Int J Speech Technol (2008) 11: 147–156
DOI 10.1007/s10772-009-9043-7

The automatic generation of thesauri of related words for English, French, German, and Russian

Reinhard Rapp

Received: 15 July 2009 / Accepted: 8 October 2009 / Published online: 29 October 2009
© Springer Science+Business Media, LLC 2009

Abstract A method for the automatic extraction of words with similar meanings is presented which is based on the analysis of word distribution in large monolingual text corpora. It involves compiling matrices of word co-occurrences and reducing the dimensionality of the semantic space by conducting a singular value decomposition. This way problems of data sparseness are reduced and a generalization effect is achieved which considerably improves the results. The method is largely language independent and has been applied to corpora of English, French, German, and Russian, with the resulting thesauri being freely available. For the English thesaurus, an evaluation has been conducted by comparing it to experimental results as obtained from test persons who were asked to give judgements of word similarities. According to this evaluation, the machine generated results come close to native speaker's performance.

Keywords Thesaurus generation · Word similarity · Singular value decomposition · Lexical acquisition · Knowledge discovery from text · Corpus linguistics

1 Introduction

In his paper "Distributional Structure" Zelig S. Harris (1954) hypothesized that words that occur in the same contexts tend to have similar meanings. This finding is often referred to as *distributional hypothesis*. It was put into practice by Ruge (1992) who showed that the semantic similarity of two words can be computed by looking at the agreement of their lexical neighborhoods. For example, as illustrated in Fig. 1, a certain degree of semantic similarity between the words *red* and *blue* can be derived from the fact that they both frequently co-occur with words like *color, dress, flower*, etc. although there are other context words that only occur with one of them. If on the basis of a large corpus a matrix of word co-occurrences is compiled, then the semantic similarities between words can be determined by comparing the vectors in the matrix. This can be done using any of the standard vector similarity measures such as the cosine coefficient.

Since Ruge's pioneering work, many researchers, e.g. Schütze (1997), Rapp (2002), and Turney (2006) used this type of distributional analysis as a basis to determine semantically related words. An important characteristic of some algorithms (e.g. Ruge 1992; Grefenstette 1994, and Lin 1998a) is that they parse the corpus and only consider co-occurrences of word pairs that are in a certain relation to each other, e.g. a head-modifier, verb-object, or subject-object relation. Others do not parse but perform a *singular value decomposition* (SVD) on the co-occurrence matrix, which also could be shown to improve results (Landauer and Dumais 1997; Landauer et al. 2007). As an alternative to the SVD, Sahlgren (2001) uses *random indexing* which is computationally less demanding.

In the current paper we use an improved variant of the Landauer and Dumais (1997) algorithm as described in Rapp (2007). It does not perform a syntactic analysis but reduces the dimensionality of the semantic space by performing an SVD. The program is applied to corpora of four languages, namely English, French, German, and Russian, and large thesauri of related words are generated for each of these languages.

The rest of the paper is organized as follows: Section 2 introduces an evaluation method that has often been used

R. Rapp (✉)
University of Tarragona, Tarragona, Spain
e-mail: reinhard.rapp@urv.cat

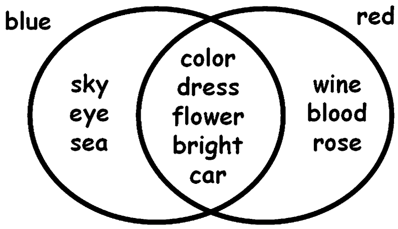

Fig. 1 Words co-occurring with *red* and *blue*

to measure the quality of automatic methods for synonym extraction. Section 3 presents an overview on the related literature and the current state of the art. In Section 4 the English, French, German, and Russian corpora that were used in the current work are introduced. Section 5 describes our algorithm which is a modified version of Latent Semantic Analysis as described in Landauer and Dumais (1997). Sections 6 and 7 present, evaluate, and discuss our results. And finally Section 8 summarizes our findings and gives an outlook on future work.

2 TOEFL synonym data for evaluation

As described in Rapp (2004), it is desirable to evaluate the results of the different algorithms for computing semantic similarity. For doing so, many possibilities can be thought of and have been applied in the past. For example, Grefenstette (1994:81) used available dictionaries as a gold standard, Lin (1998a) compared his results to WordNet, and Landauer and Dumais (1997) used experimental data taken from the synonym portion of the *Test of English as a Foreign Language* (TOEFL). As described by Turney (2006), the advantage of the TOEFL data is that it has gained considerable acceptance among researchers. Turney (2001) also introduced a similar but somewhat smaller test set (50 rather than 80 test items), namely the ESL (*English as a Second Language*) synonym data, which, however, is less widely used. Therefore, in this paper we concentrate on the TOEFL synonym data as available from Thomas K. Landauer.[1]

The TOEFL is an obligatory test for non-native speakers of English who intend to study at a university with English as the teaching language. The data used by Landauer and Dumais had been acquired from the *Educational Testing Service* and comprises 80 test items. Each item consists of a problem word embedded in a sentence and four alternative words, from which the test taker is asked to choose the one with the most similar meaning to the problem word. For example, given the test sentence "*Both boats and trains are used for transporting the materials*" and the four alternative words *planes*, *ships*, *canoes*, and *railroads*, the subject would be expected to choose the word *ships*, which is supposed to be the one most similar to *boats*.

