 12.6% of the subjects. By comparison, an associative response of an arbitrary subject was given by 21.9% of the remaining subjects.

When only those 27 stimulus words are considered, whose primary response was given by at least 500 subjects, an arbitrary response was given by 45.5% of the subjects on average. By comparison, the predicted response to one of these 27 stimulus words was given by 32.6% of the subjects. This means, that for stimulus words where the variation among subjects is small, the predictions improve.

On the other hand, 35 of the predicted responses were given by no subject at all, whereas an average subject gives only 5.9 out of 100 responses that are given by no other subject. In about half of the cases we attribute this poor performance to the lack of representativity of the corpus. For example, the predictions *combustion* to the stimulus *bed* or *brokerage* to *house* can be explained by specific verbal usage in the DOE scientific abstracts respectively in the Wall Street Journal.

In most other cases instead of paradigmatic associations (words that are used in similar contexts) syntagmatic associations (words that are often used together) are predicted. Examples are the prediction of *term* to the stimulus *long*, where most subjects answered with *short*, or the prediction of *folk* to *music*, where most subjects responded with *song*.

stim	freq	par	f (par)	pred	f (pred)
afraid	692	fear	261	am	0
anger	615	mad	351	expression	0
baby	1157	boy	162	mother	71
bath	244	clean	314	hot	10
beautiful	812	ugly	209	love	0
bed	1295	sleep	584	combustion	0
Bible	593	God	236	Society	0
bitter	541	sweet	652	sweet	652
black	4250	white	751	white	751
blossom	50	flower	672	flower	672
blue	1676	sky	175	green	125
boy	1174	girl	768	girl	768
bread	863	butter	610	wheat	4
butter	426	bread	637	bread	637
butterfly	68	moth	144	fish	0
cabbage	116	head	165	potatoes	0
carpet	138	rug	460	red	27
chair	577	table	493	clock	0
cheese	566	crackers	108	milk	47
child	8897	baby(ies)	159	care	10
citizen	525	U.S.(A.)	114	senior	0
city	8125	town	353	pop	0
cold	2003	hot	348	hot	348
comfort	386	chair	117	ease	76
command	799	order	196	army	102
cottage	137	house	298	cheese	111
dark	1695	light	829	brown	1
deep	2418	shallow	318	sea	77
doctor	766	nurse	238	patient	11
dream	629	sleep	453	sleep	453
eagle	92	bird	550	bird	550
earth	1429	round	130	rare	0
eating	2823	food	390	habits	0
foot	1169	shoe(s)	232	square	0
fruit	1841	apple	378	vegetable	114
girl	1096	boy	704	boy	704
green	1686	grass	262	blue	122
hammer	173	nail(s)	537	string	0
hand	5146	foot(ee)	255	On	0
hard	3502	soft	674	hit	1
head	5350	hair	129	tail	17
health	11433	sickness	250	mental	0
heavy	3497	light	583	ion	0
high	25220	low	675	low	675
house	3059	home	247	brokerage	0
hungry	268	food	362	eat	174
joy	246	happy	209	fear	5
justice	1314	peace	250	criminal	1
king	1983	queen	751	emperor	1
lamp	330	light	633	light	633

Table 2, part 1. Observed and predicted associative responses to stimulus words 1 to 50. The abbreviations in the headline mean: stim = stimulus word; freq = corpus frequency of stimulus word; par = primary associative response; f (par) = number of subjects who gave the primary associative response; pred = predicted associative response; f (pred) = number of subjects who gave the predicted associative response.

