Distributional Semantics and Compositionality 2011: Shared Task Description and Results

Chris Biemann
UKP lab, Technical University of Darmstadt
Hochschulstr. 10
64289 Darmstadt, Germany
biemann@tk.informatik.tu-darmstadt.de

Eugenie Giesbrecht
FZI Forschungszentrum Informatik
Haid-und-Neu-Str. 10-14
76131 Karlsruhe, Germany
giesbrecht@fzi.de

Abstract

This paper gives an overview of the shared task at the ACL-HLT 2011 DiSCo (Distributional Semantics and Compositionality) workshop. We describe in detail the motivation for the shared task, the acquisition of datasets, the evaluation methodology and the results of participating systems. The task of assigning a numerical score for a phrase according to its compositionality showed to be hard. Many groups reported features that intuitively should work, yet showed no correlation with the training data. The evaluation reveals that most systems outperform simple baselines, yet have difficulties in reliably assigning a compositionality score that closely matches the gold standard. Overall, approaches based on word space models performed slightly better than methods relying solely on statistical association measures.

1 Introduction

Any NLP system that does semantic processing relies on the assumption of semantic compositionality: the meaning of a phrase is determined by the meanings of its parts and their combination. However, this assumption does not hold for lexicalized phrases such as idiomatic expressions, which causes troubles not only for semantic, but also for syntactic processing (Sag et al., 2002). In particular, while distributional methods in semantics have proved to be very efficient in tackling a wide range of tasks in natural language processing, e.g., document retrieval, clustering and classification, question answering, query expansion, word similarity, synonym extraction, relation extraction, textual advertisement matching in search engines, etc. (see Turney and Pantel (2010) for a detailed overview), they are still strongly limited by being inherently word-based. While dictionaries and other lexical resources contain multiword entries, these are expensive to obtain and not available for all languages to a sufficient extent. Furthermore, the definition of a multiword varies across resources, and non-compositional phrases are often merely a subclass of multiword units.

This shared task addressed researchers that are interested in extracting non-compositional phrases from large corpora by applying distributional models that assign a graded compositionality score to a phrase, as well as researchers interested in expressing compositional meaning with such models. The score denotes the extent to which the compositionality assumption holds for a given expression. The latter can be used, for example, to decide whether the phrase should be treated as a single unit in applications. We emphasized that the focus is on automatically acquiring semantic compositionality and explicitly did not invite approaches that employ prefabricated lists of non-compositional phrases.

It is often the case that compositionality of a phrase depends on the context. Though we have used a sentence context in the process of constructing the gold standard, we have decided not to provide it with the dataset: we have asked for a single compositionality score per phrase. In an application, this could play the role of a compositionality prior that could, e.g., be stored in a dictionary. There is a long-living tradition within the research

community working on multiword units (MWUs) to automatically classify MWUs into either compositional or non-compositional ones. However, it has been often noted that compositionality comes in degrees, and a binary classification is not valid enough in many cases (Bannard et al., 2003; Katz and Giesbrecht, 2006). To the best of our knowledge, this has been the first attempt to offer a dataset and a shared task that allows to explicitly evaluate the models of graded compositionality.

2 Shared Task Description

For the shared task, we aimed to get compositionality scores for phrases frequently occurring in corpora. Since distributional models need large corpora to perform reliable statistics, and these statistics are more reliable for frequent items, we chose to restrict the candidate set to the most frequent phrases from the freely available WaCky[1] web corpora (Baroni et al., 2009). Those are currently downloadable for English, French, German and Italian. They have already been automatically sentence-split, tokenized, part-of-speech (POS) tagged and lemmatized, which reduces the load on both organizers and participants that decide to make use of these corpora. Further, WaCky corpora provide a good starting point for experimenting with distributional models due to their size, ranging between 1-2 billion tokens, and extensive efforts to make these corpora as clean as possible.

