How Naked is the Naked Truth?
A Multilingual Lexicon of Nominal Compound Compositionality

Carlos Ramisch[1], **Silvio Cordeiro**[1,2], **Leonardo Zilio**[2]
Marco Idiart[3], **Aline Villavicencio**[2], **Rodrigo Wilkens**[2]
[1] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[3] Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr lzilio@inf.ufrgs.br
marco.idiart@gmail.com avillavicencio@inf.ufrgs.br rswilkens@inf.ufrgs.br

Abstract

We introduce a new multilingual resource containing judgments about nominal compound compositionality in English, French and Portuguese. It covers 3 × 180 noun-noun and adjective-noun compounds for which we provide numerical compositionality scores for the head word, for the modifier and for the compound as a whole, along with possible paraphrases. This resource was constructed by native speakers via crowdsourcing. It can serve as basis for evaluating tasks such as lexical substitution and compositionality prediction.

1 Introduction

Multiword expressions (MWEs) are notoriously challenging for NLP, due to their many potential levels of idiosyncrasy, from lexical to semantic and pragmatic to statistical (Sag et al., 2002; Ramisch, 2015). One widely known problem is the semantic interpretation of noun compounds, which in English are noun phrases composed by a sequence of nouns. These MWEs often lack a structure from which to identify implicit semantic relations unambiguously. For instance, there is no indication that a *brick wall* is a wall *made of* bricks, while a *cheese knife* is not a knife *made of* cheese, but rather a knife *for cutting* cheese (Girju et al., 2005).

Noun compounds are often idiomatic or non-compositional. That is, the meaning of the whole does not come directly from the meaning of the parts. For instance, a *black Friday* is not any Friday that is somehow black, but is the day following Thanksgiving Day in the United States. Moreover, the contribution of the semantics of each element for the meaning of the compound may vary considerably (e.g. *police car* vs. *crocodile tears*). Any NLP application that intends to deal with phrasal semantics adequately must be able to distinguish fairly compositional from fully idiomatic compounds. For example, automatically translating *dead end* literally into French (?*fin morte*) or Portuguese (?*fim morto*) would drastically alter the meaning of the original expression. In this paper we introduce a resource with human judgments about the semantics of compounds and their individual elements.

Eliciting quantitative judgments about compositionality from non-linguists may be too abstract, even with accompanying guidelines and training. We propose a more constrained way of obtaining these judgments, with the participation of non-experts through crowdsourcing. We first focus the participants' attention on compound interpretation in context, by requesting paraphrases in example sentences. Then, we inquire about the degree to which the meaning of a given compound arises from each of its elements. The assumption is that if the interpretation of the compound comes from both nouns (e.g. *access road*), then it is fully compositional, whereas if it is unrelated to both nouns (e.g. *nut case*), then it is fully idiomatic. This indirect annotation does not require expert knowledge and provides reliable and stable data.

This paper presents a multilingual resource that models compounds compositionality, including both numerical scores and free paraphrases. Data is currently available for 180 compounds in 3 different languages: English, French and Portuguese. Such resources are extremely valuable, as they enable the development and evaluation of techniques for automatic compositionality prediction and lexical substitution. This paper is structured as follows: §2 discusses related work; §3 discusses the target compounds, the annotation schema and interface; §4 presents the results and §5 the conclusions and future work.

2 Related Work

There are many proposals in the literature to represent the semantics of nominal compounds. Lauer (1995) argues that prepositions (such as *from*, *for*, *in*) provide information about the role of each noun in a compound (e.g. *olive oil* is *oil from olives*). These prepositions are explicitly part of some nominal compounds in Romance languages (e.g. *huile* ***d'****olive* in French and *azeite* ***de*** *oliva* in Portuguese). Girju et al. (2005) present and compare several inventories of semantic relations between nouns, from fine-grained to coarse senses. These relations include syntactic and semantic classes such as *subject*, *instrument* and *location*. Free paraphrases have also been used to model noun compound semantics. Nakov (2008) suggests using unsupervised generation of paraphrases combined with web

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 156–161,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

search engines to classify nominal compounds. This was further extended in SemEval 2013, in a task where free paraphrases were ranked according to their relevance for explicitly describing the underlying semantic relations in the compounds (Hendrickx et al., 2013). For instance, for the MWE *flu virus*, paraphrases involving the verbs *cause, spread* and *create* (*virus that causes/spreads/creates flu*) were in the top of the rank.

