The Computation of Associative Responses to Multiword Stimuli

Reinhard Rapp
Universitat Rovira i Virgili
Plaza Imperial Tarraco, 1
43005 Tarragona, Spain
reinhardrapp@gmx.de

Abstract

It is shown that the behaviour of test persons as observed in association experiments can be simulated statistically on the basis of the common occurrences of words in large text corpora, thereby applying the law of association by contiguity which is well known from psychological learning theory. In particular, the focus of this work is on the prediction of the word associations as obtained from subjects on presentation of multiword stimuli. Results are presented for applications as diverse as crossword puzzle solving and the identification of word translations based on non-parallel texts.

1 Introduction

The idea that human memory functions associatively goes back to Aristotle who formulated that the sequence of our memories is determined by the concepts of similarity and proximity (Strube, 1984:34). As early as 1879, Francis Galton tried to systematically observe human associative behaviour by introducing an association experiment. In this experiment, given a particular stimulus word, subjects had to respond with the first other word that occurred to them spontaneously. The resulting tables of associative responses are called association norms.

To explain the behavior documented in the association norms, in the literature a multiplicity of different mechanisms underlying human memory are proposed, thereby, for example, assuming phonological, morphological, syntactical, semantical, and contextual relations between words (Wettler, 1980). However, as yet there is no agreement whether these mechanisms should be considered of equal status, or if some may be derived from others.

Already in 1750 the physiologist David Hartley suggested that it may be possible to reduce the multiplicity of proposed association laws to only a single one based on temporal contiguity. This was formulated as one of the earliest psychological laws by William James (1890: 561): "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."

Assuming that the "objects" referred to in this law are words, the law of association by contiguity implies the following two phases:

1) *Learning phase:* When perceiving language, strong associative connections are developed between words that frequently occur in close temporal succession.
2) *Retrieval phase:* These associations determine the words that come to mind during generation. Only words that are strongly interconnected or have strong associations to external stimuli can be uttered or written down.

Pre-supposing the validity of the law of association, it should be possible to derive free word associations from the distribution of words in texts. Following Church & Hanks (1990), Rapp (2004), and Wettler et al. (2005) this actually seems to be successful. The recent simulation algorithms generate results which largely agree with the free word associations as found in the association norms. An example is shown in Table 1, where the observed and the simulated responses to the stimulus word *cold* are compared.

© 2008. Licensed under the *Creative Commons Attribution-Noncommercial-Share Alike 3.0 Unported* license (http://creativecommons.org/licenses/by-nc-sa/3.0/). Some rights reserved.

Coling 2008: Proceedings of the workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), pages 102–109
Manchester, August 2008

OBSERVED RESPONSE	NUMBER OF SUBJECTS	PREDICTED RESPONSE	NUMBER OF SUBJECTS
hot	34	hot	34
ice	10	winter	2
warm	7	weather	0
water	5	warm	7
freeze	3	water	5
wet	3	heat	1
feet	2	ice	10
freezing	2	wet	3
nose	2	wind	0
room	2	temperature	0
sneeze	2	shiver	0
sore	2	freeze	3
winter	2	rain	0

Table 1: Observed and predicted associative responses to the stimulus word *cold*.

When judging these results it should be kept in mind that among subjects there is some variation of responses. Therefore, the simulation results can be considered satisfactory if the difference between the predicted and the observed answers is on average not larger than the difference between an answer of an average test subject and the answers of the remaining test subjects.

In the current paper we try to build on these results. However, while most previous work considered only associations to individual stimulus words, the question to be dealt with here is whether the associative responses to several stimuli can likewise be predicted from the co-occurrences of words in texts. This is of considerable interest as all utterances and texts can be considered as accumulations of stimulus words, which together lead to a systematic activation of other words and concepts in the mind of the listener or reader.

How uniform the reactions of test subjects can be upon presentation of several stimulus words can be seen from examples like the word pairs *circus – laugh* or *King – girl* where subjects tend to think of *clown* and *princess*, respectively. Starting from the association norms for individual stimuli, the observed results are not always obvious. For example, in a large database of association norms, namely the *Edinburgh Associative Thesaurus* (Kiss et al., 1973), among the responses to *King* the word *princess* is completely missing, and the same is true for *girl*.