Landauer and Dumais (1997) found that their algorithm for computing semantic similarities between words has a success rate comparable to the human test takers when applied to the TOEFL synonym test. Whereas the algorithm got 64.4% of the questions right (i.e. the correct solution obtained the best rank among the four alternative words), the average success rate of the human subjects was 64.5%.[2] Other researchers were able to improve the performance to 69% (Rapp 2002), 72% (Sahlgren 2001), 74% (Turney 2001) and 81.25% (Terra and Clarke 2003). This gives the impression that the quality of the simulation is above human level.

However, it has sometimes been overlooked that the 64.5% performance figure achieved by the test takers relates to non-native speakers of English, and that native speakers perform significantly better. On the other hand, the simulation programs are usually not designed to make use of the context of the test word, so they neglect some information that may be useful for the human subjects.

In order to approach both issues, we presented a test sheet with the TOEFL test words, together with the alternative words, but without the sentences, to five native and five non-native speakers of English, drawn from staff members of Macquarie University in Sydney (Rapp 2004). We asked them to select among the alternative words the one that, according to their personal judgment, was closest in meaning to the given word. Two of the native speakers got all 80 items correct, another two got 78 correct, and one got 75 correct. As expected, the performance of the non-native speakers was considerably worse. Their numbers of correct choices were 75, 70, 69, 67, and 66.

On average, the performance of the native speakers was 97.75%, whereas the performance of the non-native speakers was 86.75%. Remember that the performance of the non-native speakers in the TOEFL test, although they had the context of each test word as an additional clue, was only 64.5%. The discrepancy of more than 20% between our non-native speakers and the TOEFL test takers can be explained by the fact that most of our subjects had spent many years in English speaking countries and thus had a language proficiency far above average. More importantly, our native speakers' results indicate that the performance of the above mentioned algorithms is clearly below human performance. So the impression from the Landauer and Dumais (1997) paper that human-like quality has been obtained is obviously wrong unless one only looks at second language learners with a relatively poor proficiency.

[1] http://www.pearsonkt.com/bioLandauer.shtml.

[2] This performance figure was provided by the Educational Testing Service, with the number of test takers being unknown.

Table 1 Comparison of lexicon-based approaches

Reference for TOEFL experiment	Score	Information on algorithm
Jarmasz and Szpakowicz (2003)	21.88%	Leacock and Chodrow (1998)
Jarmasz and Szpakowicz (2003)	77.91%	Hirst and St-Onge (1998)
Jarmasz and Szpakowicz (2003)	78.75%	Jarmasz and Szpakowicz (2003)

Table 2 Comparison of corpus-based approaches

Reference for TOEFL experiment	Score	Information on algorithm
Landauer and Dumais (1997)	64.38%	Latent semantic analysis
Rapp (2002)	69.00%	Raw co-occurrences with city-block metric
Pado and Lapata (2007)	73.00%	Dependency space
Turney (2001)	73.75%	Pointwise mutual information
Turney (2008)	76.25%	PairClass
Terra and Clarke (2003)	81.25%	Pointwise mutual information
Ruiz-Casado et al. (2005)	82.55%	Context window overlapping
Bullinaria and Levy (2007)	85.00%	Positive pointwise mutual information with cosine
Matveeva et al. (2005)	86.25%	Generalized latent semantic analysis
Rapp (2004)	90.90%	Pantel and Lin (2002)
Approach described here	92.50%	Modified latent semantic analysis

As we did not have any data comparable to the TOEFL synonym test for French, German, and Russian, the evaluation was only conducted for English.

3 State of the art on solving TOEFL synonym questions

In the literature, there are essentially three basic approaches to automatically solve the TOEFL synonym questions. One is lexicon-based, another is corpus-based, and the third, which is usually referred to as hybrid, is a mixture of the first two.

With the lexicon-based approaches, a given word is looked up in a large lexicon of synonyms and it is determined whether there is a match between any of the retrieved synonyms and the four alternative words presented in the TOEFL question. If there is a match, the respective word is considered to be the solution to the question.

This procedure works rather well if the lexicon has a good coverage of the respective vocabulary. So in the literature typically very large dictionaries such as WordNet (Fellbaum 1998) have been used. On the other hand, both the TOEFL questions and the lexicons are hand crafted and therefore reflect human intuitions. So it is not surprising that a high correspondence between these two closely related types of human intuitions can be observed. Therefore, although the lexicon-based approach is of practical relevance, in this paper we concentrate on the second approach which has been sketched in Section 1. It is a corpus-based machine learning approach which appears to be more interesting from a cognitive perspective as it probably better resembles some aspects of human language acquisition.