Some authors model the meaning of compounds using numerical compositionality scores: low values mean completely idiomatic compounds while high values represent compositional ones. Separate scores can be provided for the amount of meaning provided by each individual word. For instance, *olive oil* could be 80% related to *olives* and 100% related to *oil*, whereas *dead end* is 5% *dead* and 90% an *end*. Some datasets that employ a numerical representation for different types of MWE are:

- Baldwin and Villavicencio (2002): binary type-level judgments for 3,078 English phrasal verbs, from which 14% are considered idiomatic.
- McCarthy et al. (2003): type-based scores on a scale from 0 to 10 provided by three experts for 116 English phrasal verbs.
- Reddy et al. (2011): average of 30 judgments on a scale from 0 to 5 provided by native speakers via crowdsourcing for 90 English noun compounds.
- Gurrutxaga and Alegria (2013): three-way classification (idiom, collocation, free combination) provided by three experts for 1,200 Basque noun-verb expressions.
- Roller et al. (2013): average of around 30 judgments on a scale from 1 to 7 obtained through crowdsourcing for 244 German noun compounds.
- Farahmand et al. (2015): individual binary judgments for non-compositionality and conventionality for 1,042 English noun compounds, annotated by 4 experts.

One possible source of divergence among annotators is that some datasets do not take polysemy into account. Authors ask annotators to think about the most common sense of an MWE without providing context. Some of these datasets address this issue by providing example sentences to attenuate this problem. We also employ this strategy in our questionnaires. The most similar datasets to ours are the ones presented by Reddy et al. (2011) and Hendrickx et al. (2013). Our dataset combines the methodology from both of these, extending it to French and Portuguese.

3 Dataset Construction

Although noun-noun compounds are rare in some languages mainly due to syntactic reasons, these languages present alternatives to this type of configuration. In French (FR) and Brazilian Portuguese (PT), the equivalents of English (EN) compounds of the form N_1 N_2 are usually:

1. N_2 PREP N_1, connecting the nouns through a preposition and optional determiner; e.g. *lung cancer* (EN) → *cancer du poumon* (FR), *câncer de pulmão* (PT).
2. N_2 ADJ_1, using a denominal adjective which is derived from N_1; e.g. *cell death* (EN) → *mort cellulaire* (FR), *morte celular* (PT).

We describe the construction of datasets for English, French and Brazilian Portuguese. Given the two syntactic forms above, we focus on N_2 ADJ_1 for French and Portuguese, as its simpler structure resembles more closely the English noun-noun compound structure, and also because we have some ADJ_1 N_2 compounds in English as well (e.g. *sacred cow*). We collectively call our target constructions *nominal compounds*, as they have nouns as head of the phrase.

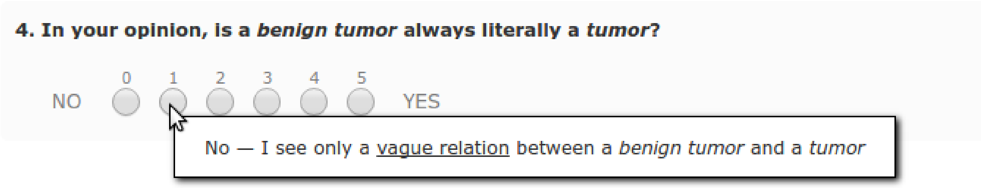
For each language, data collection involves the following steps: (1) compound selection; (2) sentence selection; and (3) questionnaire design.