This means that the combination of stimulus words can lead to associations which are only weakly linked to the individual words and therefore cannot easily be deduced from conventional association norms. Accordingly it is not obvious whether the method used for the simulation of the associative behavior to single words can be extended in a straightforward way in the case of several stimulus words.

The organization of this paper is as follows: We first look at association norms collected for pairs of stimulus words. We then introduce a corpus-based algorithm that simulates the observed behavior which is applicable in the case of single or multiple stimuli. We then present some results of the algorithm and apply it to some related problems.

2 Association norms for word pairs

For individual English words, several association norms have been published, with the largest being the Edinburgh Associative Thesaurus. However, in the case of several stimulus words hardly any data seems to exist, with Rapp (1996, 1998) being an exception. This is a study that collected the responses of 31 subjects to pairs of German[2] nouns. In compiling these association norms, a list of 10 common German nouns had been selected, namely Mädchen (girl), Krankheit (illness), Junge (boy), Musik (music), Bürger (citizen), Erde (earth), Straße (street), König (King), Freude (joy), Sorge (worry). Then all 90 possible pairs of these words were constructed, and the answers of the subjects upon presentation of these pairs were collected. The subjects were asked to come up with the first word spontaneously coming to mind. In addition, associations to the individual words were also collected.

As for the pairs it turned out that word order did not have a noticeable effect on the responses, the responses to pairs differing only in word order were merged.

In Table 2 the associative responses as given by the test subjects for two sample pairs of stimulus words are listed. In comparison to responses to individual stimulus words, the responses to pairs of words are generally less uniform, i.e. there is considerably more variation in the case of word pairs. For example 25 of 31 test subjects come up with the association *Mädchen* (girl) given the stimulus word *Junge* (boy). In contrast, the most frequently mentioned associative response upon presentation of the stimulus pair *Junge Mädchen* (boy girl), which is *Kinder* (children), is given by only seven test persons.

[2] As we are not aware of such data for English, the current study was conducted for German, with translations given throughout the paper.

STIMULUS PAIR	ASSOCIATIVE RESPONSES
Erde (earth) Sorge (worry)	Umwelt (environment) 8, Umweltverschmutzung (environmental pollution) 5, Weltuntergang (end of the world) 2, ai (AI), Ausbeutung (exploitation), Katastrophen (catastrophe), Klimakatastrophe (climatical catastrophe), Krieg (war), Luft (air), Macht (might), Müll (garbage), Mutter (mother), Ozonloch (ozone hole), Resignation (resignation), Überbevölkerung (overpopulation), Umweltzerstörung (destruction of the environment), unfruchtbar (infertile), Verschmutzung (pollution), Zerstörung (destruction)
König (King) Mädchen (girl)	Prinzessin (princess) 15, Königin (queen) 3, Tochter (daughter) 2, Abhängigkeit (dependency), Dienerin (maid), Hochzeit (wedding), Kinderspiele (children's games), Kitsch (kitsch), Königspaar (royal couple), Märchen (fairy tale), Mißbrauch (abuse), Pferd (horse), Vater (father), Vorbild (model)

Table 2: Associations to the stimulus pairs "*Erde Sorge*" (earth worry) and "K*önig Mädchen*" (King girl). Figures indicate the number of subjects with the respective response, with the default being one.

For a more exact quantitative analysis of this observation a measure is needed for the homogeneity of the answers. For this purpose, it was computed how many subjects gave the same answer to a particular stimulus pair. On average, this was the case for 4% of the subjects. In comparison, the corresponding value for individual stimulus words is 15%. Thus the impression of a substantially larger homogeneity of the associative answers for individual stimuli is confirmed.

3 Simulation program

The simulation is based on the detection of statistical regularities of the common occurrences between the words in a large text corpus. As we did not have a large and at the same time balanced corpus of German at our disposal, we decided to use a corpus of the newspaper *Frankfurter Allgemeine Zeitung* (FAZ) comprising the years 1993 to 1996 (135 million words). As in the association experiment the subjects rarely answer with inflected forms or function words, for computational reasons we lemmatized this corpus (Lezius, Rapp & Wettler, 1998) and – based on a list of stop words – removed closed class words such as articles, pronouns, and particles.

