Nominal Compound Compositionality: A Multilingual Lexicon and Predictive Model

Working Group 3

Silvio Ricardo Cordeiro[1,2], **Carlos Ramisch**[2], **Aline Villavicencio**[1]
[1] Aix Marseille Université, CNRS, LIF UMR 7279 (France)
[2] Institute of Informatics, Federal University of Rio Grande do Sul (Brazil)
silvioricardoc@gmail.com carlos.ramisch@lif.univ-mrs.fr avillavicencio@inf.ufrgs.br

1 Introduction

Nominal compounds (NCs) like *ivory tower* and *first lady* are recurrent MWEs in many languages. While their syntactic interpretation has been studied (Nakov, 2008), their semantics pose problems to NLP systems (Hendrickx et al., 2013). This is due to the fact that the contribution of the sense of each element for the meaning of the NC may vary considerably along a spectrum of idiomaticity (e.g. *police car* vs. *crocodile tears*). Applications dealing with phrasal semantics, like MT, must correctly deal with NCs, e.g. to avoid translating *dead end* literally into French (?*fin morte*) or Portuguese (?*fim morto*).

The meaning of NCs can be modeled using numerical scores (Reddy et al., 2011; Farahmand et al., 2015; Roller et al., 2013). Low values mean idiomatic NCs while high values represent compositional ones. For example, *olive oil* could be 90% compositional and *dead end* only 30%. Separate scores can be provided for each component; e.g. *olive oil* could be 80% related to *olives* and 100% *oil*, whereas *dead end* is 5% *dead* and 90% *end*.

We present a multilingual lexicon and experiments for predicting the numerical degree of compositionality of NCs in French (FR), English (EN) and Portuguese (PT). We believe that our lexicon and method can guide syntactic and semantic applications by reducing ambiguity of these frequent constructions.

2 Nominal Compound Lexicon

Our lexicon enriches Reddy et al. (2011) with other 90 English (EN) compounds, extending the work to encompass other compounds in French (FR) and Portuguese (PT), for a total of 180 compounds per language. The compounds were manually chosen by the authors so as to balance the expected number of compositional, partially compositional and idiomatic compounds. For EN, the final lexicon contains noun-noun compounds (e.g. *snail mail*) and adjective-noun compounds (e.g. *sacred cow*). For FR and PT, it contains noun-adjective and adjective-noun compounds. The lexicon contains the average of 10–30 numerical compositionality scores for each NC, as provided by native speakers through crowdsourcing.

Each annotation page presents a compound and 3 example sentences where the compound has the same meaning. They help preventing disagreements when an NC is polysemous (e.g. *cordon bleu* in French is a good cook and a meat dish). Users are asked to provide at least 2 synonyms.[1] Then, using a Likert scale from 0 to 5, they must judge how much of the meaning of the compound (1) comes from the head, (2) comes from the modifier, and (3) comes from both components.

We first request paraphrases in example sentences. Then, we inquire about the degree to which the meaning of a given NC arises from each of its elements. We assume that, if the interpretation 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

[1] Synonyms are not used for the moment, but help focusing the attention on the meaning of the NC. In the future, we would like to use them to evaluate automatic NC paraphrasing systems.

assumption is derived from the definition of compositionality itself as being the extent to which the meaning of the whole comes from the meanings of the parts.

We collect data for each compound through Amazon Mechanical Turk tasks. Each 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 language. At the end, they are also given example questions with annotated answers for training.

After averaging the scores over several annotations, we filter out individual annotations that are more than 2.3 deviations away from the average score and remove annotators whose Spearman correlation with the average of other annotators is below 0.6, following Roller et al. (2013). These arbitrary thresholds were tuned in order to maximise the data retention rate while reducing the average score deviation (Cordeiro et al., 2016b). The average scores of the 180 NCs plotted against the arithmetic and geometric means of head-modifier scores, for PT, are shown in Figure 1. 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 NC score seems to be a lower bound to each member word's score. A more detailed analysis of the scores is presented in Ramisch et al. (2016).

