Translating Highly Untranslatable Words: MUSE at the Edge of Translatability

Martin Pömsl

August 18, 2019

Contents

1 Introduction

Bilingual dictionaries are the basis for most translation methods of today both by machines and by human experts. That is why building and maintaining reliable bilingual dictionaries is such an important task. Although it may seem as if today's dictionaries are already very comprehensive, an automatic procedure to assemble bilingual dictionaries would still be useful for low-resource language pairs as well as to keep existing dictionaries up-to-date in the fast-moving world of today, where new words are coined every day.

For a long time, automatic methods for bilingual dictionary generation focused on statistically extracting translation equivalents from aligned parallel corpora. While very effective, this method requires a lot of human expert work to create aligned parallel corpora in the first place.

With the arrival of computational procedures to efficiently estimate word vectors (so-called word embeddings), another method emerged: Due to the inherent similarities in word embeddings across languages, it was suddenly possible not to align raw texts, but instead to align monolingual word embeddings by learning a linear mapping from source space to target space. This mapping was learned in a supervised way by comparing the vector representations of a few known translation equivalents.[1]

Recently, another improvement has been made to this method: The MUSE[2] library provides a fully unsupervised method to learn a mapping between two monolingual vector spaces. It is the first method to determine translation equivalents that does not use any existing human translation knowledge, but only relies on the structural similarities it finds in the vector spaces.[3]

Since this unsupervised method is never exposed to any kind of human-translated content, but only to monolingual word embeddings, it is completely ignorant of human approaches to translation. Therefore it remains unaffected by the biases and personal connotations a human translator might have. In that sense, it can be considered the first fully "neutral" method for word translation.

Another interesting property of the MUSE method is that it cannot pick and choose which words to translate based on how good the available translation equivalents are, as a human translator might do. Since it learns a single mapping that is applied to all words in the source space, it automatically generates translation equivalents for all words in the source space. The method can thus be

[1] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781 (2013).
[2] https://github.com/facebookresearch/MUSE
[3] Alexis Conneau and Guillaume Lample. *Word Translation Without Parallel Data*. 2017. arXiv: 1710.04087.

considered "complete", since it provides a translation equivalent for all words.

The property of completeness makes it possible to apply the unsupervised MUSE method to words which are usually considered "untranslatable", because no stable translation equivalent with the same meaning exists in the target language. Human translators and other automatic methods would usually either paraphrase these words or skip them altogether.

The property of neutrality makes it semantically interesting to apply the method to untranslatable words, since the method is not influenced by how human translators usually paraphrase these words. It is therefore expected that the results are approximations of the true meaning in which the words are most frequently used, as the method only relies on the distributional statistics encoded in the word embeddings.

This paper will produce a set of such translation attempts by using the MUSE method to translate German noun-noun compounds that are considered highly untranslatable into English. It will then provide both a qualitative and quantitative evaluation of the results. The code used to produce the translation attempts and to automatize the evaluation procedure will be released on GitHub.[4]

It is hoped that these experiments can shed some light on the usage and meaning of some words which have so far proven untranslatable. However, the main purpose of this work is the evaluation of the unsupervised MUSE method in extreme cases to test its usability in large-scale applications, where highly untranslatable words are bound to occur frequently.

2 Related Work

2.1 Word Embeddings

In 2013, Mikolov et al. proposed techniques to learn numerical word representations from large quantities of raw text which are known as word2vec.[5]. The resulting word embeddings assign each word in the source corpus a vector of fixed dimensionality. The final word embeddings correspond to the weights of a neural network that was trained on the task of predicting surrounding words from a given word (skip-gram) or predicting a missing word from its surrounding words (continuous bag-of-words). Since both methods aim to predict co-occurrences, they are effectively applications of the distributional hypothesis, which states that words occurring in similar contexts tend to have similar meanings.[6] Indeed the resulting word embeddings encode some semantic relationships very

[4]https://github.com/mpoemsl/THUW

[5]Mikolov et al., “Efficient Estimation of Word Representations in Vector Space”.

[6]Zellig S Harris. “Distributional structure”. In: *Word* 10.2-3 (1954), pp. 146–162.

efficiently, so that they are often used as the basis for downstream tasks.

2.2 Aligning Word Embeddings

It was soon noted that some regularities are constant in word embeddings across languages, presumably since there are some semantics common to all languages which will always have similar co-occurence statistics, e.g. the concepts "mother" and "son". Mikolov et al. proposed a method to align the word embeddings of two languages by computing a linear mapping based on these anchor points.[7] From these aligned word embeddings a translation dictionary can be created by simply applying the mapping to all vectors in the source space and taking the nearest neighbors of the resulting points the target space as most likely translations.