To determine word co-occurrences, for each word in the corpus it was counted how often its close neighbors occurred within a text window of plus and minus six words. Assuming that approximately every second word is a function word, a window size of plus and minus six words after removal of the function words roughly corresponds to a window size of plus and minus 12 words without such pre-processing. This is a window size that corresponds with what had been found appropriate for the computation of associations in other studies (e.g. Rapp, 2004).

As the co-occurrence counts largely depend on overall word frequency, some association measure needs to be applied to eliminate this undesired influence. Many previous studies have shown that the log-likelihood ratio is well suited for this purpose (Dunning, 1993). It successfully eliminates word-frequency effects and emphasizes significant word pairs by comparing their observed co-occurrence counts with their expected co-occurrence counts. It can be expected that the log-likelihood ratio produces an accurate ranking of word pairs that highly correlates with human judgment (Dunning, 1993), although there are other measures which come close in performance (e.g. Rapp, 1998).

To compute the associations to pairs of stimulus words, it would in principle be possible to consider text positions where both stimulus words occur together, and to count the co-occurrence frequencies with their neighboring words. This would result in a three-dimensional association matrix whose first two dimensions correspond to the two stimulus words and whose third dimension corresponds to their associations. However, the problem of data sparseness would be very severe with such an approach, and it would not scale well if more than two stimulus words were considered.

We therefore propose another approach, which to our knowledge is novel in this context: The idea is that a potential associative response to a pair of stimulus words should have a strong and preferably symmetric associative connection to each of the stimulus words, and that a strong association to only one of them does not suffice. Such a behavior can usually be ensured by a multiplication.

However, we do not multiply the association strengths, as the log-likelihood ratio has an inappropriate (exponential) value characteristic. This value characteristic has the effect that a weak association to one of the stimuli can easily be overcompensated by a very strong association to

the other stimulus, which is not desirable. Instead of multiplying the association strengths, we therefore suggest to multiply their ranks. This improves the results considerably.

These considerations lead us to the following procedure: Given an association matrix of vocabulary *V* containing the log-likelihood ratios between all possible pairs of words, to compute the associative response given words *a* and *b*, the following steps are conducted:

1) For each word in *V* (by applying a search-and-compare operation on the association matrix) look up the ranks of words *a* and *b* in its list of associations, and compute the product of these ranks.
2) Sort the words in *V* according to these products, with the sort order such that the lowest value obtains the top rank (i.e. conduct a reverse sort).

Note that this procedure is somewhat time consuming as computations are required for each word in a large vocabulary.[3] On the plus side, the procedure is applicable to any number of stimulus words, and with increasing number of stimuli there is only a moderate increase in computational requirements. (The application presented in section 5.2 successfully processes 30 stimulus words.)

A minor issue is the assignment of ranks to words that have identical log-likelihood scores, especially in the frequent case of zero co-occurrence counts. In such cases, the assignment of possibly almost arbitrary ranks could adversely affect the results. We therefore suggest assigning corrected ranks, which are to be chosen as the average ranks of all words with identical scores.

With large numbers of stimuli, depending on the application it can be helpful to introduce a limit to the maximum rank, thereby reducing the effects of the sparse-data problem. The benefit of this measure is similar to smoothing, but more sophisticated smoothing methods can of course also be considered (as described, e.g. in Church & Gale, 1991). Note that for the current work we only used a rank limit of 10,000, but did not apply any sophisticated smoothing as this usually has little impact if the focus is mainly on the top ranks, as is the case here.

[3] Considerable time savings are possible by using an index of the non-zero co-occurrences.