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 NC.

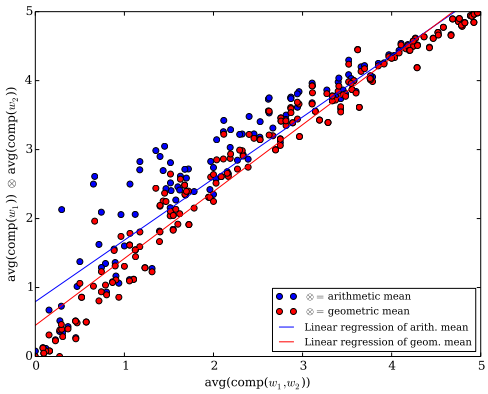

Figure 1: Average compositionality for PT NCs vs. mean scores for heads & modifiers.

	EN	FR	PT
compound $\sigma > 1.5$	22	41	30
head $\sigma > 1.5$	23	44	33
modifier $\sigma > 1.5$	35	55	34

Table 1: # NCs with high standard dev. 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 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.

3 Compositionality Prediction

We created CBOW word embeddings for each compound and its member words using word2vec (Mikolov et al., 2013) and the automatically lemmatized versions of UkWaC, FrWaC and BrWaC. The parameters are the

default ones, except for the following: no hierarchical softmax; negative sampling of 25; frequent-word downsampling weight of 10^{-6}; runs 15 training iterations. We use a minimum word count threshold of 5, and windows of +/- 8 words around the target. We test vectors with 250, 500 and 750 dimensions. More parameters were studied in an extensive evaluation (Cordeiro et al., 2016a).

To predict the compositionality of a NC w_1w_2, we use as a measure the cosine similarity between the NC vector representation $v(w_1w_2)$ and the sum of the representation vectors of the component words: $\cos(\ v(w_1w_2),\ v(w_1+w_2)\ )$ where for $v(w_1+w_2)$ we use the normalized sum $v(w_1+w_2) = \frac{v(w_1)}{||v(w_1)||} + \frac{v(w_2)}{||v(w_2)||}$. A NC is compositional if its vector is close to the sum of its components vectors (cosine is close to 1), and it is idiomatic otherwise (Cordeiro et al., 2016c).

Dimensions →	250	500	750
EN	0.67	0.71	0.72
FR	0.57	0.59	0.62
PT	0.55	0.54	0.55

Table 2: Results per language and dimension.

Table 2 shows Spearman rank-correlation of predicted NC compositionality scores with human judgments in the lexicon. The predicted scores reach considerably high results, specially for EN and FR. We believe that the lower performance for PT is due to the use of a considerably smaller corpus to build the vectors. This kind of score can be used to (a) pregroup NCs prior to parsing, (b) disambiguate attachment ambiguities during parsing and (c) identify NCs that should be treated as units, e.g. translated as unique phrases in MT.

References

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

Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline Villavicencio. 2016a. Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of ACL 2016*, Berlin, Germany.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. 2016b. Filtering and measuring the intrinsic quality of human compositionality judgments. In *The ACL 2016 Workshop on Multiword Expressions (MWE 2016)*, Berlin, Germany.

Silvio Cordeiro, Carlos Ramisch, and Aline Villavicencio. 2016c. Mwetoolkit+sem: Integrating word embeddings in the mwetoolkit for semantic MWE processing. In *Proceedings of LREC 2016*, Portoroz, Slovenia.

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.

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.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

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, Silvio Cordeiro, Leonardo Zilio, Aline Villavicencio, Marco Idiart, and Rodrigo Wilkens. 2016. How naked is the naked truth? a multilingual lexicon of nominal compound compositionality. In *Proceedings of ACL 2016: Short Papers*, Berlin, Germany.

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

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.