The third approach (hybrid) is in essence a fall-back strategy for the first approach: That is, by default the lexicon-based approach is used as its results tend to be more reliable. However, if the relevant words can not be found in the lexicon, then it is of course better to also take indirect synonyms into account or to fall back to the corpus-based approach rather than to guess randomly.

Tables 1 to 3[3] list the results on the TOEFL questions as found in the literature for the three types of algorithms. To allow a better judgment of the results, Table 4 shows a number of relevant baselines. It should be noted that with regard to the performance figures given in the tables a bit of caution is in order. Firstly, the figures are based on corpora of very different sizes and nature, secondly, some algorithms make distinctions between various parts of speech while others do not, and thirdly some but not all of the algorithms have been optimized using the TOEFL test set. Finally, it can be argued that with its 80 questions the TOEFL synonym test is rather small and therefore susceptible to statistical variation. Also, it was not designed to measure strengths and weaknesses of the various algorithms with regard to particular properties of the input words, e.g. their frequency, part of speech, or ambiguity.

[3] Part of this information has been adapted from the ACL wiki as of July 10, 2009: http://aclweb.org/aclwiki/index.php?title=State_of_the_art.

Table 3 Comparison of hybrid approaches

Reference for TOEFL experiment	Score	Information on algorithm
Jarmasz and Szpakowicz (2003)	20.31%	Resnik (1995)
Jarmasz and Szpakowicz (2003)	24.06%	Lin (1998b)
Jarmasz and Szpakowicz (2003)	25.00%	Jiang and Conrath (1997)
Turney et al. (2003)	97.50%	Product Rule

Table 4 Baselines for judgment of TOEFL scores

Reference	Score	Description
Rapp (2004)	25.00%	Random guessing. Four alternatives per given word, of which one is correct
Landauer and Dumais (1997)	64.50%	Average non-English US college applicant taking TOEFL
Rapp (2004)	86.75%	Non-native speakers of English living in Australia
Rapp (2004)	97.75%	Native speakers of English living in Australia

4 Corpora and corpus pre-processing

Since our algorithm is based on a similarity measure relying on co-occurrence data, corpora are required from which the co-occurrence counts can be derived. If—as in this case—a measure for the success of the system is the results' plausibility to human judgment, it is advisable to use corpora that are as typical as possible for the language environment of native speakers.

For English, we chose the *British National Corpus* (BNC), a 100-million-word corpus of written and spoken language that was compiled with the intention of providing a balanced sample of British English (Burnard and Aston 1998). As for the German corpus, due to lack of a balanced corpus, we used 135 million words of the newspaper *Frankfurter Allgemeine Zeitung* (years 1993 to 1996). For French, only small newspaper corpora were available, so a corpus comprising the French version of "Wikipedia" (Denoyer and Gallinari 2006) and "ABU—La Bibliothéque Universelle"[4] (together about 70 million words) were acquired by downloading them from the Internet. For each of these corpora, a specific cleanup-program had to be written and applied. For Russian, we used the *Russian Reference Corpus*, a corpus of about 50 million words.[5]

Since these corpora are relatively large, to save disk space and processing time we decided to remove all function words from the texts. This was done on the basis of a list of approximately 600 German, 500 French, and another list of about 200 English function words. These lists were compiled by looking at the closed class words (mainly articles, pronouns, and particles) in morphological lexica (for details see Lezius et al. 1998) and at word frequency lists derived from our corpora. In the case of the Russian Reference Corpus, part-of-speech information was given for all words, and function words were removed on this basis. By eliminating function words, we assumed we would lose little information: Function words are often highly ambiguous and their co-occurrences are mostly based on syntactic rather than semantic patterns.

We also decided to lemmatize our corpora. Since we were interested in the similarities of base forms only, it was clear that lemmatization would be useful as it reduces the sparse-data problem. For English and German we conducted a partial lemmatization procedure that was based only on a morphological lexicon and did not take the context of a word form into account. This means that we could not lemmatize those ambiguous word forms that can be derived from more than one base form. However, this is a relatively rare case. (According to Lezius et al. 1998, 93% of the tokens of a German text had only one lemma.) For lemmatization of French we used the context sensitive lemmatization function of the TreeTagger (Schmid 1995) with the parameter file provided by Achim Stein.[6] In the case of Russian, no special processing was necessary since lemma information was already given in the corpus as provided by Serge Sharoff.

5 Algorithm

Our algorithm is a modified version of Landauer and Dumais (1997) and consists of the following four steps (see also Rapp 2003, 2004):

1. Counting word co-occurrences
2. Applying an association measure to the raw co-occurrence counts

[4] http://abu.cnam.fr/.