MUSE takes this concept one step further by aligning to vector spaces without any anchor points in a fully unsupervised way.[8] This is mainly achieved through the implementation of a Generative Adversarial Network[9], where the generator is the proposed mapping and the discriminator tries to distinguish whether given points come from the source or target space. On top of this, MUSE also employs the closed-form solution to the orthogonal Procrustes problem by Schönemann[10] to further refine the solution iteratively. The Procrustes problem is the problem of finding a linear mapping that maps a vector space to another with as little error as possible.

MUSE operates on fastText word embeddings, which are in principle generated in a similar way to word2vec skip-gram embeddings, but also include subword information. That means that the relation of words on a character-level is also taken into account by the word embeddings.[11]

2.3 Measures of Untranslatability

In the task of word translation, there are frequently some words in the source language for which there is no established translation equivalent in the target language. It is of interest for document translation to identify such words beforehand and single them out for special treatment, e.g. paraphrasing. In order to do this, an automatic measure of translatability is required. A theoretical

[7] Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. *Exploiting Similarities among Languages for Machine Translation*. 2013. arXiv: `1309.4168 [cs.CL]`.

[8] Conneau and Lample, *Word Translation Without Parallel Data*.

[9] Ian Goodfellow et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680. URL: `http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf`.

[10] Peter H. Schönemann. “A generalized solution of the orthogonal procrustes problem”. In: *Psychometrika* 31.1 (Mar. 1966), pp. 1–10. ISSN: 1860-0980. DOI: `10.1007/BF02289451`. URL: `https://doi.org/10.1007/BF02289451`.

[11] Piotr Bojanowski et al. “Enriching word vectors with subword information”. In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

framework for the quantification of translatability has been established by Zakaria, however without providing an implementation.[12]

Johnson also proposes a method to quantify the translatability of noun-noun compounds as the mean of two separate scoring methods[13]: One determines the stability of common translation attempts by measuring their frequency similarily to Tufis et al.[14] The other considers how easily the meaning of the whole compound is recoverable by translating its components, following Weller and Heid.[15]

3 Methodology

3.1 Data Selection

The translation of noun-noun compounds is a major issue in translation, as they are often poorly documented and the meaning of the compound tends to transcend the meaning of its components in many languages.[16]

The German language is notoriously noun-noun compound productive, which makes it an ideal source language for the purposes of this study.[17] As a target language, English was chosen due to the availability of reliable bilingual resources for this language pair.

Consequently, a list of German noun-noun compounds that are usually considered "untranslatable" was compiled. Candidates for this list were gathered from various blogs on the internet, which is rich in articles about "untranslatable words".

The untranslatability of the words was validated by only accepting those words with a translatability score of less than 0.6 by Johnson's scoring system, which can be determined using the code in the GitHub Repository Assessing Translatability.[18] As a preprocessing step, the words were split into their noun components manually to avoid introducing noise at this stage. The threshold of 0.6 was

[12] Ingie Zakaria. "Quantifying a successful translation: A cognitive frame analysis of (un)translatability". In: (Jan. 2017).

[13] Christian Johnson. "Assessing Translatability: An Algorithmic Procedure for Identifying and Scoring Difficult-to-Translate German to English Multi-Noun Compounds by Leveraging Bilingual Corpus Data". In: (Aug. 2019).

[14] Dan Tufis, Ana-Maria Barbu, and Radu Ion. "Extracting Multilingual Lexicons from Parallel Corpora". In: *Computers and the Humanities* 38 (May 2004), pp. 163–189.

[15] Marion Weller and Ulrich Heid. "Analyzing and Aligning German Compound Nouns". In: *in Proceedings of LREC*. 2012.

[16] Takaaki Tanaka and Timothy Baldwin. "Noun-noun compound machine translation: A feasibility study on shallow processing". In: (Jan. 2003).

[17] Hinde Metsenaere, Sonia Vandepitte, and Marc Van de Velde. "Dutch and German noun-noun compounds in translation". In: 61 (Apr. 2016), pp. 175–205. DOI: 10.1515/les-2016-0006.

[18] https://github.com/christianj6/Assessing-Translatability

determined by computing tje translatability scores for a test group of ten randomly chosen German noun-noun compounds for which suitable translation equivalents exist. 0.6 was found to be a suitable lower bound for the translatability scores of translatable words. The translatability scores of both the untranslatable words and the test group can be found in the appendix.