Compound selection The initial set of idiomatic and partially compositional candidates was constructed by introspection, independently for each language, since these may be harder to find in corpora because of lower frequency. This list of compounds was complemented by selecting entries from lists of frequent *adjective+noun* and *noun+noun* pairs. These were automatically extracted through POS-sequence queries using the mwetoolkit (Ramisch, 2015) from ukWaC (Baroni et al., 2009), frWaC and brWaC (Boos et al., 2014). We removed all compounds in which the complement is not an adjective in Portuguese/French (e.g. PT noun-noun *abelha rainha*), those in which the head is not necessarily a noun (e.g. FR *aller simple*, as *aller* is also a verb) and those in which the literal sense is very common in the corpus (e.g. EN *low blow*). For each language, we attempted to select a balanced set of 60 idiomatic, 60 partially compositional and 60 fully compositional compounds by rough manual pre-annotation.[1]

Sentence selection For each compound, we selected 3 sentences from a WaC corpus where the compound is used with the same meaning. These sentences are used during the data collection process (described later) as disambiguating context for the annotators. We sort them by sentence length, in order to favor shorter sentences, and manually select 3 examples that satisfy these criteria:

- The occurrence of the compound must have the same meaning in all sentences.
- A sentence must contain enough context to enable mental disambiguation of the compound.
- Inter-sentence variability can be used to provide more information to the reader.

[1] We have not attempted to select equivalent compounds for all three languages. A compound in a given language may correspond to a single word in the other languages. Even when it does translate as a compound, its POS pattern and level of compositionality may be widely different.



4. In your opinion, is a *benign tumor* always literally a *tumor*?

NO ○0 ○1 ○2 ○3 ○4 ○5 YES

No — I see only a vague relation between a *benign tumor* and a *tumor*

Figure 1: Evaluating compositionality regarding a compounds' head.

Questionnaire design We collect data for each compound through a separate HIT (Human Intelligence Task). Each HIT page contains a list of instructions followed by the questionnaire associated with that compound. In the instructions, we briefly describe the task and require that the users fill in an external identification form, following Reddy et al. (2011). This form provides us with demographics about the annotators, ensuring that they are native speakers of the target language. At the end of the form, they are also given extra example questions with annotated answers for training. After filling in the identification form, users can start working on the task. This section of the HIT is structured in 5 subtasks:

1. Read the compound itself.
2. Read 3 sentences containing the compound.
3. Provide 2 to 3 synonym expressions for the target compound seen in the sentences.
4. Using a Likert scale from 0 to 5, judge how much of the meaning of the compound comes from word$_1$ (mod) and word$_2$ (head) separately, as shown in Figure 1.
5. Using a Likert scale from 0 to 5, judge how much of the meaning of the compound (comp) comes from its components.

We have been consciously careful about requiring answers in an even-numbered scale (0–5 makes for 6 reply categories), as otherwise, undecided annotators would be biased towards the middle score. As an additional help for the annotators, when the mouse hovers over a reply to a multiple-choice question, we present a guiding tooltip, as in Figure 1. We avoid incomplete HITs by making Subtasks 3–5 mandatory.

The order of subtasks has also been taken into account. During a pilot test, we found that presenting the multiple-choice questions (Subtasks 4–5) before asking for synonyms (Subtask 3) yielded lower agreement, as users were often less self-consistent in the multiple-choice questions (e.g. replying "non-compositional" for Subtask 4 but "compositional" for Subtask 5), even if they carefully selected their synonyms in response to Subtask 3. Asking for synonyms prior to the multiple-choice questions helps the user focus on the target meaning for the compound and also have more examples (the synonyms) when considering the semantic contribution of each element of the compound.

For EN and FR, annotators were recruited and paid via Amazon Mechanical Turk. The quality of FR results was manually controlled by only accepting HITs with reasonable paraphrases. During a pilot, we noticed the lack of qualified PT native speakers on the platform. For PT only, judgments were provided by volunteers through a standalone web interface that simulated the HIT page.

4 Results

For each compound, we have collected judgments from around 15 HITs. The average of these scores, for EN[2], FR and PT, are shown in Figure 2. The compositionality judgments for the compounds confirm that they are balanced with respect to idiomaticity. Moreover, there seems to be a greater agreement between the score for the compound and that of its head (or modifier) for the two extremes (totally idiomatic and fully compositional). For PT and FR, in particular, the compound score seems to be a lower bound to each member word's score.

