LexSubNC: a Dataset of Lexical Substitution for Nominal Compounds

Rodrigo Wilkens[1], Leonardo Zilio[1], Silvio Cordeiro[2,3],
Felipe S. F. Paula[2], Carlos Ramisch[3], Marco Idiart[4], Aline Villavicencio[2]

[1]CENTAL, Université catholique de Louvain (Belgium)
[2]Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
[3]Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[4]Institute of Physics, Federal University of Rio Grande do Sul (Brazil)
{rodrigo.wilkens,leonardo.zilio}@uclouvain.be
{silvio.cordeiro,carlos.ramisch}@lif.univ-mrs.fr
{felipesfpaula,marco.idiart,alinev}@gmail.com

Abstract

In the context of NLP tasks such as text simplification, lexicons containing information about semantically related words are an important resource for evaluating the quality of the system output. Existing resources containing lexical substitutes have been built with a focus on single words. In this paper, we present a lexical substitution dataset for Portuguese nominal compounds. The compounds have varying degrees of compositionality, conventionality and frequency, and we investigate the impact of these characteristics on the suggestions of lexical substitution made by native speakers. No strong correlations are found for these factors on the number or type of responses provided. However, a significant effect of compositionality is found in the use of one of the component words (head or modifier) as a substitute. The resulting resource, LexSubNC, contains over 1,500 manually validated substitutes for 180 compounds, further classified according to the type of response.

1 Introduction

In tasks involving lexical substitution, alternatives need to be identified for a given target word (McCarthy and Navigli, 2007, 2009), usually in a particular context. Candidates can be chosen to maximize word properties that are relevant for the particular task, such as unigram and n-gram frequencies, concreteness, imageability, and conventionality. Depending on the target word, more than one possible alternative substitution may fulfill the criteria and produce an acceptable result (e.g. *acquire/buy/purchase a painting*). Resources such as thesauri, containing semantically related words (Fellbaum, 1998; Lin, 1998), and word norms, with information about word properties (Nelson et al., 2004), may be used to inform these tasks.

Various initiatives for collecting word norms resulted in datasets such as the South Florida Association Norms (Nelson et al., 2004), SimLex-999 (Hill et al., 2015), Hyperlex (Vulić et al., 2016) and Rare Words (Luong et al., 2013). These resources form valuable gold standards for evaluating the quality of a variety of tasks and applications, including text simplification and machine translation. However, they often concentrate on single words as targets.

The collection of norms for longer units is particularly challenging due to the arbitrary interactions between their member words, particularly if they involve multiword expressions (MWEs), such as nominal compounds or verbal idioms. Available MWE datasets often target specific types of MWEs such as verb-particle constructions (McCarthy et al., 2003) and noun compounds (Reddy et al., 2011), and tend to focus on numerical scores that model compositionality and conventionality. Resources with lexical substitutes or paraphrases for MWEs are rare and often only include compositional expressions (Hendrickx et al., 2013).

However, MWEs may also involve some degree of semantic or statistical idiosyncrasy with respect to regular combinations (Baldwin and Kim, 2010) and these may have an impact on the quality of the collected data. For instance, there may be less agreement among annotators for an idiomatic nominal compound like *Black Friday* as it may be perceived as being related to various different concepts like *Friday*, *promotion* and *Thanksgiving*, which may all be possible substitutes for the compound but are not synonyms among themselves.

In this paper we introduce LexSubNC, a dataset that contains the responses of human annotators about lexical substitutes for a set of nominal compounds of varying degrees of compositionality, frequency and conventionality in Brazilian Portuguese. The raw data was collected using a dedicated web interface, allowing the participation of many volunteer non-expert native speakers. The responses were then manually validated and classified according to the particular semantic relations involved. We examine the impact of factors like frequency, conventionality and compositionality on the number and type of responses collected for the construction of the dataset.[1]

LexSubNC is potentially useful for the evaluation and development of several NLP tasks and applications. For example, it could be used to tune the development of distributional semantic models that maximize the similarity between an MWE and its substitutes, similarly to what is currently done for single words (Hill et al., 2015; Levy et al., 2015). It could also be used for the evaluation of machine translation methods that focus on non-compositional expressions, similarly to what is currently done for instance in METEOR (Denkowski and Lavie, 2014). For automatic text simplification, paraphrases could be used to replace non-compositional expressions by more explicit paraphrases (Specia et al., 2012).