4 Results

The algorithm as described above was applied to the FAZ corpus. That is, based on a window size of plus and minus six words, an association matrix with log-likelihood scores and (in both rows and columns) comprising all words with a corpus frequency of 200 or higher was computed. For each of the 45 word pairs, the top associations as resulting from the product of ranks were computed. To give some examples, the following tables show the outcome for a few stimulus pairs. Hereby, the columns in the tables have the following meanings:

1) rank
2) corpus frequency of association
3) score (product of stimulus ranks)
4) association

Junge Mädchen (boy girl)

1	247	11.33	fünfzehnjährig (15 year old)
2	2960	9.81	dreizehn (13)
3	398	9.72	gleichaltrig (same age)
4	86559	9.72	alt (old)
5	850	9.66	blond (blond)

Bürger Mädchen (citizen girl)

1	1276	11.51	brav (well behaved)
2	1268	7.26	unschuldig (innocent)
3	223	6.73	verängstigt (scared)
4	979	6.41	anvertrauen (to intrust)
5	362	5.97	belästigen (to molest)

Straße Mädchen (street girl)

1	2509	7.50	tanzen (to dance)
2	242	7.12	pflastern (to pave)
3	272	6.96	Bürgersteig (sidewalk)
4	529	6.87	Prostitution (prostitution)
5	4367	6.76	begegnen (to encounter)

Sorge Mädchen (worry girl)

1	317	7.03	elterlich (parental)
2	210	6.62	Burschen (fellows)
3	222	6.23	Beschneidung (concision)
4	7508	5.81	Eltern (parents)
5	271	5.77	zwölfjährig (12 year old)

Junge Krankheit (boy illness)

1	8891	7.33	leiden (to suffer)
2	3553	7.14	tödlich (lethal)
3	16468	7.04	sterben (to die)
4	423	6.83	Heilung (cure)
5	261	6.62	Schizophrenie (schizophrenia)

Straße Krankheit (street illness)

1	308	6.94	Tuberkulose (tuberculosis)
2	4704	6.74	Unfall (accident)
3	276	6.71	tückisch (malicious)
4	232	6.34	heimtückisch (malignant)
5	620	6.07	anstecken (to infect)

Straße Bürger (street citizen)

1	272	7.21	Bürgersteig (sidewalk)
2	235	7.18	Gibraltar (Gibraltar)
3	207	7.09	flanieren (to stroll)
4	242	7.02	pflastern (to pave)
5	366	6.58	Fußgängerzone (pedestrian zone)

Sorge Freude

1	6331	1.11	bereiten (to cause)
2	8747	9.21	Anlaß (occasion)
3	950	8.54	überwiegen (to outweigh)
4	27136	7.54	Grund (reason)
5	248	7.21	ungetrübt (untroubled)

If we look at all 45 word pairs, we obtain the following evaluation: Whereas an associative answer given by a subject is on average also given by 4% of the other subjects, only about 0.8% of the subjects give the answer produced in the simulation, i.e. the word ending up on the top rank. However, due to the low number of cases, this value may be subject to some sampling error.

A method less sensitive to sampling errors is to look at the overall simulation ranks of the subjects' responses. Hereby it is better to consider the median of the ranks rather than the mean, as the median's treatment of outliers is more appropriate. Note that when computing the median, associative responses given by n subjects obtain an n-fold higher weight. To further reduce the effects of outliers, only responses that are given by at least two subjects are taken into account.

Under these assumptions, the overall median (computed over all stimulus pairs) has a value of 245. With the total vocabulary of corpus frequency 200 and higher comprising about 25000 words, this value is at the 1% level. This compares to 12500 at the 50% level, which could be expected in the case of random behaviour.

5 Applications

5.1 Crossword puzzle solver

As crossword puzzles have definitions which usually consist of several words, the proposed algorithm can be applied as a crossword puzzle solver. In order not to reduce this task to a (for computers) relatively simple combinatorial problem, we hereby only restrict the ranked list of words as produced by the simulation program to those words that have the correct number of characters, but do not utilize as clues the common characters of horizontal and vertical words.

As an example, Figure 1 shows a crossword puzzle which is attributed to be the world's first one. It was designed by Arthur Wynne and published on December 21, 1913 in *The New York World*. Table 3 shows the definitions of this crossword puzzle together with the supposed solutions and the ranks of the respective words as computed by our algorithm based on three different corpora, namely the *British National Corpus* (BNC), the years 1990 to 1994 of the newspaper *The Guardian*, and the English part of the *Wikipedia XML Corpus* (Denoyer & Gallinari, 2006). These three corpora have a size of roughly 100, 150, and 300 million words, respectively. To allow a better judgment of the simulation results, the number of words of the respective length in the underlying vocabulary is specified in column 5.