2.1 Candidate Selection

There is a wide range of subsentential units that can function as a non-compositional construction. These units do not have to be realized continuously in the surface realization and can consist of an arbitrary number of lexical items. While it would be interesting to examine unrestricted forms of multiwords and compositional phrases, we decided to restrict candidate selection to certain grammatical constructions to make the task more tangible. Specifically, we use word pairs in the following relations:

- ADJ_NN: Adjective modifying a noun, e.g. "red herring" or "blue skies"
- V_SUBJ: Noun in subject position and verb, e..g. "flies fly" or "people transfer (sth.)"
- V_OBJ: Noun in object position and verb, e.g. "lose keys", "play song"

While it is possible to extract the relations fairly accurately from parsed English text, there is – to our knowledge – no reliable, freely available method that can tell verb-subjects from verb-objects for German. Thus, we employed a three-step selection procedure for producing a set of candidate phrases per grammatical relation and language that involved heavy manual intervention.

1. Extract candidates using (possibly overgenerating) patterns over part-of-speech sequences and sort by frequency
2. Manually select plausible candidates for the target grammatical relation in order of decreasing frequency
3. Balance the candidate set to select enough non-compositional phrases

For English, we used the following POS patterns: ADJ_NN: "JJ* NN*"; V_SUBJ: "NN* VV*"; V_OBJ: "VV* DT|CD NN*" and "VV* NN*". The star * denotes continuation of tag labels: e.g. VV* matches all tags starting with "VV", such as VV, VVD, VVG, VVN, VVP and VVZ.

For German, we used "ADJ* NN*" for ADJ_NN. For relations involving nouns and verbs, we extracted all noun-verb pairs in a window of 4 tokens and manually filtered by relation on the aggregated frequency list. Frequencies were computed on the lemma forms.

This introduces a bias on the possible constructions that realize the target relations, especially for the verb-noun pairs. Further, the selection procedure is biased by the intuition of the person that performs the selection. We only admitted what we thought were clear-cut cases (only nouns that are typically found in subject respectively object position) to the candidate set at this stage.

Since non-compositional phrases are much less in numbers than compositional phrases, we tried to somewhat balance this in the third step in the selection. If the candidates would have been randomly

[1] http://wacky.sslmit.unibo.it

selected, an overwhelming number of compositional phrases would have rendered the task very hard to evaluate, since a baseline system predicting high compositionality in all cases would have achieved a very high score. We argue that since we are especially interested in non-compositional phrases in this competition, it is valid to bias the dataset in this way.

After we collected a candidate list, we randomly selected seven sentences per candidate from the corpus. Through manual filtering, we checked whether the target word pair was in fact found in the target relation in these sentences. Further we removed incomplete and too long sentences, so that we ended up with five sentences per target phrase. Some candidate phrases that only occurred in very fixed contexts (e.g. disclaimers) or did not have enough well-formed sentences were removed in this step.

Figure 1 shows the sentences for "V_OBJ: buck trend" as an example output of this procedure.

2.2 Annotation

The sample usages of target phrases now had to be annotated for compositionality. We employed the crowdsourcing service Amazon Turk[2] for realizing these annotations. The advantage of crowdsourcing is its scalability through the large numbers of workers that are ready to perform small tasks for pay. The disadvantage is that tasks usually cannot be very complex, since quality issues (scammers) have to be addressed either with test items or redundancy or both – mechanisms that only work for types of tasks where there is clearly a correct answer.

Previous experiences in constructing linguistic annotations with Amazon Turk (Biemann and Nygaard, 2010) made us stick to the following two-step procedure that more or less ensured the quality of annotation by hand-picking workers:

1. *Gather high quality workers*: In an open task for a small data sample with unquestionable decisions, we collected annotations from a large number of workers. Workers were asked to provide reasons for their decisions. Workers that performed well, gave reasons that demonstrated their understanding of the task and completed a significant amount of the examples were invited for a closed task. Net pay was 2 US cents for completing a HIT.

2. *Get annotations for the real task*: In the closed task, only invited workers were admitted and redundancy was reduced to four workers per HIT. Net pay was 3 US cents for completing a HIT.

Figure 2 shows a sample HIT (human intelligence task) for English on Amazon turk, including instructions. Workers were asked to enter a judgment from 0-10 about the literacy of the highlighted target phrase in the respective context. For the German data, we used an equivalent task definition in German.