This paper is structured as follows: we discuss similar resources and the techniques used to collect them (§2), and describe the protocol used for collecting human responses (§3). The responses are analyzed for possible correlations between characteristics of the compounds and the responses provided (§4). We finish with conclusions and a discussion of future work (§5).

2 Related Work

A variety of protocols have been adopted for collecting specific word norms, usually targeting single words. For the lexical substitution of single words in English, McCarthy and Navigli (2007) asked annotators to provide up to three substitutes, preferably also single words, for 210 target words (nouns, verbs, adjectives and adverbs) in 10 sentences each, in a total of 2.100 sentences. They adopted no fixed inventory of words for the task, and the results reflected this inherent variation, as several substitutes are possible for a specific target in a particular context. This dataset was later used as the starting point for rating the alternatives provided in terms of simplicity (Specia et al., 2012).

Lexical substitution datasets for languages other than English include a dataset for German containing substitutes for 153 target words (51 nouns, 51 adjectives, and 51 verbs) in 2,040 sentences (Cholakov et al., 2014). Cholakov et al. (2014) also collected information about the difficulty of the annotation from a pilot study, where 78% of the tasks were considered of easy or medium difficulty. For multilingual lexical substitution Mihalcea et al. (2010) asked annotators to find Spanish alternatives for English words in a given context, allowing the presence of multiword substitutes.

Related resources that target MWEs include collecting paraphrases, which can be used as a form of MWE substitution. Due to the particular structure of nominal compounds, their semantics can often be approximated using paraphrasing verbs and prepositions in combination with their component nouns (Lauer, 1995; Nakov, 2008; Butnariu et al., 2010; Hendrickx et al., 2013). Nakov (2008) collected annotations for about 250 noun compounds, where the annotators were asked to use the component words of a compound along with verbs and prepositions to form paraphrases (e.g., *malaria mosquito* is a *mosquito that **causes** malaria*). Girju et al. (2005) model similar information using a restricted set of categories from a semantic inventory. Arguing for a less restricted task, Hendrickx et al. (2013) requested free paraphrases from the annotators for 355 compounds in English. The paraphrasing terms could have

[1]The full resource is publicly available at `http://pageperso.lif.univ-mrs.fr/~carlos.ramisch/?page=downloads/compounds&lang=en`.

any part of speech, as long as the resulting paraphrase was a well-formed noun phrase. In all of these cases, the compounds were compositional: they could be paraphrased using combinations of their parts, and lexical substitution could be performed for each component individually.

Other MWE datasets contain compositionality assessment, and, as a consequence, they also include idiomatic cases. Datasets for nominal compound compositionality are available in English (Reddy et al., 2011; Ramisch et al., 2016; Farahmand et al., 2015), German (Roller et al., 2013), Portuguese and French (Ramisch et al., 2016), and for noun-verb expressions in Basque (Gurrutxaga and Alegria, 2013). For instance, Reddy et al. (2011) collected numerical scores for 90 English nominal compounds regarding their compositionality. Data for each compound and component word was gathered through crowdsourcing using a 6-point scale ranging from totally idiomatic (0) to fully compositional (5). Ramisch et al. (2016) extended this set with additional compounds and also applied it to other languages, generating a total of 180 compounds per language for English, Portuguese, and French. Farahmand et al. (2015) performed a similar dataset collection with 1,048 compounds annotated for compositionality and conventionality by 4 expert judges using a binary scale. However, these do not contain information about lexical substitutes. Moreover, to date, no analysis has been published on the impact of compositionality on the selection of substitutes. This paper examines this question for the substitutes proposed for compounds of varying degrees of compositionality in Brazilian Portuguese.