For ten of the previously assembled candidates, the translatability score was less than 0.6. As a sanity check, the dictionary compiled by the authors of MUSE for evaluation purposes was searched for translation equivalents for the accepted candidates, but none were found. Consequently, the ten candidate words were accepted and will from now on be references by the moniker "untranslatable words".

Word	**Translatability Score**	**In Dictionary**
Kopfkino	0.5348	False
Morgenmuffel	0.5143	False
Schnapsidee	0.4849	False
Ohrwurm	0.3590	False
Weltschmerz	0.2369	False
Sitzfleisch	0.1988	False
Treppenwitz	0.1938	False
Lebensmüdigkeit	0.1885	False
Geisterfahrer	0.1533	False
Torschlusspanik	0.0540	False

Since the untranslatable words are by their very nature infrequent, some of them were not contained in the word embeddings used by the authors of the original papers, which used two hundred thousand word vectors pre-trained on the Wikipedia corpus with fastText. To alleviate this, a large set of fastText word embeddings with two million words pre-trained on Wikipedia and Common Crawl was used.[19] These word embeddings contained all untranslatable words. The Wikipedia corpus is the raw text of all articles on the collaborative multi-lingual online encyclopedia of the same name. The Common Crawl corpus consists of over 3 billion internet pages crawled by the Common Crawl bot.

3.2 Experimental Setup

The German and English word embeddings were aligned by running the unsupervised MUSE method on an Amazon Web Services EC2 instance. The code was directly cloned from the MUSE GitHub repository[20] and the unsupervised alignment was performed using both adversarial mapping gener-

[19] Edouard Grave et al. "Learning Word Vectors for 157 Languages". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

[20] https://github.com/facebookresearch/MUSE

ation and Procrustes refinement with the default parameters.

The computation took 2.3 hours on a p2.xlarge EC2 instance running the Deep Learning AMI on Linux. This corresponds to using a NVIDIA Tesla K80 GPU with CUDA. Including file storage expenses, the total costs of this experiment amounted to roughly 3 USD.

For each of the previously established untranslatable words, the 10 nearest neighbors were computed both in the source and in the target space. This corresponds to a setup of k=10 in terms of the original paper. On the evaluation dictionaries provided by the authors, the aligned word embeddings score a precision of 81.6 in the k=10 setup and of 63.3 in the k=1 setup. This is roughly consistent with the 69.6 precision in the k=1 setup reported by the authors of the original paper for the de-en language pair. The discrepancy might be explained by the fact that the data set in the original paper used different and much smaller word embeddings.

When compiling the set of nearest neighbors, the constraint of semantic uniqueness is imposed by only accepting words with a Levenshtein distance of less than 3 to each other. The purpose of this post-processing step is to filter out inflected forms and spelling errors.

3.3 Evaluation

The question of how to evaluate the success of translating an untranslatable word is of some difficulty. This is due to the fact that the main property of an untranslatable word is that there does not exist a commonly used translation equivalent to which the translation attempt could be compared for evaluation purposes.

Since there is no corresponding word in the target language, it is impossible to judge the position in the target vector space against the position of the true translation equivalent, as it would normally be done. However, since the nearest neighbors of the source word usually are quite translatable, it is possible to compare the nearest neighbors of the source word with the nearest neighbors in the target space. According to the distributional hypothesis, the neighbors should have similar semantics as well if translation has succeeded.

For these purposes, for each untranslatable word a co-occurrence matrix of nearest neighbors in source space and target space are compiled. An algorithmic approach to compiling this co-occurrence matrix has been implemented by evaluating neighbors in a pairwise manner. If the translation of the neighbor in the target space has a Levenshtein distance of less than three to the neighbor in the source space, they are judged to have the same meaning. The translations are procured by means of an established procedure (Google Translate API).

However, this automatic method fails to identify some semantic equivalents, possibly due to the translations failing to capture the polysemy of some words. Because of this, the co-occurence matrices are complemented manually by a human evaluator as a post-processing step. The quality of the translation is then quantified by the number of semantic co-occurrences between the nearest neighbors in the source and target space.

	durchhalte-vermögen	steh-vermögen	ausdauer	nerven-kostüm	stand-vermögen	geduld
brain-power	0	0	0	0	0	0
stamina	1	1	1	0	0	0
lie-down	0	0	0	0	0	0
intrepidness	0	0	0	0	0	0
muscle-power	0	0	0	0	0	0
brain-work	0	0	0	0	0	0
bone-weary	0	0	0	0	0	0
slackens	0	0	0	0	0	0
brawniness	0	0	0	0	0	0
grizzle	0	0	0	0	0	0

Table 1: Partial co-occurrence matrix of the untranslatable word "Sitzfleisch"

4 Discussion

The computed nearest neighbors for all untranslatable words in source and target space as well as co-occurrence scores can be found in the appendix.