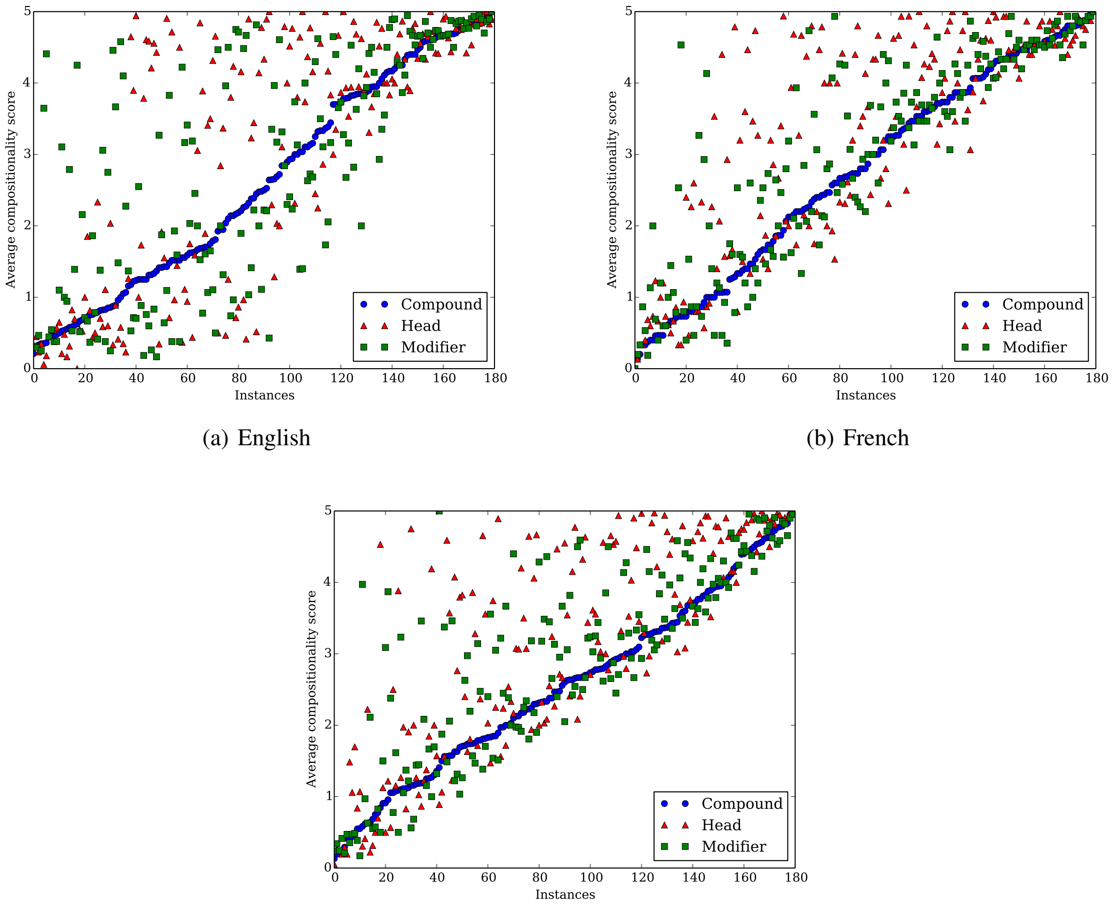
We also looked at the distribution of each of the scores around the mean in terms of the standard deviation (σ). Ideally, if all the annotators agreed on compositionality, σ should be low. We calculated for each language the number of compounds, heads and modifiers with standard deviations greater than 1.5 (Table 1). The largest variations are for modifiers, which may reflect their potentially accessory role in the meaning of the compound in relation to the head.