This vocabulary was chosen to consist of all words that have a corpus frequency of 100 or higher in the BNC but did not occur in our list of about 200 function words. To this vocabulary, all words occurring in the crossword puzzle were added. The purpose of limiting the vocabulary was solely for computational reasons, as our algorithm is rather demanding with regard to both execution time and memory requirements.

Note that the BNC-based vocabulary was also used for the other somewhat larger corpora as not many words were missing there: In the Guardian corpus of the altogether 34,448 words all but 390 occurred at least one time, and in the larger Wikipedia corpus all but 131. We did not lemmatize the English corpora as in several cases inflected forms of words occurred in the definitions or in the solutions of the crossword puzzle.

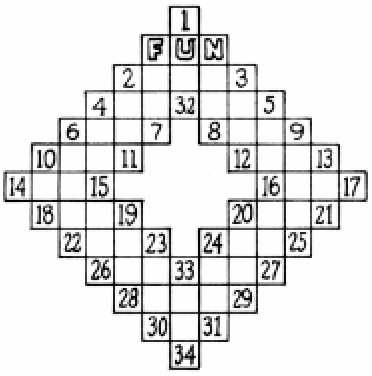

Figure 1: Crossword puzzle by Arthur Wynne.

As described in section 2, for counting the co-occurrences of words a window of plus and minus six words around a given word was considered, and for the computation of the associative strengths the log-likelihood ratio was used. Stop words were also removed from the corpora beforehand, but no lemmatization was conducted.

As many of the words used in the crossword puzzle are rare and several are outdated, solving this problem by a simulation is a non-trivial task. Nevertheless, for the Wikipedia corpus the algorithm got 8 of 31 answers ranked among the top five. When inspecting the examples that the algorithm got wrong, it appears that these are often the ones where humans would also have difficulties. For example, the solution "*side"* for "*to agree with*" got consistently poor ranks with all three corpora. On the other hand, rather surprisingly, the solution for "*such and nothing more*", namely "*mere*", received top rankings despite the fact that there are no salient content words in the description. This may be an indication that the algorithm grasps something that is related to cognitive processes. However, a similar example, namely "*what we all should be*" (→ *moral*) only obtains a reasonable ranking with the Wikipedia corpus. According to the average rankings (bottom line of Table 2), this corpus seems to be better suited for this task than the other two corpora.

5.2 Identifying word translations

The proposed core algorithm also has applications that may come somewhat unexpectedly. What we suggest here is to identify word transla-

POS.	DEFINITION	SOLUTION	LENGTH	WORDS OF THIS LENGTH	RANK BNC	RANK GUARDIAN	RANK WIKIPEDIA
2-3	what bargain hunters enjoy	sales	5	4254	1014	70	338
4-5	a written acknowledgement	receipt	7	5371	2	44	355
6-7	such and nothing more	mere	4	2916	16	17	4
10-11	a bird	dove	4	2916	17	87	4
14-15	opposed to less	more	4	2916	42	34	5
18-19	what this puzzle is	hard	4	2916	1486	115	384
22-23	an animal of prey	lion	4	2916	84	16	324
26-27	the close of a day	evening	7	5371	603	494	185
28-29	to elude	evade	5	4254	80	64	38
30-31	the plural of is	are	3	1424	238	119	412
8-9	to cultivate	farm	4	2916	2316	2783	1070
12-13	a bar of wood or iron	rail	4	2916	1658	1419	925
16-17	what artists learn to do	draw	4	2916	227	1437	86
20-21	fastened	tied	4	2916	15	2335	2078
24-25	found on the seashore	sand	4	2916	124	19	757
10-18	the fibre of the gomuti palm	doh	3	1424	585	279	711
6-22	what we all should be	moral	5	4254	4107	1163	51
4-26	a day dream	reverie	6	5371	489	572	2
2-11	a talon	sere	4	2916	676	803	492
19-28	a pigeon	dove	4	2916	36	8	1
F-7	part of your head	face	4	2916	63	20	143
23-30	a river in Russia	Neva	4	2916	174	413	3
1-32	to govern	rule	4	2916	48	9	13
33-34	an aromatic plant	nard	4	2916	616	2753	393
N-8	a fist	neif	4	2916	---	---	---
24-31	to agree with	side	4	2916	2836	2393	1387
3-12	part of a ship	spar	4	2916	2693	1932	90
20-29	one	tane	4	2916	2814	2773	2680
5-27	exchanging	trading	7	5371	3444	5216	2347
9-25	sunk in mud	mired	5	4254	3	2	1
13-21	a boy	lad	3	1424	3	2	2
AVERAGE RANK					891.6	922.3	520.2