All five contexts per target phrase were scored by four workers each. A few items were identified as problematic by the workers (e.g. missing highlighting, too little context), and one worker was excluded during the English experiment for starting to deliberately scam. For this worker, all judgments were removed and not repeated. Thus, the standard number of judgments per target phrase was 20, with some targets receiving less judgments because of these problems. The minimum number of judgments per target phrase was 12: four HITs with three judgments each.

From this, we computed a score by averaging over all judgments per phrase and multiplying the overall score by 10 to get scores in the range of 0-100. This score cannot help in discriminating moderately compositional phrases like "V_OBJ: make decision" from phrases that are dependent on the context like "V_OBJ: wait minute" which had two HITs for the idiomatic use of "wait a minute!" and three HITs with literally minutes to spend idling.

As each HIT was annotated by a possibly different set of workers, it is not possible to compute inter-annotator agreement. Eyeballing the scores revealed that some workers generally tend to give higher respectively lower scores than others. Overall, workers agreed more for clearly compositional or clearly non-compositional HITs. We believe that using this comparatively high number of judgments per target, averaged over several contexts, should give us fairly reliable judgments, as worker biases should cancel out each other.

[2] http://www.mturk.com

- I would like to **buck** the **trend** of complaint !
- One company that is **bucking** the **trend** is Flowcrete Group plc located in Sandbach , Cheshire .
- ” We are now moving into a new phase where we are hoping to **buck** the **trend** .
- With a claimed 11,000 customers and what look like aggressive growth plans , including recent acquisitions of Infinium Software , Interbiz and earlier also Max international , the firm does seem to be **bucking** the **trend** of difficult times .
- Every time we get a new PocketPC in to Pocket-Lint tower , it seems to offer more features for less money and the HP iPaq 4150 is n’t about to **buck** the **trend** .

Figure 1: sentences for V_OBJ: buck trend after manual filtering and selection. The target is **highlighted**.

How literal is this phrase?

Can you infer the meaning of a given phrase by only considering their parts literally, or does the phrase carry a ’special’ meaning?
In the context below, how literal is the meaning of the phrase in bold?
Enter a number between 0 and 10.

- 0 means: this phrase is not to be understood literally at all.
- 10 means: this phrase is to be understood very literally.
- Use values in between to grade your decision. Please, however, try to *take a stand as often as possible.*

In case the context is unclear or nonsensical, please enter ”66” and use the comment field to explain. However, please try to make sense of it even if the sentences are incomplete.

Example 1 :
There was a red truck parked curbside. It looked like someone was living in it.
YOUR ANSWER: 10
reason: the color of the truck is red, this can be inferred from the parts ”red” and ”truck” only - without any special knowledge.

? Example 2 :
What a tour! We were on cloud nine when we got back to headquarters but we kept our mouths shut.
YOUR ANSWER: 0
reason: ”cloud nine” means to be blissfully happy. It does NOT refer to a cloud with the number nine.

Example 3 :
Yellow fever is found only in parts of South America and Africa.
YOUR ANSWER: 7
reason: ”yellow fever” refers to a disease causing high body temperature. However, the fever itself is not yellow. Overall, this phrase is fairly literal, but not totally, hence answering with a value between 5 and 8 is appropriate.

We take rejection seriously and will not reject a HIT unless done carelessly. Entering anything else but numbers between 0 and 10 or 66 in the judgment field will automatically trigger rejection.

YOUR CONTEXT with **big day**
Special Offers : Please call FREEPHONE 0800 0762205 to receive your free copy of ’ Groom ’ the full colour magazine dedicated to dressing up for the **big day** and details of Moss Bros Hire rates .

How literal is the bolded phrase in the context above between 0 and 10?
[ ]

OPTIONAL: leave a comment, tell us about what is broken, help us to improve this type of HIT:
[ ]

Figure 2: Sample Human Intelligence Task on Amazon Turk with annotation instructions

EN	ADJ_NN	V_SUBJ	V_OBJ	Sum
Train	58 (43)	30 (23)	52 (41)	140 (107)
Vali.	10 (7)	9 (6)	16 (13)	35 (26)
Test	77 (52)	35 (26)	62 (40)	174 (118)
All	145 (102)	74 (55)	130 (94)	349 (251)

Table 1: English dataset: number of target phrases (with coarse scores)