	EN	FR	PT
Pearson r head-compound	0.75	0.81	0.80
Pearson r mod-compound	0.74	0.89	0.84
compound $\sigma > 1.5$	22	41	30
head $\sigma > 1.5$	23	44	33
modifier $\sigma > 1.5$	35	55	34

Table 1: Pearson correlation r and number of cases of high standard deviation σ.

Out of all human judges, 3 of them annotated a large subset of 119 compounds in PT. For this subset, we report inter-annotator agreement. Pairwise weighted κ values range from .28 to .58 depending on the question (head, mod or comp) and on the annotator pair. Multi-rater α agreement (Artstein and Poesio, 2008) values are $\alpha = .52$ for head, $\alpha = .36$ for mod and $\alpha = .42$ for comp scores. We have also calculated the α score of an expert annotator with himself, performing the same task a few weeks later. The score ranges from

[2] We include the 90 compounds from Reddy et al. (2011), which are compatible with the new dataset.



(a) English

(b) French

(c) Portuguese

Figure 2: Average compositionality for compounds, heads and modifiers.

0.59 for modifiers and compounds to 0.69 for heads. This seems to confirm the hypothesis that modifiers are harder to annotate than heads.

Table 2 presents the most controversial compounds, along with the ones that had highest agreement (lowest σ). The most consensual compounds are mostly 100% compositional and sometimes 100% idiomatic. Low σ values are consistent among the three questions, indicating that some compounds are simply easier to judge than others.

There are multiple reasons for divergences in the judgment scores. For some MWEs, our sentences were not enough for disambiguation; e.g. one of the *fish story* sentences talked about a whale and prompted literal interpretations of *fish* for some judges). Other differences have been caused by the interpretation of uncommon words; e.g. the PT noun *olhado* does not appear by itself very often; some judges seem to have interpreted it as an adjective and thus concluded that *mau-olhado* (*evil eye*, lit. *bad-glance*) has a fully non-compositional head. Finally, some differences have been caused by whether speakers had incorporated a new meaning into their lexicon; e.g. EN speakers agreed on the level of head and head+modifier compositionality for *dirty word*, but disagreed when judging the modifier: it is fully idiomatic for some, while just containing an uncommon sense of *dirty* for others.

5 Conclusions and Future Work

We presented a multilingual dataset of nominal compounds containing human judgments about compositionality. It contains 180 compounds for each of the 3 target languages: English, French and Portuguese. Annotations are collected through crowdsourcing. Since the task is performed by native speakers who may not have a background in linguistics, it needs to be appropriately constrained not to require expert knowledge. The resulting resource can be used for applications and tasks involving some degree of semantic processing, such as lexical substitution and text simplification. For the cases where the numerical judgments alone are not enough for a given task, our dataset also provides sets of paraphrases, which serve as a symbolic counterpart to those scores. The complete resource will be made freely available.[3] As future work, we plan to validate these scores through compositionality prediction (Yaz-