[5] See http://corpus.leeds.ac.uk/serge/bokrcorpora/index-en.html. This corpus was kindly provided by Serge Sharoff.

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html.

3. Dimensionality reduction using singular value decomposition
4. Computing vector similarities

In the following subsections we exemplify the application of this algorithm for the case of English, i.e. based on the British National Corpus. However, its application on the French, German, and Russian corpora is straightforward and minor differences are described together with the results in Section 6.

5.1 Counting word co-occurrences

For counting word co-occurrences, as in most other studies a fixed window size is chosen and it is determined how often each pair of words occurs within a text window of this size. Choosing a window size usually means a trade-off between two parameters: specificity versus the sparse-data problem. The smaller the window, the more salient the associative relations between the words inside the window, but the more severe the sparse data problem. In this work we chose a window size of ±2 words, which on first glance may look rather small. However, this can be justified since we have reduced the effects of the sparse data problem by using a large corpus and by lemmatizing the corpus. It also should be noted that a window size of ±2 applied after elimination of the function words is comparable to a window size of ±4 applied to the original texts (assuming that roughly every second word is a function word).

Based on the window size of ±2, we computed the co-occurrence matrix for the corpus. By storing it as a sparse matrix, it was feasible to include all of the approximately 375 000 lemmas occurring in the BNC.

5.2 Applying an association measure

Although semantic similarities can be successfully computed based on raw word co-occurrence counts, the results can be improved when the observed co-occurrence-frequencies are transformed by some function that reduces the effects of different word frequencies. For example, by applying a significance test that compares the observed co-occurrence counts with the expected co-occurrence counts (e.g. the log-likelihood ratio as proposed by Dunning 1993) significant word pairs are strengthened and incidental word pairs are weakened. Other measures applied successfully include TF/IDF and mutual information. In the remainder of this paper, we refer to co-occurrence matrices that have been transformed by such a function as *association matrices*. However, in order to further improve similarity estimates, in this study we apply a *singular value decomposition* (SVD) to our association matrices (see Section 5.3). To our surprise, our experiments clearly showed that the log-likelihood ratio, which was the transformation function that gave good similarity estimates without SVD (Rapp 2002), was not optimal when using SVD. Following Landauer and Dumais (1997), we found that with SVD some entropy-based transformation function gave substantially better results than the loglikelihood ratio. This is the formula that we use:

$$A_{ij} = \log(1 + f_{ij}) \cdot \left(-\sum_k p_{kj} \log(p_{kj}) \right)$$

$$\text{with } p_{kj} = \frac{f_{kj}}{c_j}.$$

Hereby f_{ij} is the co-occurrence frequency of words i and j and c_j is the corpus frequency of word j. Indices i, j, and k all have a range between one and the number of words in the vocabulary n. The right term in the formula (sum) is entropy. As usual with entropy, it is assumed that $0 \log(0) = 0$. The entropy of a word reaches its maximum of $\log(n)$ if the word co-occurs equally often with all other words in a vocabulary, and it reaches its minimum (zero) if it co-occurs only with a single other word.

Let us now look at how the formula works. The important part is taking the logarithm of f_{ij} thus dampening the effects of large differences in frequency. Adding 1 to f_{ij} provides some smoothing and prevents the logarithm from becoming infinite if f_{ij} is zero. A relatively modest, but noticeable improvement (in the order of 5% when measured using the TOEFL-data) can be achieved by multiplying this by the entropy of a word. This has the effect that the weights of rare words that have only few (and often incidental) co-occurrences are reduced.

Note that this is in contrast to Landauer and Dumais (1997) who suggest not to multiply but to divide by entropy. The reasoning is that words with a salient co-occurrence distribution should have stronger weights than words with a more or less random distribution. However, as shown empirically, in our setting multiplication leads to clearly better results than division.

5.3 Singular value decomposition

Landauer and Dumais (1997) showed that the results can be improved if before computing semantic similarities the dimensionality of the association matrix is reduced. An appropriate mathematical method to do so is singular value decomposition. As this method is rather sophisticated, we can not go into the details here. A good description can be found in Landauer and Dumais (1997). The essence is that by computing the Eigenvalues of a matrix and by truncating the smaller ones, SVD allows to significantly reduce the number of columns, thereby (in a least squares sense) optimally preserving the Euclidean distances between the lines (Schütze 1997:191).