DE	ADJ_NN	V_SUBJ	V_OBJ	Sum
Train	49 (42)	26 (23)	44 (33)	119 (98)
Vali.	11 (8)	9 (8)	9 (7)	29 (23)
Test	63 (48)	29 (28)	57 (44)	149 (120)
All	123 (98)	64 (59)	110 ()	297 (241)

Table 2: German dataset: number of target phrases (with coarse scores)

Additionally to the numerical scores, we've also provided coarse-grained labels. This is motivated by the following: for some applications, it is probably enough to decide whether a phrase is always compositional, somewhat compositional or usually not compositional, without the need of more fine-grained distinctions. For this, we've transformed the numerical scores in the range of 0-25 to coarse label "low", those between 38-62 have been labeled as "medium", and the ones from 75 to 100 have received the value "high". All other phrases have been excluded from the corresponding training and test datasets for "coarse evaluation" (s. Section 2.4.2): 28.1% of English and 18.9% of German phrases.

2.3 Datasets

Now we describe the datasets in detail. Table 1 summarizes the English data, Table 2 describes the German data quantitatively. Per language and relation, the data was randomly split in approximatively 40% training, 10% validation and 50% test.

2.4 Scoring of system responses

We provided evaluation scripts along with the training and validation data. Additionally, we report correlation values (Spearman's rho and Kendall's tau) in Section 4.

2.4.1 Numerical Scoring

For numerical scoring, the evaluation script computes the distance between the system responses $S = \{s_{target1}, s_{target2}, ...s_{targetN}\}$ and the gold standard $G = \{g_{target1}, g_{target2}, ...g_{targetN}\}$ in points, averaged over all items:

$NUMSCORE(S, G) = \frac{1}{N} \sum_{i=1..N} |g_i - s_i|$.

Missing values in the system scores are filled with the default value of 50. A perfect score is 0, indicating no difference between the system responses and the gold standard.

2.4.2 Coarse Scoring

We use precision on coarse label predictions for coarse scoring:

$$COARSE(S, G) = \frac{1}{N} \sum_{i=1..N} \begin{cases} s_i == g_i : 1 \\ otherwise : 0 \end{cases}.$$

As with numerical scoring, missing system responses are filled with a default value, in this case 'medium'. A perfect score would be 1.00, connoting complete congruence of gold standard and system response labels.

3 Participants

Seven teams participated in the shared task. Table 3 summarizes the participants and their systems. Four of the teams (Duluth, UoY, JUCSE, SCSS-TCD) submitted three runs for the whole English test set. One team participated with two systems, one of which was for the entire English dataset and another one included entries only for English V_SUBJ and V_OBJ relations. A team from UNED provided scores solely for English ADJ_NN pairs. UCPH was the only team that delivered results for both English and German.

Systems can be split into approaches based on statistical association measures and approaches based on word space models. On top, some systems used a machine-learned classifier to predict numerical scores or coarse labels.

4 Results

The results of the official evaluation for English are shown in Tables 4 and 5.

Table 4 reports the results for numerical scoring. *UCPH-simple.en* performed best with the score of 16.19. The second best system *UoY: Exm-Best* achieved 16.51, and the third was *UoY:Pro-Best* with 16.79. It is worth noting that the top six systems

Systems	Institution	Team	Approach
Duluth-1 Duluth-2 Duluth-3	Dept. of Computer Science, University of Minnesota	Ted Pedersen	statistical association measures: t-score and pmi
JUCSE-1 JUCSE-2 JUCSE-3	Jadavpur University	Tanmoy Chakraborty, Santanu Pal Tapabrata Mondal, Tanik Saikh, Sivaju Bandyopadhyay	mix of statistical association measures
SCSS-TCD:conf1 SCSS-TCD:conf2 SCSS-TCD:conf3	SCSS, Trinity College Dublin	Alfredo Maldonado-Guerra, Martin Emms	unsupervised WSM, cosine similarity
submission-ws submission-pmi	Gavagai	Hillevi Hägglöf, Lisa Tengstrand	random indexing association measures (pmi)
UCPH-simple.en	University of Copenhagen	Anders Johannsen, Hector Martinez, Christian Rishøj, Anders Søgaard	support vector regression with COALS-based endocentricity features
UoY: Exm UoY: Exm-Best UoY: Pro-Best	University of York, UK; Lexical Computing Ltd., UK	Siva Reddy, Diana McCarthy, Suresh Manandhar, Spandana Gella	exemplar-based WSMs prototype-based WSM
UNED-1: NN UNED-2: NN UNED-3: NN	NLP and IR Group at UNED	Guillermo Garrido, Anselmo Peas	syntactic VSM, dependency-parsed UKWaC, SVM classifier

Table 3: Participants of DiSCo'2011 Shared Task

in the numerical evaluation are all based on different variations of word space models.