[3] http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds

	compound	head	mod	comp
English	brass ring	3.9 ±2.0	3.7 ±1.9	3.7 ±1.8
	fish story	4.8 ±0.4	1.5 ±1.8	1.7 ±1.8
	tennis elbow	4.3 ±1.3	2.2 ±1.8	2.5 ±1.8
	brick wall	3.5 ±1.9	3.2 ±2.2	3.8 ±1.7
	dirty word	4.1 ±1.4	2.0 ±1.4	2.5 ±1.7
	prison guard	4.8 ±0.4	4.9 ±0.3	4.9 ±0.3
	graduate student	5.0 ±0.0	4.7 ±0.5	4.9 ±0.3
	engine room	5.0 ±0.0	4.9 ±0.3	4.9 ±0.3
	climate change	4.8 ±0.4	4.9 ±0.3	5.0 ±0.2
	insurance company	4.9 ±0.5	5.0 ±0.0	5.0 ±0.0
French	match nul	4.4 ±1.3	2.2 ±2.3	2.5 ±2.1
	mort né	4.6 ±1.1	3.5 ±1.8	3.2 ±2.0
	carte grise	4.5 ±0.9	3.2 ±2.0	3.1 ±1.9
	second degré	1.7 ±1.9	2.4 ±2.1	1.4 ±1.9
	grippe aviaire	4.6 ±1.4	3.8 ±1.9	3.6 ±1.9
	eau chaude	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	eau potable	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	matière grasse	4.8 ±0.4	5.0 ±0.0	5.0 ±0.0
	poule mouillée	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0
	téléphone portable	4.9 ±0.5	4.9 ±0.3	5.0 ±0.0
Portuguese	pavio curto	1.6 ±1.8	1.1 ±1.9	1.9 ±2.3
	sexto sentido	4.0 ±1.4	2.5 ±2.1	2.8 ±2.2
	gelo-seco	3.2 ±1.6	3.2 ±1.8	3.0 ±2.1
	mau-olhado	1.8 ±1.2	4.2 ±1.5	2.3 ±2.1
	câmara fria	3.6 ±2.2	5.0 ±0.0	3.4 ±2.1
	núcleo atômico	5.0 ±0.0	4.4 ±1.8	5.0 ±0.0
	pão-duro	0.0 ±0.0	1.0 ±1.7	0.0 ±0.0
	sentença judicial	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	tartaruga-marinha	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0
	vôo internacional	5.0 ±0.0	5.0 ±0.0	5.0 ±0.0

Table 2: Most polemic and consensual compounds in each language (average±σ score).

dani et al., 2015; Salehi et al., 2015) and by incorporating the scores and paraphrases into a machine translation system. We also envisage extending the dataset for each of the languages and for additional languages.

Acknowledgments

This work has been partly funded by projects PARSEME-FR (ANR-14-CERA-0001), AIM-WEST (FAPERGS-INRIA 1706- 2551/13-7), CNPq 482520/2012-4, 312114/2015-0, "Simplificação Textual de Expressões Complexas", sponsored by Samsung Eletrônica da Amazônia Ltda. under the terms of Brazilian federal law No. 8.248/91.

References

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Timothy Baldwin and Aline Villavicencio. 2002. Extracting the unextractable: A case study on verb-particles. In *Proceedings of CoNLL 2002*, COLING-02, pages 1–7. ACL.

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*, 43(3):209–226, September.

Rodrigo Boos, Kassius Prestes, and Aline Villavicencio. 2014. Identification of multiword expressions in the brWaC. In *Proceedings of LREC 2014*, pages 728–735. ELRA, May. ACL Anthology Identifier: L14-1429.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression data set: Annotating non-compositionality and conventionalization for english noun compounds. In *Proceedings of NAACL-HLT*, pages 29–33.

Roxana Girju, Dan Moldovan, Marta Tatu, and Daniel Antohe. 2005. On the semantics of noun compounds. *Computer speech & language*, 19(4):479–496.

Antton Gurrutxaga and Iñaki Alegria. 2013. Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 116–125. ACL, June.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. Semeval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of *SEM 2013, Volume 2 – SemEval*, pages 138–143. ACL, June.

Mark Lauer. 1995. How much is enough?: Data requirements for statistical NLP. *CoRR*, abs/cmp-lg/9509001.

Diana McCarthy, Bill Keller, and John Carroll. 2003. Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions*, MWE '03, pages 73–80. ACL.

Preslav Nakov. 2008. Paraphrasing verbs for noun compound interpretation. In *Proc. of the LREC Workshop Towards a Shared Task for MWEs (MWE 2008)*, pages 46–49.

Carlos Ramisch. 2015. *Multiword Expressions Acquisition - A Generic and Open Framework*. Theory and Applications of Natural Language Processing. Springer.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An empirical study on compositionality in compound nouns. In *Proceedings of IJCNLP 2011*, November.

Stephen Roller, Sabine Schulte im Walde, and Silke Scheible. 2013. The (un)expected effects of applying standard cleansing models to human ratings on compositionality. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 32–41. ACL, June.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of CICLing-2002*, pages 1–15.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2015. A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of NAACL-HLT 2015*, pages 977–983. ACL, May–June.

Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of EMNLP 2015*, pages 1733–1742. ACL, September.