For computational reasons, we were not able to conduct the SVD for a matrix of all 374,244 lemmas occurring in the

Table 5 Semantic similarities for some English words as computed. The lists are ranked according to the cosine coefficient

enormously	greatly (0.52), immensely (0.51), tremendously (0.48), considerably (0.48), substantially (0.44), vastly (0.38), hugely (0.38), dramatically (0.35), materially (0.34), appreciably (0.33)
flaw	Shortcomings (0.43), defect (0.42), deficiencies (0.41), weakness (0.41), fault (0.36), drawback (0.36), anomaly (0.34), inconsistency (0.34), discrepancy (0.33), fallacy (0.31)
issue	question (0.51), matter (0.47), debate (0.38), concern (0.38), problem (0.37), topic (0.34), consideration (0.31), raise (0.30), dilemma (0.29), discussion (0.28)
build	building (0.55), construct (0.48), erect (0.39), design (0.37), create (0.37), develop (0.36), construction (0.34), rebuild (0.34), exist (0.29), brick (0.27)
discrepancy	disparity (0.44), anomaly (0.43), inconsistency (0.43), inaccuracy (0.40), difference (0.36), shortcomings (0.35), variance (0.34), imbalance (0.34), flaw (0.33), variation (0.33)
essentially	primarily (0.50), largely (0.49), purely (0.48), basically (0.48), mainly (0.46), mostly (0.39), fundamentally (0.39), principally (0.39), solely (0.36), entirely (0.35)

BNC. Therefore, we restricted our vocabulary to all lemmas with a BNC frequency of at least 20. To this vocabulary all problem and alternative words occurring in the TOEFL synonym test were added. This resulted in a total vocabulary of 56,491 words. In the association matrix corresponding to this vocabulary all 395 lines and 395 columns that contained only zeroes were removed which led to a matrix of size 56,096 by 56,096.

By using a version of Mike Berry's SVDPACK[7] software that had been modified by Hinrich Schütze, we transformed the 56,096 by 56,096 association matrix to a matrix of 56,096 lines and 300 columns. This smaller matrix has the advantage that all subsequent similarity computations tend to be considerably faster.[8] As discussed in Landauer and Dumais (1997), the process of dimensionality reduction, by combining similar columns (relating to words with similar meanings), is believed to perform a kind of generalization that is hoped to improve similarity computations (even critics concede at least a smoothing effect).

5.4 Computation of semantic similarity

The computation of the semantic similarities between words is based on comparisons between their dimensionality reduced association vectors. Our experience is that the sparse data problem is usually by far not as severe for the computation of vector similarities (second-order dependency) as it is—for example—for the computation of mutual information (first-order dependency). The reason is that for the computation of vector similarities a large number of association values are taken into account, and although each value is subject to a sampling error, these errors tend to cancel out over the whole vector. Since association measures such as mutual information usually only take a single association value into account, this kind of error reduction cannot be observed here.

For vector comparison, among the many similarity measures found in the literature, we decided to use the cosine coefficient. The cosine coefficient computes the cosine of the angle between two vectors.

6 Results and evaluation

The processing of the French, German, and the Russian corpora was done in analogy to English as described above. That is, the size of vocabulary has been chosen to be in the same order of magnitude,[9] and all parameters remained the same. Only for French and Russian the number of dimensions was chosen to be 250 instead of 300. The reason is that the French and Russian corpora are considerably smaller than the English and the German ones. They therefore carry less information and require fewer dimensions.[10]

To give a first impression, Table 5 shows the top most similar words to a few English examples as computed using SVD, the cosine-coefficient, and a vocabulary of 56,096 words. Although these results look plausible, a quantitative evaluation is always desirable. For this reason we used our system for solving the TOEFL synonym test and compared the results to the correct answers as provided by the Educational Testing Service, which had been lemmatized in the same way as the English corpus. Remember that the subjects had to choose the word most similar to a given stimulus word from a list of four alternatives. In the simulation, we assumed that the system made the right decision if the correct answer was ranked best among the four alternatives. This was the case for

[7] http://www.netlib.org/svdpack/.

[8] This is subject to details of implementation. As SVD transformed matrices tend to show very little data sparseness, algorithms that take advantage of this property may only be effective when applied to the original matrices.

[9] For French the 37,362 words with a corpus frequency above 49 were chosen, for German the 39,745 words with a corpus frequency above 99, and for Russian the 57,058 words with a frequency above 19.

[10] After removal of the function words, the English corpus contained 50,486,400 tokens, the French corpus 27,224,905, the German corpus 59,307,629, and the Russian corpus 42,792,750.

Table 6 Semantic similarities for some French words as computed. For each word, an English translation is given in square brackets