The outcome of evaluation for coarse scores is displayed in Table 5. Here, *Duluth-1* performs highest with 0.585, followed closely by *UoY:ExmBest with 0.576* and *UoY: ProBest* with 0.567. *Duluth-1* is an approach purely based on association measures.

Both tables also report ZERO-response and RANDOM-response baselines. ZERO-response means that, if no score is reported for a phrase, it gets a default value of 50 (fifty) points in numerical evaluation and 'medium' in coarse evaluation. Random baselines were created by using random labels from a uniform distribution. Most systems beat the RANDOM-response baseline, only about half of the systems are better than ZERO-response.

Apart from the officially announced scoring methods, we provide Spearman's rho and Kendall's tau rank correlations for numerical scoring. Rank correlation scores that are not significant are noted in parentheses. With correlations, the higher the score, the better is the system's ability to order the phrases according to their compositionality scores. Here, systems *UoY: Exm-Best*, *UoY: Pro-Best* / *JUCSE-1* and *JUCSE-2* achieved the first, second and third best results respectively.

Overall, there is no clear winner for the English dataset. However, across different scoring mechanisms, *UoY: Exm-Best* is the most robust of the systems. The *UCPH-simple.en* system has a stellar performance on V_OBJ but apparently uses a suboptimal way of assigning coarse labels. The *Duluth-1* system, on the other hand, is not able to produce a numerical ranking that is significant according to the correlation measures, but excels in the coarse scoring.

When comparing word space models and association measures, it seems that the former do a slightly better job on modeling graded compositionality, which is especially obvious in the numerical evaluation.

Since word space models and statistical association measures are language-independent approaches and most teams have not used syntactic preprocessing other than POS tagging, it is a pity that only one team has tried the German task (see Tables 6 and 7). The comparison to the baselines shows that the *UCPH* system is robust across languages and performs (relatively speaking) equally well in the numerical scoring both for the German and the English tasks.

numerical scores	responses	ρ	τ	EN all	EN_ADJ_NN	EN_V_SUBJ	EN_V_OBJ
number of phrases				174	77	35	62
0-response baseline	0	-	-	23.42	24.67	17.03	25.47
random baseline	174	(0.02)	(0.02)	32.82	34.57	29.83	32.34
UCPH-simple.en	174	0.27	0.18	**16.19**	14.93	21.64	**14.66**
UoY: Exm-Best	169	**0.35**	**0.24**	16.51	15.19	**15.72**	18.6
UoY: Pro-Best	169	0.33	0.23	16.79	**14.62**	18.89	18.31
UoY: Exm	169	0.26	0.18	17.28	15.82	18.18	18.6
SCSS-TCD: conf1	174	0.27	0.19	17.95	18.56	20.8	15.58
SCSS-TCD: conf2	174	0.28	0.19	18.35	19.62	20.2	15.73
Duluth-1	174	(-0.01)	(-0.01)	21.22	19.35	26.71	20.45
JUCSE-1	174	0.33	0.23	22.67	25.32	17.71	22.16
JUCSE-2	174	0.32	0.22	22.94	25.69	17.51	22.6
SCSS-TCD: conf3	174	0.18	0.12	25.59	24.16	32.04	23.73
JUCSE-3	174	(-0.04)	(-0.03)	25.75	30.03	26.91	19.77
Duluth-2	174	(-0.06)	(-0.04)	27.93	37.45	17.74	21.85
Duluth-3	174	(-0.08)	(-0.05)	33.04	44.04	17.6	28.09
submission-ws	173	0.24	0.16	44.27	37.24	50.06	49.72
submission-pmi	96	-	-	-	-	52.13	50.46
UNED-1: NN	77	-	-	-	17.02	-	-
UNED-2: NN	77	-	-	-	17.18	-	-
UNED-3: NN	77	-	-	-	17.29	-	-