Table 2: Crossword puzzle definitions and the computed ranks of their solutions based on three corpora. ('---' means that a solution does not occur in a corpus (not taken into account when computing average ranks).

tions from monolingual English and German corpora, i.e. from corpora that are not translations of each other (Rapp, 1999). This is a rather difficult task.

As our textual basis, for German we use the FAZ corpus as described above, with exactly the same pre-processing. For English we use a similarly sized corpus of the newspaper "The Guardian", with analogous pre-processing.

We apply a two-stage procedure to compute the translation of a source language word: First, by considering the log-likelihood ratios, its strongest source language associations are determined and translated to the target language using a small pocket dictionary. Hereby, associations that are missing in the dictionary are discarded, and of the remaining associations only the top 30 are selected.

The second step exactly corresponds to the computation of associations when given multiple stimulus words as described above. That is, for each word in the target language vocabulary (comprising all words that in the Guardian corpus occur with a frequency of 100 or higher) the ranks of the 30 translations are determined, and the product of these ranks is computed. The word obtaining the smallest value for the product is considered to be the translation of the source language word. This algorithm turned out to be a significant improvement over the previous algorithm described in Rapp (1999) as it provides a better accuracy and at the same time a considerably higher robustness.

Based on this novel algorithm, a large dictionary for German to English was computed. As for the translation of the source language vectors a base dictionary is required, we adapted for this purpose a small Collins pocket dictionary which comprised in the order of 20 000 entries. In essence, the adaptation procedure involves deriving word equations from the dictionary, each consisting of the source word and its first translation as mentioned in the dictionary.

To give an impression of the results, the following tables show the top ten computed translations for the six words *Historie* (history), *Leibwächter* (bodyguard), *Raumfähre* (space shuttle), *spirituell* (spiritual), *ukrainisch* (Ukranian), and *umdenken* (rethink). Hereby, the columns have the following meanings:

1) Rank of a potential translation
2) Corpus frequency of translation
3) Score assigned to translation
4) Computed translation

Historie (history)

1	29453	13.73	history
2	4997	12.87	literature
3	4758	8.74	historical
4	2670	0.67	essay
5	6969	0.11	contemporary
6	18909	-1.72	art
7	18382	-2.81	modern
8	15728	-4.31	writing
9	1447	-5.52	photography
10	2442	-5.53	narrative

Leibwächter (body guard)

1	949	40.02	bodyguard
2	5619	23.34	policeman
3	2535	8.18	gunman
4	26347	3.69	kill
5	9180	2.92	guard
6	401	-0.56	bystander
7	815	-1.24	POLICE
8	8503	-2.33	injured
9	2973	-3.23	stab
10	1876	-3.58	murderer

Raumfähre (space shuttle)

1	1259	46.20	shuttle
2	666	26.25	Nasa
3	473	25.95	astronaut
4	287	25.76	spacecraft
5	1062	16.92	orbit
6	16086	11.72	space
7	525	9.50	manned
8	125	7.69	cosmonaut
9	254	5.24	mir
10	7080	3.70	plane

spirituell (spritual)

1	2964	56.10	spiritual
2	1380	8.34	Christianity
3	7721	8.08	religious
4	9525	4.10	moral
5	1414	0.63	secular
6	5685	0.06	emotional
7	4678	-1.04	religion
8	6447	-1.49	intellectual
9	8749	-2.25	belief
10	8863	-4.07	cultural

ukrainisch (Ukrainian)