colonialiste [colonialist]	nationaliste [nationalist] (0.28), libéral [liberal] (0.27), communisme [Communism] (0.27), communiste [Communist] (0.27), colonisation [colonization] (0.26), impérialiste [imperialist] (0.26), critiquer [to criticize] (0.26), révolutionnaire [revolutionist] (0.26), anti [anti] (0.26), démocratie [democracy] (0.26)
dérapage [to skid]	accident [accident] (0.26), révéler [to reveal] (0.24), rupture [break] (0.24), apparition [appearance] (0.23), changement [change] (0.23), déplacement [displacement] (0.23), violent [violent] (0.23), conflit [conflict] (0.23), prévoir [envisage] (0.23), catastrophe [catastrophe] (0.23)
ingrédient [ingredient]	recette [receipt] (0.56), sucre [sugar] (0.53), beurre [butter] (0.53), farine [flour] (0.53), pâte [pastry] (0.53), lait [milk] (0.52), préparation [preparation] (0.52), fromage [cheese] (0.52), crème [cream] (0.51), légume [vegetable] (0.51)
microbe [microbe]	bactérie [bacterium] (0.36), pathogène [pathogenic] (0.36), virus [virus] (0.32), infecter [to infect] (0.29), organisme [organism] (0.29), parasite [parasite] (0.29), substance [substance] (0.29), germe [germ] (0.29), maladie [disease] (0.28), végétal [vegetable] (0.27)
réglementation [regulation]	législation [legislation] (0.55), règlement [regulation] (0.54), règle [rule] (0.53), norme [norm] (0.52), procédure [procedure] (0.52), légal [legal] (0.52), sécurité [safety] (0.51), contrôle [control] (0.51), marché [market] (0.51), protection [protection] (0.51)
prairie [meadow]	forêt [forest] (0.61), plaine [plain] (0.59), arbre [tree] (0.56), montagne [mountain] (0.55), rivière [river] (0.55), colline [hill] (0.55), désert [desert] (0.55), bois [wood] (0.54), jardin [garden] (0.54), humide [humid] (0.53)

Table 7 Semantic similarities for some German words as computed. For each word, an English translation is given in square brackets

ärgerlich [angry]	peinlich [embarrassing] (0.51), bedauerlich [regrettable] (0.47), unangenehm [unpleasant] (0.45), empörend [infuriating] (0.44), bedenklich [precarious] (0.42), ärgern [to annoy] (0,41), unverständlich [incomprehensible] (0.39), nerven [to annoy] (0.39), skandalös [scandalous] (0.39), deprimieren [to depress] (0.38)
Darsteller [performer]	Schauspieler [actor] (0.64), Regisseur [director] (0.57), Sänger [vocalist] (0.56), Hauptrolle [leading part] (0.55), inszenieren [to stage-manage] (0.49), Filmemacher [moviemaker] (0.49), Tänzer [dancer] (0.48), Choreograph [choreographer] (0.47), Komödie [comedy] (0.47), filmen [filming] (0.45)
dennoch [nevertheless]	gleichwohl [nonetheless] (0.76), trotzdem [although] (0.64), indes [however] (0,60), zwar [indeed] (0.59), allerdings [but] (0.57), deshalb [therefore] (0.56), immerhin [anyhow] (0.56), freilich [sure enough] (0.55), indessen [meanwhile] (0.54), zudem [furthermore] (0.51)
Gesang [singing]	Lied [song] (0.62), singen [to sing] (0.53), Klang [sound] (0.52), Melodie [melody] (0.49), Musik [music] (0.47), Trommeln [drumming] (0.46), Orgel [organ] (0.46), Hymnus [hymn] (0.46), Arie [aria] (0.45), Ballade [ballad] (0.45)
Magistrat [magistrate]	Stadtparlament [city parliament] (0.60), Stadtverordnetenversammlung [city council meeting] (0.58), Stadtrat [city council] (0.55), Stadtverordnete [city councillor] (0.48), Gemeinderat [municipal council] (0.48), Abgeordnetenhaus [house of representatives] (0.46), Bürgermeisterin [mayoress] (0.46), Kreistag [district council] (0.45), Bürgerschaft [citizenship] (0.41), Stadtoberhaupt [mayor] (0.41)
Spott [ridicule]	Häme [malice] (0.54), Empörung [outrage] (0.46), Neid [enviousness] (0.43), Mitleid [compassion] (0.42), Unverständnis [lack of understanding] (0.41), Polemik [polemic] (0.41), Wut [fury] (0.41), Zorn [anger] (0.41), ironisch [ironical] (0.39), Unmut [displeasure] (0.39)

74 of the 80 test items which gives us an accuracy of 92.5%. In comparison, recall that the performance of our human subjects had been 97.75% for the native speakers and 86.75% for our highly proficient non-native speakers. This means our program's performance is in between these two levels with about equal margins towards both sides.