Table 4: Numerical evaluation scores for English: average point difference and correlation measures (not significant values in parentheses)

coarse values	responses	EN all	EN_ADJ_NN	EN_V_SUBJ	EN_V_OBJ
number of phrases		118	52	26	40
zero-response baseline	0	0.356	0.288	0.654	0.250
random baseline	118	0.297	0.288	0.308	0.300
Duluth-1	118	**0.585**	0.654	0.385	0.625
UoY: Exm-Best	114	0.576	0.692	0.500	0.475
UoY: Pro-Best	114	0.567	**0.731**	0.346	0.500
UoY: Exm	114	0.542	0.692	0.346	0.475
SCSS-TCD: conf2	118	0.542	0.635	0.192	**0.650**
SCSS-TCD: conf1	118	0.534	0.64	0.192	0.625
JUCSE-3	118	0.475	0.442	0.346	0.600
JUCSE-2	118	0.458	0.481	0.462	0.425
SCSS-TCD: conf3	118	0.449	0.404	0.423	0.525
JUCSE-1	118	0.441	0.442	0.462	0.425
submission-ws	117	0.373	0.346	0.269	0.475
UCPH-simple.en	118	0.356	0.346	0.500	0.275
Duluth-2	118	0.322	0.173	0.346	0.500
Duluth-3	118	0.322	0.135	**0.577**	0.400
submission-pmi	-	-	-	0.346	0.550
UNED-1-NN	52	-	0.289	-	-
UNED-2-NN	52	-	0.404	-	-
UNED-3-NN	52	-	0.327	-	-

Table 5: Coarse evaluation scores for English

numerical scores	responses	ρ	τ	DE all	DE_ADJ_NN	DE_V_SUBJ	DE_V_OBJ
number of phrases				149	63	29	57
0-response baseline	0	-	-	32.51	32.21	38.00	30.05
random baseline	149	(0.005)	(0.004)	37.79	36.27	47.45	34.54
UCPH-simple.de	148	0.171	0.116	24.03	27.09	15.55	24.06

Table 6: Numerical evaluation scores for German

heightcoarse values	responses	DE all	DE_ADJ_NN	DE_V_SUBJ	DE_V_OBJ
number of phrases		120	48	28	44
0-response baseline	0	0.158	0.208	0.071	0.159
random baseline	120	0.283	0.313	0.214	0.295
UCPH-simple.de	119	0.283	0.375	0.286	0.182

Table 7: Coarse evaluation scores for German

For more details on the systems as well as fine-grained analysis of the results, please consult the corresponding system description papers.

5 Conclusion

DiSCo Shared Task attracted seven groups that submitted results for 19 systems. We consider this a success, taking into consideration that the task is new and difficult. The opportunity to evaluate language-independent models for languages other than English was unfortunately not taken up by most participants.

The teams applied a variety of approaches that can be classified into lexical association measures and word space models of various flavors. From the evaluation, it is hard to decide what method is currently more suited for the task of automatic acquisition of compositionality, with a slight favor for approaches based on word space model.

A takeaway message is that a pure corpus-based acquisition of graded compositionality is a hard task. While some approaches clearly outperform baselines, further advances are needed for automatic systems to be able to reproduce semantic compositionality automatically.

Acknowledgments

We thank Emiliano Guevara for helping with the preparation of the evaluation scripts and the initial task description. This work was partially supported by the German Federal Ministry of Economics (BMWi) under the project Theseus (number 01MQ07019).

References

Colin Bannard, Timothy Baldwin, and Alex Lascarides. 2003. A statistical approach to the semantics of verb-particles. In *Proc. of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 65–72, Sapporo, Japan.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*.

Chris Biemann and Valerie Nygaard. 2010. Crowdsourcing WordNet. In *Proc. of the 5th International Conference of the Global WordNet Association (GWC-2010)*, Mumbai, India.

Graham Katz and Eugenie Giesbrecht. 2006. Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In *Proceedings of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties*, pages 12–19, Sydney, Australia.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15, Mexico City, Mexico.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.