1	1753	50.69	Ukrainian
2	22626	39.88	Russian
3	3205	29.25	Ukraine
4	34572	23.63	Soviet

5	978	21.13	Lithuanian
6	1005	18.88	Kiev
7	10968	15.07	Gorbachev
8	10209	14.51	Yeltsin
9	16616	13.38	republic
10	502	11.71	Latvian

umdenken (rethink)

1	1119	20.76	rethink
2	248	15.46	reassessment
3	84109	13.39	change
4	12497	12.13	reform
5	236	10.00	reappraisal
6	9220	9.97	improvement
7	5212	9.48	implement
8	1139	8.25	overhaul
9	13550	7.89	unless
10	9807	7.88	immediate

6 Summary

It could be shown that word associations to multiple stimuli as collected from test persons can be predicted with reasonable accuracy using a simulation program that analyzes the co-occurrences of words in texts.

This result makes the automatic construction of an associative thesaurus of responses to multiple stimuli feasible. Note that such a thesaurus could not realistically be compiled by collecting the responses of human subjects as there are too many possible combinations of stimuli.

Finally, by looking at two sample applications we showed the pracical utility of the method. In principle, there should be many more applications, as all utterances and texts can be considered as collections of stimulus words. A notable one is search word generation in the context of internet search engines.

Of course, all existing algorithms for speech and text processing, although often not claiming any cognitive plausibility, necessarily also have some implicit mechanisms that deal with multiword stimuli. We nevertheless hope that the specific perspective that we presented here may add to a better understanding of the underlying cognitive mechanisms, and that it offers a systematic way of approaching these challenges.

7 Acknowledgments

This research was in part supported by a Marie Curie Intra European Fellowship within the 6th European Community Framework Programme.

References

Church, Kenneth W.; Gale, William (1991). A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5(1), 19–54.

Church, Kenneth W., Hanks, Patrick (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Denoyer, Ludovic; Gallinari, Pattrick (2006). The Wikipedia XML Corpus. *SIGIR Forum*, 40(1), 64–69.

Dunning, Ted (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Galton, Francis. (1879). Psychometric experiments. *Brain,* 1, 149–162.

James, William (1890). *The Principles of Psychology*. New York: Dover Publications.

Kiss, George R.; Armstrong, Christine; Milroy, Robert; Piper, James (1973). An associative thesaurus of English and its computer analysis. In: A. Aitken, R. Beiley, N. Hamilton-Smith (eds.): *The Computer and Literary Studies*. Edinburgh University Press.

Lezius, Wolfgan; Rapp, Reinhard; Wettler, Mafred (1998). A freely available morphological analyzer, disambiguator, and context sensitive lemmatizer for German. *Proceedings of COLING ACL 1998*, Montreal, 743–747.

Rapp, Reinhard (1996). *Die Berechnung von Assoziationen: ein korpuslinguistischer Ansatz.* Hildesheim: Olms.

Rapp, Reinhard (1998). Das Kontiguitätsprinzip und die Simulation des Assoziierens auf mehrere Stimuluswörter. In: B. Schröder, W. Lenders, W. Hess, T. Portele: *Computer, Linguistik und Phonetik zwischen Sprache und Sprechen.* Frankfurt am Main: Peter Lang, 261–272.

Rapp, Reinhard (1999). Automatic identification of word translations from unrelated English and German corpora. In: *Proceedings of the 37th Annual Meeting of the ACL,* College Park, MD, 519–526.

Rapp, Reinhard (2004). Word Sense Induction as Statistical Pattern Recognition. In: Ernst Buchberger (ed.): *Tagungsband der 7. Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, Universität Wien, 161–168.

Strube, Gerhard (1984). *Assoziation.* Berlin: Springer.

Wettler, Manfred (1980). *Sprache, Gedächtnis, Verstehen*. Berlin: de Gruyter.

Wettler, Manfred; Rapp, Reinhard; Sedlmeier, Peter (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics*, 12(2), 111–122.