Show Me Your Variance and I Tell You Who You Are – Deriving Compound Compositionality from Word Alignments

Fabienne Cap
Department of Linguistics and Philology
Uppsala University
fabienne.cap@lingfil.uu.se

Abstract

We use word alignment variance as an indicator for the non-compositionality of German and English noun compounds. Our work-in-progress results are on their own not competitive with state-of-the art approaches, but they show that alignment variance is correlated with compositionality and thus worth a closer look in the future.

1 Introduction

A compound is a combination of two or more words to build a new word. Many languages (e.g. German) allow for the productive creation of new compounds from scratch. While most of such newly created compounds are compositional, i.e. the meaning of the whole can be predicted based on the meaning of its parts, there also exist lexicalised compounds which have partly or completely lost their compositional meaning (or never had one in the first place).

For many NLP applications, it is crucial to distinguish compositional from non-compositional compounds, e.g. in order to decide whether or not to split a German closed compound into its parts in order to reduce data sparsity.

This paper presents some first results on calculating compositionality scores for German and English noun compounds based on the variance of translations they exhibit when word-aligned to another language. We assume that non-compositional compounds exhibit a greater alignment variance than compositional constructions, because many non-compositional compounds...

i) are lexicalised and lexicalised counterparts are sometimes missing in the other language. The translators will instead describe the semantic content of the compound and these descriptions are very likely to differ for each occurrence. In contrast, if a compositional compound does not exist in the other language, it can most probably be created ad hoc by the translator. E.g: *Herzblut* (non-comp.: "passion/commit-ment/dedication", lit.: "heart blood") vs. *Herzbus*[1] (lit: "heart bus").

ii) may occur in contexts where they are used literally. The found translations cover occurences in both kinds of contexts and thus exhibit a larger variance than purely compositional contructions. E.g.: *Blütezeit* (non-comp.: "heyday", comp.: "blossom", lit.: "bloom time") vs. *Blütenhonig* (lit.: "blossom honey").

iii) may occur mostly (sometimes only) within larger idiomatic expressions, which in turn, similar to i), often lack an exact counterpart in the other language and are thus translated with more variance. E.g.: *auf gleicher Augenhöhe sein* (non-comp.: "to be on equal terms" lit.: "to be on the same eye level")

In our experiments, we find that translational variance in fact is a possible indicator for the compositionality of both German and English compounds and worth further improvement and investigation in the future.

2 Related Work

There has been a tremendous interest and a wide range of proposed solutions to the automatic extraction of multiword expressions (MWEs)

[1]This example has been made up from scratch. It could denote a bus providing healthcare for people suffering from heart diseases, following the pattern of "Blutbus" - a bus in which blood can be donated or alternatively a bus with a heart on it.

Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 102–107,
Valencia, Spain, April 4.

and/or the prediction of their semantic non-compositionality, which is one of their most prominent features. We restrict our review to a selection of word-alignment or translation-based approaches.

Villada Moirón and Tiedemann (2006) used word alignments to predict the idiomaticity of Dutch MWEs (preposition+NP). They calculated the variance of the alignments for each component word, and we follow their approach in the present work. Moreover, they compared the alignments of the words when occurring within an MWE vs. when occurring independently. Medeiros de Caseli et al. (2010) used alignment assymetries to identify MWEs of Brazilian Portuguese.

More recently, Salehi and Cook (2013) used string similarity to compare the translations of English MWEs with the translations of their parts. Translations were obtained lexicon-based. Salehi et al. (2014) use distributional similarity measures to identify MWE candidates in the source language. In order to determine the compositionality of the constructions they then translate the components (using a lexicon) and calculate distributional similarity for their translations. This approach was evaluated for English and German MWEs.

3 Methodology

3.1 Compound Splitting

In German, noun compounds are written as one word without spaces, e.g. *Schriftgröße* ("font size"). In order to access the word alignments of its component parts (*Schrift* ("font") and *Größe* ("size")) they have to be split prior to the word alignment process. We do so using a rule-based morphological analyser for German (Schmid et al., 2004) whose analyses are disambiguated using corpus heuristics in a two-step approach (Fritzinger and Fraser, 2010). In order to improve word alignment accurracy between German and English, we lemmatise all German nouns using the same rule-based morphological analyser.

For our experiments on English noun compounds, no preprocessing on the English data is performed.

3.2 Measuring Translational Variance

German We run word alignment on the English and the modified German parallel corpus. After the alignment, we mark the German compounds which have previously been split in the German section of the parallel corpus, e.g. *Schrift* → *Schrift_MOD*, *Größe* → *Größe_HEAD*. Then, alignments for all occurrences of e.g. *Schrift* ("font") are collected in which *Schrift* occurs in the modifier position of the word *Schriftgröße* ("font size"). The same procedure applies to all occurrences of the head *Größe* ("size"). Table 1 (a) illustrates to which words *Schrift* and *Größe* have been aligned to, we call these alignments **local alignments**.