4.1 Quantitative Analysis

When looking at the co-occurrence scores of the neighbors, it becomes apparent that there is a high variance in the scores: Values range from 1 to 9, the mean is 3.9.

The theoretical upper bound for the co-occurrence score is the maximum possible number of co-occurrences, which is 100 when considering the ten nearest neighbors. However, since this would require all of the neighbors to be synonyms, a more realistic upper bound is 20.

The words can intuitively be clustered into three groups based on the co-occurrence count: Those with few occurrences (1-3), those with medium occurrences (4-6), and those with many occurrences (7-9).

4.2 Qualitative Analysis

The quantitative groups also seem to exhibit similar qualitative properties. In the following, the neighbors of one representative word from each group will be semantically analyzed.

Few Co-occurrences: Torschlusspanik

Nearest Neighbors in Source Space:

torschluss; untergangsstimmung; torscharte; unentschiedenheit; liebesbeteuerungen; stimmungss-chwankung; zukunftsangst; orgasmusstörungen; steuererhöhungspläne; stimmungsumschwünge

Nearest Neighbors in Target Space:

desperation; unhappiness; infatuation; post-honeymoon; wedding-planning; bridezillas; mangage-ment; marriage; nuptials; newlywed

Torschlusspanik is the fear of running out of time in achieving ones life goals and is often associated with marital problems. While this is in principle reflected semantically in the neighbors both in source and in target space, there is a crucial difference which causes the co-occurrence count to be so low: The German nearest neighbors are all terms indirectly associated with marriage problems, however the German word for marriage is never explicitly mentioned. In the English nearest neighbors, however, marriage and closely related terms like "nuptials" and "wedding" are mentioned all the time. One interpretation of this could be a cultural difference: It is possible that German users write about marriage problems more indirectly than English users. In general it can be concluded that few co-occurences indicate semantic discrepancies in the neighbors of the source and the target language.

Medium Co-occurrences: Geisterfahrer

Nearest Neighbors in Source Space:

falschfahrer; geisterfahrerin; autofahrer; geisterfahrt; geisterradler; pkw-fahrer; verkehrsteilnehmer; motorradfahrer; lasterfahrer; raser

Nearest Neighbors in Target Space:

motorcyclist; motorbiker; motorist; hit-run; t-boning; bike-riders; drink-driver; bicyclist; road-sign; speeding

Geisterfahrer is the German term for a driver who goes in the wrong direction and thus endangers fellow road users. Both the German and the English neighbors roughly follow the same theme: Different kinds of road users. However, the English neighbors deviate from this theme in that they not only refer to the road users, but also to accidents ("hit-run", "t-boning") and their possible causes ("speeding", "drink-driver"). This can be interpreted as being an indication that driving in the wrong direction is more closely associated with accidents for English users than for German users. In general it seems to be the case that a medium number of co-occurrences indicates that the semantics are covered quite well by the translation, but they are still differing in some aspects.

Many Co-occurrences: Morgenmuffel

Nearest Neighbors in Source Space:

frühaufsteher; morgenmensch; langschläfer; frühstücksmuffel; wachmacher; spätaufsteher; muntermacher; nachtmensch; weckerklingeln; bewegungsmuffel

Nearest Neighbors in Target Space:

non-morning; early-riser; sleepyhead; night-owl; oversleep; mornings; grumpy; pick-me-ups; grouchy; groggy

Morgenmuffel is the German term for a person that does not like waking up in the morning. Both English and German neighbors are filled with the themes of waking up and sleep habits, including a surprisingly high number of semantic overlaps. For example both languages cover terms for people that like to sleep late, rise early or stay up late. Substances that help waking up ("pick-me-ups") are also frequent. This indicates a nearly complete semantic overlap, which could be caused by the fact that sleep is something common to users of all languages and does not vary so much across cultures. It can be concluded that a high number of co-occurrences indicates a large semantic overlap in neighbors and thus a high quality of translation.