Results analogous to Table 5 are given for French, German, and Russian in Tables 6 to 8. To make the interpretation easier, for all words appearing in the tables translations are provided. For the Russian data this has been kindly

Table 8 Semantic similarities for some Russian words as computed. For each word, an English translation is given in square brackets

festivalq [festival]	koncert [concert] (0.62), vystavka [exhibition] (0.58), spektaklq [performance, show] (0.57), teatr [theater] (0.52), balet [ballet] (0.51), konkurs [competition] (0.49), forum [forum] (0.49), opera [opera] (0.49), gastrolq [(performance) tours] (0,49), kamernyj [chamber (adj.)] (0.48)
dolzhnostq [post]	post [post] (0.59), zvanie [rank] (0.49), naznachatq [to appoint] (0.46), otstavka [resignation] (0.46), naznachenie [appointment] (0.45), oklad [salary] (0.45), uvolqnjatq [to fire] (0.42)
dom [house]	kvartira [flat] (0.59), ulica [street] (0.54), zdanie [building] (0.54), dvor [garden] (0.53), gorod [town] (0.48), komnata [room] (0.45), domik [small house] (0.43), dacha [summer cottage] (0.39), semqja [family] (0.38), derevnja [village] (0.38)
domashnij [domestic]	domashnie [domestic] (0.55), koshka [cat] (0.32), zhivotnoe [animal] (0.31), kompqjuter [computer] (0.31), privychnyj [accustomed] (0.30), ujut [coziness] (0.29), kuxnja [kitchen] (0.29), xozjajstvo [facilities] (0.28), piwa [food] (0.27), kurica [hen] (0.27)
begatq [to run]	xoditq [to walk] (0.53), prygatq [to jump] 0.53, gonjatq [to race] (0.52), metatqsja [to race in panic] 0.50, nositqsja [to rush] (0.49), broditq [to wander] (4.5), gonjatqsja [to chase] (0.40), pobezhalyj [the color of a burnt steel] (0.40), polzatq [to creep] (0,40), chistitq [to clean] (0.40)
davno [long ago]	pora [period] (0.55), nedavno [recently] (0.46), davnym-davno [very long ago] (0.44), uzkij [narrow] (0,43), sej [sow! (imperative)] (0,42), kogda-to [once upon a time] (0.39), navsegda [forever] (0.38), nikogda [never] (0.38), rano [early] 0.35, teperq [now, nowadays] 0.34

done by Alexander Perekrestenko. Note that for Russian the transliteration from the Cyrillic to the Latin alphabet as already provided by the Russian Reference Corpus has been used.

7 Discussion

In Table 3 the performances of some other corpus-based systems that also had been evaluated on the TOEFL synonym test have been given. Although some of the other systems used much larger corpora (e.g. Terra and Clarke 2003, use a corpus of 53 billion words), the current knowledge-poor approach was able to outperform these results. This is an indication that the generalization effect claimed for SVD actually works in practice. This finding is also confirmed by our previous performance of only 69% achieved on the BNC without SVD (Rapp 2002).

Given the results of Table 3, for the task of computing semantically related words it also seems not essential to perform a syntactical analysis beforehand in order to determine specific dependency relations between words, as for example done by Pantel and Lin (2002). Although—depending on the corpus used—the results with our knowledge-poor approach can be somewhat noisy, it should be noted that—in case only paradigmatic relations are of interest—there is the possibility of filtering the output lists according to part of speech which should remove most of the noise.

But even if no filtering is performed, the results of our fully unsupervised approach which only relies on algebra largely agree with human intuition. Neglecting the preprocessing step of partial lemmatization, which essentially served the purpose of keeping our co-occurrence matrix small enough for SVD processing, no linguistic resources, neither a lexicon nor syntactic rules, are required. The algorithm considers any string of characters that is delimited by blanks or punctuation marks as a word, applies the SVD to an association matrix derived from the co-occurrences of the words in a corpus, and finally comes up with lists of similar words that highly agree with human intuitions.

8 Summary and prospects

We have presented a statistical method for the corpus-based automatic computation of related words which has been evaluated on the TOEFL synonym test. Its performance on this task favorably compares to other purely corpus-based approaches and suggests that sophisticated and language dependent syntactic processing is not essential.

The automatically generated sample thesauri of related words for English, French, German and Russian, each comprising in the order of 50,000 entries, are freely available from the author. Although, unlike other thesauri, at the current stage they do not distinguish different kinds of relationships between words, there is one advantage over manually created thesauri: Given a certain word, not only a few related words are listed. Instead, all words of a large vocabulary are ranked according to their similarity to the given word. Since, as indicated by the mostly correct rankings of the TOEFL alternatives, even at the higher ranks the distinctions obtained seem meaningful, this is an important feature that is indispensable for some kinds of machine processing, e.g. for word sense disambiguation and induction.

Future work that we envisage includes applying our method to corpora of other languages, adding multi-word units to the vocabulary, and to find solutions to the problem of word ambiguity that has not been dealt with here.

We also intend to increase the validity of comparisons between different algorithms by introducing a large WordNet-based gold standard which makes distinctions with regard to word frequency, part of speech, and ambiguity. To further investigate the virtues of dimensionality reduction, we plan to compare the behavior of SVD-based methods to conventional smoothing, thereby trying to find out whether the optimal number of dimensions simply corresponds to an optimal intensity of smoothing. Hereby an interesting question will be to find out whether this optimal intensity varies with the corpus frequency of the words under consideration.