(a) Schriftgröße (102 occurrences, TE: 1.451)

Word		Alignments
Schrift	=	font (65), text (7), fonts (3), size (3), type (2), character (2), sizes (2), font text (1), record (1) (... 16 more singletons ...)
Größe	=	size (74), sizes (13), relative size (1), (... 14 more singletons ...)

(b) Schriftzug (89 occurrences, TE: 3.827)

Word		Alignments
Schrift	=	lettering (10), logo (6), label (5), logotype (4), text (3), writing (3), texts (3), inscription (2), sticker (2), etched (2), word (1) , imprints (1), (... 47 more singletons ...)
Zug	=	lettering (10), label (5), logo (5), logotype (4), of (4), inscription (3), sticker (2), letters (2), writings (1), nameplate (1), handwriting (1), (... 51 more singletons ...)

Table 1: Local alignments for the compositional *Schrifgröße* ("font size") and the non-compositional *Schriftzug* ("lettering").

From these local alignments we then calculate the *translational entropy* (TE) scores as described in (Villada Moirón and Tiedemann, 2006). Details are given in Equation (1), where T_s is the compound with its two parts, $P(t|s)$ is the proportion of alignment t among all alignments of the word s in the context of the given compound.

$$H(T_s|s) = -\sum_{t \in T_s} P(t|s) \log P(t|s) \qquad (1)$$

High translational variance results in high TE scores. Recall from our hypothesis that the higher the translational variance, the more likely the present compound is non-compositional. We thus rank all compounds in descending order of their TE score. The example given in Table 1 illustrates the greater variance of local alignments for the non-compositional compound *Schriftzug* ("lettering") as opposed to the compositional compound *Schriftgröße*. It can be seen that there are dominant alignments for both parts of *Schriftgröße*, namely *Schrift* → font (65 times) and *Größe* →

size (74) times. In total the modifier is aligned to 25 different words and the head to 17 different words. Comparing these numbers to the non-compositional example *Schriftzug*, we find that the most frequent alignments are less dominant and there is an overall higher variance. The modifier *Schrift* (lit. "writing, font") is aligned to 59 different words, most of which occurred only once and the head *Zug* (lit. "characteristic") is aligned to 62 different words. This results in a TE score of 1.451 for *Schriftgröße* and a score of 3.827 for *Schriftzug*.

English For our experiments on English noun compounds, we apply the same procedure as described above for German. We use exactly the same word alignment file: the English section is left in its original shape, but German compounds are split and lemmatised for better word alignment quality. After alignment we mark English compounds. In the German experiment we split the compounds and thus knew where they occurred, but for English we do not have information about the presence of compounds. We thus rely on our evaluation data set consisting of English compounds and mark only those compounds in the English section of the parallel text which have occurred there. The remaining steps are the same as for German.

4 Experimental Settings

4.1 Data

Word Alignment We perform statistical word alignment using MGIZA++ (Gao and Vogel, 2008) based on parallel data provided for the annual shared tasks on machine translation[2]. The parallel corpus for German-English is mainly composed of Europarl and web-crawled texts, but also contains some translated newspaper texts. In total it consists of ca. 4.5 million sentences.

German Evaluation We evaluate our compositionality ranking of German noun-noun compounds against two available gold standard annotations, which are both part of the Ghost-NN dataset (Schulte im Walde et al., 2016b). The first one (VDHB) consists of 244 noun-noun compounds, originally annotated by von der Heide and Borgwaldt (2009) for both modifier and head compositionality on a 7-point scale (with 1 being opaque and 7 being compositional). It has been enriched by Schulte im Walde et al. (2016b) with more annotations (in part using Amazon's Mechanical Turk) in order to produce more and thus more reliable ratings. The second one (GHOST-NN) is the full Ghost-NN dataset consisting of 868 German noun-noun compounds annotated in the same manner as VDHB. Note that GHOST-NN includes VDHB.

English Evaluation For English, we base our evaluation on a dataset of 1048 English noun-noun compounds (Farahmand et al., 2015), annotated by 4 trained experts for a binary decision on compositionality. In the present study, we rely on these binary annotations and ignore the conventionalisation scores that come with the dataset.

4.2 Parameters

Frequency Ranges Due to the fact that we base our scores on statistical word alignment, we exclude all compounds that have occurred less than 5 times in the parallel corpus from our ranking. As word alignment becomes more reliable with more occurrences, we investigate 5 different frequency spans throughout all experiments with minimal occurrences of 5, 10, 25, 50 and 100 times.

Compositionality Ranges This parameter applies only to the English experiments, where 4 annotators assigned a binary compositionaly scores to the evaluation data set. We investigate two different compositionality ranges $\geq 50\%$ (at least two of the 4 annotators assigned non-compositional to the compound) and $\geq 75\%$, respectively.

Translational Entropy Scores We use up to three translational entropy scores: one based on the local alignments of the modifier (*mod.te*), one based on the alignments of the head (*head.te*) and finally, one for both (*te*), which is simply the average of the two.

4.3 Evaluation

We evaluate our rankings with respect to the German and English gold standards. Due to their different characteristics, we chose different evaluation metrics for the German and the English ranking, respectively.