It seems that the unsupervised MUSE method yields good results for untranslatable words denoting concepts that are common to users of both source and target language. However, if the concept behind the word is subject to ambiguity due to cultural differences, the translations yielded by the unsupervised MUSE method are colored by these cultural differences. This is an indicator that although the unsupervised MUSE method is never exposed to traditional translations, it is still subject to biases due to the co-occurrence statistics encoded in the word embeddings.

5 Conclusion

In this paper, the MUSE library was used to find approximate translations of words that are usually considered "untranslatable" by traditional means. For this purpose the nearest neighbors of German noun-noun compounds with a low translatability score were determined after aligning large German and English word embeddings with the unsupervised MUSE method. The quality of the translations was evaluated by counting the number of semantic co-occurrences of the nearest neighbors in the source space and target space. It was found that co-occurrence scores are a good indicator of the quality of translations and that differences in the semantics of source nearest neighbors and target nearest neighbors are most likely due to cultural differences between the users of the languages. From this it can be concluded that the unsupervised MUSE method yields good translation approximations even for highly untranslatable words only under the condition that the concepts behind the words are not subject to cultural differences between the users of source and target languages.

6 References

Bojanowski, Piotr et al. "Enriching word vectors with subword information". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 135–146.

Conneau, Alexis and Guillaume Lample. *Word Translation Without Parallel Data*. 2017. arXiv: 1710.04087.

Goodfellow, Ian et al. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani et al. Curran Associates, Inc., 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Grave, Edouard et al. "Learning Word Vectors for 157 Languages". In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*. 2018.

Harris, Zellig S. "Distributional structure". In: *Word* 10.2-3 (1954), pp. 146–162.

Johnson, Christian. "Assessing Translatability: An Algorithmic Procedure for Identifying and Scoring Difficult-to-Translate German to English Multi-Noun Compounds by Leveraging Bilingual Corpus Data". In: (Aug. 2019).

Metsenaere, Hinde, Sonia Vandepitte, and Marc Van de Velde. "Dutch and German noun-noun compounds in translation". In: 61 (Apr. 2016), pp. 175–205. DOI: 10.1515/les-2016-0006.

Mikolov, Tomas, Quoc V. Le, and Ilya Sutskever. *Exploiting Similarities among Languages for Machine Translation*. 2013. arXiv: 1309.4168 [cs.CL].

Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space". In: *CoRR* abs/1301.3781 (2013).

Schönemann, Peter H. "A generalized solution of the orthogonal procrustes problem". In: *Psychometrika* 31.1 (Mar. 1966), pp. 1–10. ISSN: 1860-0980. DOI: 10.1007/BF02289451. URL: https://doi.org/10.1007/BF02289451.

Tanaka, Takaaki and Timothy Baldwin. "Noun-noun compound machine translation: A feasibility study on shallow processing". In: (Jan. 2003).

Tufis, Dan, Ana-Maria Barbu, and Radu Ion. "Extracting Multilingual Lexicons from Parallel Corpora". In: *Computers and the Humanities* 38 (May 2004), pp. 163–189.

Weller, Marion and Ulrich Heid. "Analyzing and Aligning German Compound Nouns". In: *in Proceedings of LREC*. 2012.

Zakaria, Ingie. "Quantifying a successful translation: A cognitive frame analysis of (un)translatability". In: (Jan. 2017).

7 Appendix

7.1 Translatability Scores

Untranslatable Words:

Word	Translatability Score	In Dictionary
Kopfkino	0.5348	False
Morgenmuffel	0.5143	False
Schnapsidee	0.4849	False
Ohrwurm	0.3590	False
Weltschmerz	0.2369	False
Sitzfleisch	0.1988	False
Treppenwitz	0.1938	False
Lebensmüdigkeit	0.1885	False
Geisterfahrer	0.1533	False
Torschlusspanik	0.0540	False

Test Group:

Word	Translatability Score	In Dictionary
Zimmerdecke	0.7716	False
Radioantenne	0.7688	False
Schulbeginn	0.7432	False
Bananenbrot	0.6675	False
Kochtopf	0.6629	False
Briefkasten	0.6560	True
Motorrad	0.6534	True
Kinderzimmer	0.6347	True
Kreditkarte	0.6272	True
Hundehütte	0.6180	False

7.2 MUSE Parameters

- adversarial: True
- batch_size: 32
- dico_build: S2T
- dico_eval: default
- dico_max_rank: 15000
- dico_max_size: 0
- dico_method: csls_knn_10
- dico_min_size: 0
- dico_threshold: 0
- dis_clip_weights: 0
- dis_dropout: 0.0
- dis_hid_dim: 2048
- dis_input_dropout: 0.1
- dis_lambda: 1