3 Materials and Methods

As the basis for LexSubNC, we use the set of 180 nominal compounds consisting of a noun and an adjective in Brazilian Portuguese described by Ramisch et al. (2016). They include two morphosyntactic configurations: adjective-noun, such as in *alto mar* (lit. *high sea* [international waters]), and noun-adjective, such as *vinho branco* (lit. *wine white* [white wine]). To investigate possible effects of familiarity, conventionality and compositionality in the quality of the human responses about lexical substitutes, all compounds have been annotated with corpus frequency, association strength and compositionality information.

Frequency is used as a predictor of human *familiarity* with a word, assuming that the higher the frequency the more familiar the compound. The association strength is used as an indication of the *conventionality* of a compound, with the assumption that the higher the strength, the more conventional the compound is. This follows the agreement between association measures and conventionality found by Farahmand et al. (2015). In this work we use pointwise mutual information (PMI, Church and Hanks (1990)), an association measure widely used for MWEs. 12Both frequency and PMI are calculated based on counts from a combined corpus of around 1.91 billion tokens. The corpus is formed by a concatenation of the brWaC (Wagner Filho et al., 2016; Boos et al., 2014), the Brazilian Corpus (Berber Sardinha et al., 2008), and the Portuguese Wikipedia. For *compositionality* we use the scores collected by Ramisch et al. (2016). This ensures a balance of compositionality, since the dataset was designed to contain 60 compositional (e.g., *acampamento militar* – lit. *camp military* [military camp]), 60 partly compositional (e.g., *círculo vicioso* – lit. *circle vicious* [vicious circle]), and 60 idiomatic cases (e.g., *bode expiatório* – lit. *goat expiatory* [scapegoat]).

To collect lexical substitutes for nominal compounds suggested by native speakers, we invited 86 volunteer native speakers of Brazilian Portuguese to participate in the task. All participants were undergraduate and graduate students in computer science and linguistics. Prior to starting the annotation, they were required to take a training session in which examples of compounds in sentences were presented along with the expected responses.

During the annotation, participants were asked to first read 3 sentences selected from corpora. Our hypothesis is that, by reading the sentences, annotators will think about the sense of the compound. Moreover, dispersion due to polysemy is avoided,[2] since the sentences were manually selected so that a single sense of the compound is represented.

[2]Polysemous compounds are rare but do exist, for example, *braço direito* can mean *right-hand man*, that is, a reliable assistant, or literally *right arm* as a body part.

The annotators were then asked to provide between 3 and 5 substitutes per compound, preferably single words. A minimum of three substitutes was required to allow for a greater diversity of answers per user and per compound. This requirement proved to be too strict, as many annotators complained that sometimes it is extremely difficult to find more than one substitute per compound.

The annotation interface is shown in Figure 1. Each compound was shown on a separate screen, so that, after submitting the answers, the annotator could choose to continue annotating or to stop contributing. A simplified login procedure ensured that the same annotator did not annotate the same compound twice. We estimate that each compound took 1-3 minutes to annotate, therefore this design allowed for a good flexibility, adapting to the each annotator's availability.