stim	freq	par	f (par)	pred	f (pred)
light	7538	dark	647	dark	647
lion	182	tiger	261	sea	2
long	16437	short	758	term	0
loud	230	soft	541	noise	210
man	7472	woman(e)	767	woman	767
memory	3230	mind	119	deficits	0
moon	295	stars	205	sun	168
mountain	1066	hill(s)	266	ranges	0
music	3635	song(s)	183	folk	0
mutton	39	lamb	365	beef	32
needle	208	thread	464	sharing	0
ocean	1066	water	314	floor	6
priest	311	church	328	Catholic	189
quiet	673	loud	348	sleep	53
red	3029	white	221	yellow	19
religion	1224	church	285	Christianity	5
river	1624	water	246	flows	0
rough	457	smooth	439	smooth	439
salt	2158	pepper	430	sugar	83
scissors	25	cut	671	pair	1
sheep	854	wool	201	cattle	15
short	7388	tall	397	term	0
sickness	207	health	376	motion	0
sleep	1843	bed	238	hrs	0
slow	1858	fast	752	wave	0
smooth	690	rough	328	muscle	1
soft	1681	hard	445	drink	10
soldier	321	army	187	army	187
sour	154	sweet	568	sweet	568
spider	97	web	454	tail	0
square	1430	round	372	root	22
stem	796	flower	402	brain	2
stomach	501	food	211	cancer	1
stove	98	hot	235	kitchen	16
street	859	avenue	190	corner	20
sweet	700	sour	434	potatoes	0
swift	184	fast	369	rivers	0
table	2396	chair	840	honour	0
thief	63	steal	286	catch	2
thirsty	32	water	348	drink	296
tobacco	1056	smoke	515	textiles	0
trouble	1108	bad	89	ran	0
whiskey	63	drink(s)	284	beer	52
whistle	77	stop	131	train	89
white	4807	black	617	black	617
window	816	door	191	glass	171
wish	2061	want	124	I	2
woman	2995	man(e)	646	yr	0
working	5366	hard	132	class	3
yellow	1188	blue	156	green	89
MEAN:	2064.78		377.52		127.34

Table 2, part 2. Observed and predicted associative responses to stimulus words 51 to 100.

Using the corpora listed in section 4, the same simulation as described above was conducted for German. For the computation of the associative strengths, again formula 6 was used. For optimal results, only a small adjustment had to be made to parameter alpha (from 0.66 to 0.68). However, a significant change was necessary for parameters β and γ, which again for ease of parameter optimization were assumed to be identical. β and γ had to be reduced by a factor of approximately four from a value of 0.00002 to a value of 0.000005. Apart from these parameters, nothing was changed in the algorithm.

Table 3 compares the quantitative results as given above for both languages. The figures can be interpreted as follows: With an average of 21.9% of the other subjects giving the same response as an arbitrary subject, the variation among subjects is much smaller in English than it is in German (8.7%). This is reflected in the simulation results, where both figures (12.6% and 6.9%) have a similar ratio, however at a lower level.

This observation is confirmed when only stimuli with low variation of the associative responses are considered. In both languages, the decrease in variation is in about the same order of magnitude for experiment and simulation. Overall, the simulation results are somewhat better for German than they are for English. This may be surprising, since with a total of 33 million words the English corpus is larger than the German with 21 million words. However, if one has a closer look at the texts, it becomes clear, that the German corpus, by incorporating popular newspapers and spoken language, is clearly more representative to everyday language.

Description	English	German
percentage of subjects who give the predicted associative response	12.6%	6.9%
percentage of other subjects who give the response of an arbitrary subject	21.9%	8.7%
percentage of subjects who give the predicted associative response for stimuli with little response variation*	32.6%	15.6%
percentage of other subjects who give the response of an arbitrary subject for stimuli with little response variation*	45.5%	18.1%
percentage of cases where the predicted response is identical to the observed primary response	17.0%	19.0%
percentage of cases where the response of an arbitrary subject is identical to the observed primary response	37.5%	22.5%
percentage of cases where the predicted response is given by no subject **	35.0%	57.0%
percentage of cases where the response of an arbitrary subject is given by no other subject**	5.9%	19.8%

Table 3: Comparison of results between simulation and experiment for English and German. Notes: *) little response variation is defined slightly different for English and German: in the English study, only those 27 stimulus words are considered, whose primary response is given by at least 500 out of 1008 subjects. In the German study, only those 26 stimulus words are taken into account, whose primary response is given by at least 100 out of 331 subjects. **) for comparison of English and German experimental figures, it should be kept in mind, that the American experiment was conducted with 1008, but the German experiment with only 331 subjects.

7 Discussion and conclusion

In the simulation results a bias towards syntagmatic associations was found. Since the associations were computed from co-occurences of words in texts, this preference of syntagmatic associations is not surprising. It is remarkable, instead, that many associations usually considered to be paradigmatic are predicted correctly. Examples include *man* → *woman*, *black* → *white* and *bitter* → *sweet*. We believe, however, that the tendency to prefer syntagmatic associations can be reduced by not counting co-occurences found within collocations. Equivalently, the association strength between word pairs always occuring together in a strict formation (separated by a constant number of other words) could be reduced.