Acknowledgements This research was supported by the German Research Society (DFG) and by a Marie Curie Intra-European Fellowship within the 6th European Community Framework Programme. The evaluation part of this research was conducted during my stay at the Centre for Language Technology of Macquarie University, Sydney. I would like to thank Robert Dale and his research group for the pleasant cooperation. Many thanks also go to Serge Sharoff and Alexander Perekrestenko for their support of this work.

References

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526.

Burnard, L., & Aston, G. (1998). *The BNC handbook: Exploring the British national corpus with Sara*. Edinburgh: Edinburgh University Press.

Denoyer, L., & Gallinari, P. (2006). The Wikipedia XML corpus. *ACM SIGIR Forum*, *40*(1), 64–69.

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

Fellbaum, C. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge: Bradford Books, MIT Press.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Dordrecht: Kluwer.

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Hirst, G., & St-Onge, D. (1998). Lexical chains as representation of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 305–332). Cambridge: MIT Press.

Jarmasz, M., & Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of the international conference on recent advances in natural language processing (RANLP-03)*, Borovets, Bulgaria, September (pp. 212–219).

Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the international conference on research in computational linguistics*, Taiwan.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.) (2007). *Handbook of Latent Semantic Analysis*. Hillsdale: Lawrence Erlbaum.

Leacock, C., & Chodorow, M. (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 265–283). Cambridge: MIT Press.

Lezius, W., Rapp, R., & Wettler, M. (1998). A freely available morphology system, part-of-speech tagger, and context-sensitive lemmatizer for German. In *Proceedings of COLING-ACL 1998*, Montreal (Vol. 2, pp. 743–748).

Lin, D. (1998a). Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 1998*, Montreal (Vol. 2, pp. 768–773).

Lin, D. (1998b). An information-theoretic definition of similarity. In *Proceedings of the 15th international conference on machine learning (ICML-98)*, Madison, WI (pp. 296–304).

Matveeva, I., Levow, G., Farahat, A., & Royer, C. (2005). Generalized latent semantic analysis for term representation. In *Proceedings of the international conference on recent advances in natural language processing (RANLP-05)*, Borovets, Bulgaria.

Pado, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Pantel, P., & Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM SIGKDD*, Edmonton (pp. 613–619).

Rapp, R. (2002). The computation of word associations: Comparing syntagmatic and paradigmatic approaches. In *Proceedings of 19th COLING*, Taipei, ROC (Vol. 2, pp. 821–827).

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In *Proceedings of the ninth machine translation summit*, New Orleans (pp. 315–322).

Rapp, R. (2004). A freely available automatically generated thesaurus of related words. In *Proceedings of the fourth international conference on language resources and evaluation (LREC)*, Lisbon (Vol. II, pp. 395–398).

Rapp, R. (2007). The computation of semantically related words: Thesaurus generation for English, German, and Russian. In B. Sharp & M. Zock (Eds.), *Natural language processing and cognitive science* (pp. 71–80). Setúba: INSTICC Press.

Resnik, P. (1995). Using information content to evaluate semantic similarity. In *Proceedings of the 14th international joint conference on artificial intelligence (IJCAI-95)*, Montreal (pp. 448–453).

Ruge, G. (1992). Experiments on linguistically based term associations. *Information Processing and Management*, *28*(3), 317–332.

Ruiz-Casado, M., Alfonseca, E., & Castells, P. (2005) Using context-window overlapping in Synonym Discovery and Ontology Extension. In *Proceedings of the international conference recent advances in natural language processing (RANLP-2005)*, Borovets, Bulgaria.

Sahlgren, M. (2001). Vector-based semantic analysis: representing word meanings based on random labels. In A. Lenci, S. Montemagni, & V. Pirrelli (Eds.), *Proceedings of the ESSLLI workshop on the acquisition and representation of word meaning*, Helsinki.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to German. In *Proceedings of the EACL SIGDAT workshop*, Dublin (pp. 47–50).

Schütze, H. (1997). *Ambiguity resolution in language learning: computational and cognitive models*. Stanford: CSLI Publications.

Terra, E., & Clarke, C. L. A. (2003). Frequency estimates for statistical word similarity measures. In *Proceedings of HLT/NAACL*, Edmonton, Alberta (pp. 244–251).

Turney, P. D. (2001). Mining the Web for synonyms. PMI-IR versus LSA on TOEFL. In *Proc. of the twelfth European conference on machine learning*, Freiburg, Germany (pp. 491–502).

Turney, P. D. (2006). Similarity of semantic relations. *Computational Linguistics*, *32*(3), 379–416.

Turney, P. D. (2008). A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)*, Manchester, UK (pp. 905–912).

Turney, P. D., Littman, M. L., Bigham, J., & Shnayder, V. (2003). Combining independent modules to solve multiple-choice synonym and analogy problems. In *Proceedings of the international conference on recent advances in natural language processing (RANLP-03)*, Borovets, Bulgaria (pp. 482–489).