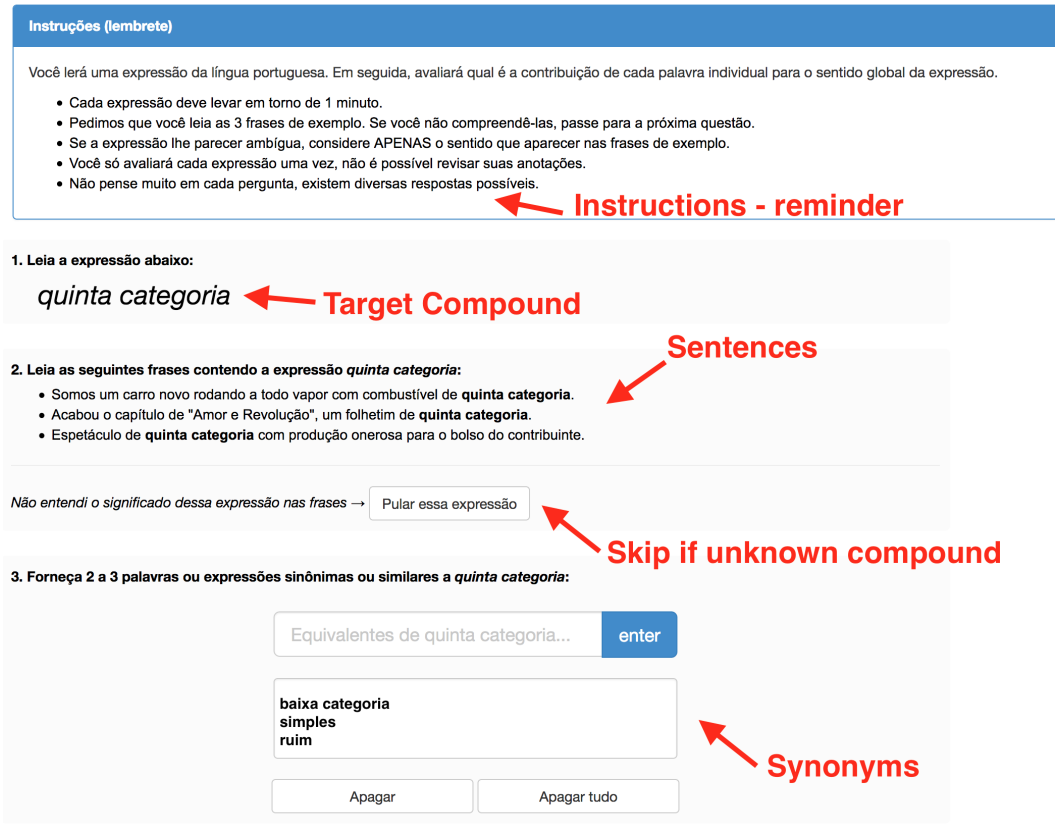

Figure 1: Annotation interface for lexical substitutes.

4 Results

A total of 5,546 responses were collected for the 180 target compounds, with 3,715 unique responses, which were manually verified by a linguist. From these, any response that could not be considered a substitute for the compound was removed: responses that expressed opinions or judgments about the compound (e.g., *país conivente com falcatruas* [country that indulges scams] for *paraíso fiscal* – lit. *paradise fiscal* [tax haven]), that used semantically related but distinct concepts (e.g., *binóculo* [binoculars] for *olho mágico* – lit. *eye magic* [peephole]) and that were tentative explanations (e.g., *recipiente de presente secreto* [recipient of secret gift] for *amigo secreto* – lit. *friend secret* [secret Santa]). One of the possible reasons for invalid cases is that users needed to enter at least three responses for a compound before being able to start the next task. The valid responses were manually classified by the expert according to the following categories:

- Synonyms: this class distinguishes between single-word synonyms (**Syn**$_{word}$, like *microchip* for

circuito integrado – lit. *circuit integrated* [integrated circuit]) and multiword synonyms (**Syn**$_{MWE}$, such as *pronto-atendimento* [urgent care] for *pronto-socorro* – lit. *ready help* [emergency services]). Moreover, where applicable, the synonyms were further identified as head (**head**, as in *vinho* [wine] for *vinho branco* [white wine]) or modifier (**mod**, as in *doce* [sweet] for *algodão-doce* – lit. *cotton sweet* [cotton candy]) of the compound.

- Near synonyms: this class identified semantically related responses, such as hypernyms, meronyms, and hyponyms, distinguishing between single word (**NearSyn**$_{word}$, like *comida* [food] for *batata-doce* – lit. *potato sweet* [sweet potato]) and multiword near synonyms (**NearSyn**$_{MWE}$, e.g., *carne de peixe* [fish meat] for *carne branca* – lit. *meat white* [white meat]).

- Paraphrases (**Paraphrases**) or definitions (**Definitions**): these two classes were used for rewrites or explanations about the target compound (e.g., *arma que não é de fogo* [weapon that is not a firearm] for *arma branca* – lit. *weapon white* [white weapon] or *passagem de ano* [passage from one year to another] for *ano-novo* – lit. *year new* [new year]).

Table 1 displays the number of total and unique responses per category, along with the number of target compounds that received responses in each category. Table 2 shows the number of cases for which the head or the modifier were proposed as substitutes for the compound.