When going from English to German, the parameters β and γ in equation 6 needed to be readjusted in such a way, that less frequent words obtained a better chance to be associated. This reflects the fact, that there is more variation in the associative responses of German than of American subjects, and that American subjects tend to respond with words of higher corpus frequency. We believe that by considering additional languages this parameter adjustment could be predicted from word-frequency-distribution.

In conclusion, the results show, that free word associations for English and German can be successfully predicted by an almost identical algorithm which is based on the co-occurence-frequencies of words in texts. Some peculiarities in the associative behavior of the subjects were confirmed in the simulation. Together, this is a good indication that the learning of word associations is governed by the law of association by contiguity.

Although our simulation results are not perfect, specialized versions of our program have already proved useful in a number of applications:

- Information Retrieval: Generation of search terms for document retrieval in bibliographic databases (Wettler & Rapp, 1989, Ferber, Wettler & Rapp, 1993).
- Marketing: Association norms are useful to predict what effects word usage in advertisements has on people (Wettler & Rapp, 1993). Multilingual assocation norms help to find a global marketing strategy in international markets (Kroeber-Riel, 1992).
- Machine Translation: In an experimental prototype we have shown that associations derived from context are useful to find the correct translations for semantically ambiguous words.

The successful prediction of different types of verbal behavior on the basis of co-occurrences of words in texts is a direct application of the classical contiguity-theory, or, in more modern neurophysiological terms, of Hebb's learning rule. Cognitive psychology has severely criticized contiguity-theory with the arguments that association theory did not produce useful results (Jenkins, 1974), and that associations are not the result of associative learning but of underlying semantic processes (Clark, 1970). Both arguments need a critical revision. Recent work with large corpora as well as a large number of connectionist studies have yielded very useful results in different psychological domains, and the high predictive power of the associationist approach makes that the intuitive appeal of cognitivist explanations is fading rapidly.

References

Clark, H.H. (1970). Word associations and linguistic theory. In: J. Lyons (ed.), *New horizons in linguistics.* Harmondsworth: Penguin, 271–286.

Church, K.W., Gale, W., Hanks, P. & Hindle, D. (1989). Parsing, word associations and typical predicate-argument relations. In: *Proceedings of the International Workshop on Parsing Technologies*, Carnegie Mellon University, PA, 389–398.

Church, K.W., Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics, Volume 16, Number 1*, March 1990, 22–29.

Ferber, R., Wettler, M., Rapp, R. (1993). An associative model of word selection in the generation of search queries. In preparation.

Foppa, K. (1965). *Lernen, Gedächtnis, Verhalten.* Köln: Kiepenheuer & Witsch.

Gale, W. A., Church, K. W. (1990). Poor estimates of context are worse than none. *DARPA Speech and Natural Language Workshop*, Hidden Valley, PA, June 1990, 283–287.

Galton, F. (1880). *Psychometric experiments.* Brain 2, 149–162.

James, W. (1890). *The principles of psychology.* New York: Dover Publications.

Jenkins, J.J. (1970). The 1952 Minnesota word association norms. In: Postman, L., Keppel, G. (eds.): *Norms of word association.* New York: Academic Press, 1–38.

Jenkins, J.J. (1974). Remember that old theory of learning? Well, forget it! *American Psychologist*, 29, 785–795.

Kent, G.H. & Rosanoff, A.J. (1910). A study of association in insanity. *American Journal of Insanity,* 67 (1910), 37-96, 317–390.

Kroeber-Riel, W. (1992). Globalisierung der Euro-Werbung. *Marketing ZFP,* Heft 4, IV Quartal, 261–267.

Rapp, R. & Wettler, M. (1991). Prediction of free word associations based on Hebbian learning. *Proceedings of the International Joint Conference on Neural Networks,* Singapore, Vol.1, 25-29.

Russell, W.A. (1970). The complete German language norms for responses to 100 words from the Kent-Rosanoff word association test. In: L. Postman & G. Keppel (eds.), *Norms of word association.* New York: Academic Press, 53–94.

Russell, W.A. & Meseck, O.R. (1959). Der Einfluß der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache. *Zeitschrift für experimentelle und angewandte Psychologie,* 6, 191–211.

Wettler, M. & Rapp, R. (1989). A connectionist system to simulate lexical decisions in information retrieval. In: Pfeifer, R., Schreter, Z., Fogelman, F. Steels, L. (eds.), *Connectionism in perspective.* Amsterdam: Elsevier, 463–469.

Wettler, M. & Rapp, R. (1993). Associative analysis of advertisements. Submitted to *Marketing and Research Today.*