- dis_layers: 2
- dis_most_frequent: 50000
- dis_optimizer: sgd,lr=0.1
- dis_smooth: 0.1
- dis_steps: 5
- emb_dim: 300
- epoch_size: 1000000
- lr_decay: 0.95
- lr_shrink: 0.5
- map_beta: 0.001
- map_id_init: True
- map_optimizer: sgd, lr=0.1
- max_vocab: 200000
- min_lr: 1e-06
- n_epochs: 5
- n_refinement: 5
- normalize_embeddings: False
- seed: -1
- src_lang: de
- tgt_lang: en

7.3 Nearest Neighbor Co-occurences

Source Word	**Nearest Neighbors in Source Space**	**Co-Occs**
geisterfahrer	falschfahrer;geisterfahrerin;autofahrer;geisterfahrt;geisterradler; pkw-fahrer;verkehrsteilnehmer;motorradfahrer;lasterfahrer;raser	4
ohrwurm	ohrwürmer;ohrwurmqualität;song;ohrwurmgarantie;ohrwurmmelodien; ohrwurmcharakter;gute-laune-song;ohrwurmpotential;pop-song;hammersong	5
sitzfleisch	durchhaltevermögen;stehvermögen;ausdauer;nervenkostüm; standvermögen;geduld;köpfchen;fingerspitzengefühl;wadeln;hirnschmalz	5
lebensmüdigkeit	lebensüberdruss;antriebslosigkeit;tagesmüdigkeit;hoffnungslosigkeit; dauermüdigkeit;depressionen;motivationslosigkeit;schwermut; depressivität;ohnmachtsgefühle	5
morgenmuffel	frühaufsteher;morgenmensch;langschläfer;frühstücksmuffel;wachmacher; spätaufsteher;muntermacher;nachtmensch;weckerklingeln;bewegungsmuffel	9
schnapsidee	idee;bierlaune;super-idee;marketing-idee;geschäftsidee; luftnummer;spitzenidee;unsinn;quatsch;saublöde	6
kopfkino	kino;hirn;sinnesrausch;gänsehautfeeling;gänsehaut;gefühlszentrum; männerfantasie;gänsehautgefühl;gedankenkarussell;leseerlebnis	2
treppenwitz	witz;ostfriesenwitz;blondinenwitz;abstrus;schildbürgerstreich;lachnummer; anachronismus;aprilscherz;riesenblamage;häschenwitz	3
torschlusspanik	torschluss;untergangsstimmung;torscharte;unentschiedenheit; liebesbeteuerungen;stimmungsschwankung;zukunftsangst; orgasmusstörungen;steuererhöhungspläne;stimmungsumschwünge	1
weltschmerz	melancholie;seelenschmerz;weltschmerzes;liebesschmerz; schmerz;herzschmerz;traurigkeit;menschenhass; selbstmitleid;lebensüberdruss	4

Source Word	Nearest Neighbors in Target Space	Co-Occs
geisterfahrer	motorcyclist;motorbiker;motorist;hit-run;t-boning; bike-riders;drink-driver;bicyclist;road-sign;speeding	4
ohrwurm	ear-worm;song;head-bopping;toe-tapper;ditty; clap-along;funkalicious;melody;back-beat;bass-line	5
sitzfleisch	brain-power;stamina;lie-down;intrepidness;muscle-power; brain-work;bone-weary;slackens;brawniness;grizzle	5
lebensmüdigkeit	deliriums;loneliness;depressiveness;hopelessness; melancholy;love-sickness;desperation;despondence; boredom;morbidness	5
morgenmuffel	non-morning;early-riser;sleepyheads;night-owl; oversleep;mornings;grumpy;pick-me-ups;grouchy;groggy	9
schnapsidee	idea;joke;pipedream;knee-slapper;pie-in-the-sky; oxymoron;crack-pot;guffaw;balderdash;cockamamie	6
kopfkino	inner-monologue;erotic;mind-fucking;pageturner; suspense;sex-scene;sexytimes;mid-book;steaminess;eroticism	2
treppenwitz	joke;factoid;irony;travesty;farce;truthiness; laughable;canard;oddity;myth	3
torschlusspanik	desperation;unhappiness;infatuation;post-honeymoon; wedding-planning;bridezillas;mangagement;marriage; nuptials;newlywed	1
weltschmerz	lonesomeness;melancholy;angst;misanthropy;neurosis; self-pity;ennui;moroseness;self-loathing;depressiveness	4