	# Total Responses	# Unique Responses	# Target Compounds
Syn$_{word}$	966	318	99
Syn$_{MWE}$	1,257	684	159
NearSyn$_{word}$	315	150	83
NearSyn$_{word}$	303	183	96
Paraphrases	54	47	24
Definitions	166	162	90
Total	3,061	1,544	

Table 1: Substitutes classified

The average number of responses per class ranges from 1 to 4.3 types and 2 to 9.5 tokens. Some compounds had more responses than others, from 14 (for *banho turco* – lit. *bath Turkish* [Turkish bath]) to 45 (for *reta final* – lit. *straight-line final* [final stretch]). The number of annotators that agreed on a response varied from 2 to 16. Indeed, the distribution of responses confirms the suggestion of Hendrickx et al. (2013) for not using a fixed inventory of options and allowing participants to propose free paraphrasing, as most responses were unique and proposed only once for the target compounds.

	# Total Responses	# Unique Responses	# Target Compounds
Head	232	56	56
Mod	5	2	2

Table 2: Heads and modifiers as substitutes

To examine whether the familiarity, conventionality and compositionality of the compounds had any impact on the number and variety of valid lexical substitutes obtained, we measured their correlation using the Spearman coefficient, reported in Table 3. We found no effect for the total number of responses. In other words, the fact that a compound is more familiar, conventional or compositional does not necessarily correlate with the number of different substitutes it has. We found a significant mild effect for either the head or the modifier being used as response, which was positively correlated with the compositionality of the compound. In other words, the heads or modifiers were used as responses more often for the compositional cases (e.g. *água* [water] for *água mineral* – lit. *water mineral* [mineral water]).

	Frequency	PMI	Compositionality
Total number of responses	0.09	-0.03	0.08
Synonyms and near synonyms	0.05	0.10	0.16^{*}
Head or modifier	0.02	0.18^{*}	0.41^{**}

Table 3: Spearman Correlation Coefficient for responses collected. Significance levels marked as * for $p \leq 0.05$ and ** for $p \leq 0.01$.

This is not surprising given that, in compositonal compounds, the meaning of the whole can be derived from the meaning of the component words to some extent.

The resulting resource, LexSubNC, contains over 1,500 manually validated substitutes for 180 compounds classified according to the type of response and annotated for frequency, PMI and compositionality.

5 Conclusions and Future Work

In this paper we presented LexSubNC, a dataset of lexical substitutes for nominal compounds in Brazilian Portuguese. The dataset contains 180 compounds, annotated with frequency, PMI and compositionality. For each compound, the dataset contains information about different substitutes, their classification and the number of times they were proposed by the annotators. We analyzed the responses so as to determine if the frequency, conventionality or compositionality of a compound had any impact on the responses given by the human annotators. The results obtained suggest that no such effects can be found for either of these factors, apart from a mild correlation between the head or the modifier used as a response and the compositionality of a compound. The resulting dataset can be used for a variety of tasks, including as a gold standard for the evaluation of the output of lexical simplification and machine translation systems.

As future work, we plan on ranking the responses according to simplicity as substitutes for these particular compounds. This would result in a resource not only for lexical substitution, but also for lexical simplification. In addition, we plan on collecting annotations on whether these compounds are concrete or abstract, so as to verify whether there is any interaction between the analyzed variables and the concreteness of compounds. Finally, we will also collect scores for the similarity of the responses, which will result in a gold standard of lexical similarity scores.

Acknowledgments

Financial support from the projects: BEWARE 1610378 and 1510637, CNPq 312114/2015-0, 423843/2016-8, PARSEME (COST IC1207), PARSEME-FR (ANR-14-CERA-0001), ORFEO (ANR-12-CORP-0005). We would also like to thank all volunteer annotators for their valuable contribution.

References

Baldwin, T. and S. N. Kim (2010). Multiword expressions. In N. Indurkhya and F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (2 ed.)., pp. 267–292. Boca Raton, FL, USA: CRC Press, Taylor and Francis Group.

Berber Sardinha, T., J. Moreira Filho, and E. Alambert (2008). O corpus brasileiro. *Comunicaçao ao VII Encontro de Lingüistica de Corpus*.

Boos, R., K. Prestes, A. Villavicencio, and M. Padró (2014). brWaC: a WaCky corpus for Brazilian Portuguese. In *International Conference on Computational Processing of the Portuguese Language*, pp. 201–206. Springer.

Butnariu, C., S. N. Kim, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale (2010, July). Semeval-2 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, Uppsala, Sweden, pp. 39–44. Association for Computational Linguistics.

Cholakov, K., C. Biemann, J. Eckle-Kohler, and I. Gurevych (2014). Lexical substitution dataset for German. In *LREC*, pp. 1406–1411.

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

Denkowski, M. and A. Lavie (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*.

Farahmand, M., A. Smith, and J. Nivre (2015, June). A multiword expression data set: Annotating non-compositionality and conventionalization for English noun compounds. In *Proceedings of the 11th Workshop on Multiword Expressions*, Denver, Colorado, pp. 29–33. Association for Computational Linguistics.

Fellbaum, C. (Ed.) (1998, May). *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. MITPRESS. 423 p.

Girju, R., D. Moldovan, M. Tatu, and D. Antohe (2005). On the semantics of noun compounds. *Computer speech & language 19*(4), 479–496.

Gurrutxaga, A. and I. n. Alegria (2013, June). Combining different features of idiomaticity for the automatic classification of noun+verb expressions in Basque. In *Proceedings of the 9th Workshop on Multiword Expressions*, Atlanta, Georgia, USA, pp. 116–125. Association for Computational Linguistics.

Hendrickx, I., Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Szpakowicz, and T. Veale (2013, June). Semeval-2013 task 4: Free paraphrases of noun compounds. In *Proceedings of *SEM 2013, Volume 2 – SemEval*, pp. 138–143. ACL.

Hill, F., R. Reichart, and A. Korhonen (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics 41*(4), 665–695.

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

Levy, O., Y. Goldberg, and I. Dagan (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics 3*, 211–225.

Lin, D. (1998). Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pp. 768–774. Association for Computational Linguistics.

Luong, T., R. Socher, and C. Manning (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 104–113. Association for Computational Linguistics.

McCarthy, D., B. Keller, and J. Carroll (2003, July). Detecting a continuum of compositionality in phrasal verbs. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 73–80. Association for Computational Linguistics.

McCarthy, D. and R. Navigli (2007). Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 48–53. Association for Computational Linguistics.

McCarthy, D. and R. Navigli (2009). The English lexical substitution task. *Language resources and evaluation 43*(2), 139–159.

Mihalcea, R., R. Sinha, and D. McCarthy (2010). Semeval-2010 task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, Stroudsburg, PA, USA, pp. 9–14. Association for Computational Linguistics.

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

Nelson, D. L., C. L. McEvoy, and T. A. Schreiber (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers 36*(3), 402–407.

Ramisch, C., S. R. Cordeiro, L. Zilio, M. Idiart, A. Villavicencio, and R. Wilkens (2016). How naked is the naked truth? A multilingual lexicon of nominal compound compositionality. In *Proc. of ACL 2016*. ACL. To appear.

Reddy, S., D. McCarthy, and S. Manandhar (2011, November). An empirical study on compositionality in compound nouns. In *Proceedings of The 5th International Joint Conference on Natural Language Processing 2011 (IJCNLP 2011)*, Chiang Mai, Thailand.

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

Specia, L., S. K. Jauhar, and R. Mihalcea (2012). Semeval-2012 task 1: English lexical simplification. In *Proceedings of the 6th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2012, Montréal, Canada, June 7-8, 2012*, pp. 347–355.

Vulić, I., D. Gerz, D. Kiela, F. Hill, and A. Korhonen (2016). Hyperlex: A large-scale evaluation of graded lexical entailment. *arXiv*.

Wagner Filho, J., R. Wilkens, L. Zilio, M. Idiart, and A. Villavicencio (2016). Crawling by readability level. In *Proceedings of 12th International Conference on the Computational Processing of Portuguese (PROPOR)*.