German The VDHB and the GHOST data sets are both annotated with a compositionality score ranging from 1 to 7. As a consequence, the values

[2] http://www.statmt.org/wmt15

GHOST	minimal frequency				
	5	**10**	**25**	**50**	**100**
#compounds	640	504	343	209	116
mod.freq	-0.0200	-0.0453	-0.0209	-0.0572	-0.0447
mod.lmi	-0.0233	-0.414	-0.0213	-0.0462	0.0358
mod.te	0.1010	0.1355	0.1509	0.1407	0.1534
head.freq	0.0200	0.0198	-0.0697	-0.0290	-0.0227
head.lmi	-0.0094	-0.0088	-0.0565	-0.0127	0.0249
head.te	0.1602	0.1885	0.2213	0.2620	0.1845

Table 2: ρ-value results for the GHOST dataset.

of these data sets present a continuum of compositionality scores. This is in line with how our lists are ranked according to the TE scores. Following previous works (e.g. Schulte im Walde et al. (2016a)), we use the Spearman Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988) to evaluate how well our ranking is correlated with the ranking of the gold annotations.

English Due to the binary nature of the English data set we use, there are only 5 possible compositionality values (0, 0.25, 0.5, 0.75 and 1.0) and thus only 5 possible ranking positions. We thus use the uniterpolated average precision (*uap*, Manning and Schütze (1999)) to indicate the quality of the ranking.

5 Results

5.1 German

GHOST data set The results for the GHOST data set are given in Table 2. We compare the rank correlations of our rankings for modifiers (*mod.te*) and heads (*head.te*) to two simple baselines: *(mod|head).freq* = ranked in decreasing frequency of the compound and *(mod|head).lmi* = ranked in decreasing local mutual information (LMI) score (Evert, 2005). Not all compounds of the GHOST data set occurred in all frequency ranges. We thus give the number of compounds for each range in Table 2. The baselines perform poorly and rarely achieve positive ρ-values. The TE rankings improve with the frequencies of the compounds. An optimal value seems to be located between 25 and 50. For the highest frequency range of 100 we get mixed results. It can be seen that the correlations are higher overall when the lists have been ranked according to the TE score of their heads.

VDHB data set The results for the VDHB data set are given in Table 3. Again, not all compounds of the original set have occurred in all frequency ranges[3]. Only 18 of the 244 compounds occurred $\geq$100 times, which makes the results less conclusive. For this data set, we had access to the ranking of (Schulte im Walde et al., 2016a) and thus compare our results to theirs (*(mod|head).vector* in Table 3). Note that the numbers given here differ from those given in (Schulte im Walde et al., 2016a) because they are not calculated on the whole VDHB dataset but only on subsets of it. We can see from the results that the TE rankings most of the time do not even come near the performance of the vector-based ranking. It comes close only for *head.te* and a minimal frequency of 100, which apply only to 18 compounds, thus this result may not be very reliable. However, these results are nevertheless useful for further attempts of using TE scores for compositionality calculations. First, we can see that the head.te values significantly outperforms the mod.te values. This shows that the alignment variance of the compound head is more important when predicting the compounds' compositionality than the alignment variance of its modifier. Second, we see again, that the TE ranking correlation improves with increased minimal frequency constraints of the compounds to be ranked.

VDHB	minimal frequency				
	5	**10**	**25**	**50**	**100**
#compounds	143	110	76	43	18
mod.vector	0.5839	0.5478	0.5237	0.4713	0.2301
mod.te	-0.0175	-0.043	-0.0524	-0.0663	-0.0877
head.vector	0.5942	0.5871	0.5946	0.4804	0.4634
head.te	0.1268	0.1205	0.1643	0.3392	0.4407

Table 3: ρ-value results for the VDHB data set.

5.2 English

Our results for the compositionality ranking of English noun-noun compounds are given in Table 4. Note that not all of the 1042 compounds of the gold standard occurred in all frequency ranges in our corpus. We give the total number of compounds together with the number of non-compositional compounds thereof, depending on the compositionality range in the first two rows of Tables 4(a)+(b). As for the German GHOST data set above, we compare our rankings here to a simple frequency-based ranking (*freq* in Table 4) using the uninterpolated average precision (*uap*). We can see from Table 4 that all TE rank-

[3]We attribute this to the fact that half of the parallel corpus is based on the Europarl corpus, where words like *Kaffeepad* ("coffee pad") do not occur.

(a) Compositionality ≥ 0.50

	minimal frequency				
	5	**10**	**25**	**50**	**100**
#compounds	610	478	332	236	155
#opaque	138	116	84	61	35
freq	0.259	0.264	0.272	0.277	0.302
mod.te	0.295	0.308	0.299	0.296	0.258
head.te	0.279	0.291	0.293	0.297	0.262
te	0.295	0.306	0.299	0.299	0.256

(b) Compositionality ≥ 0.75

	minimal frequency				
	5	**10**	**25**	**50**	**100**
#compounds	610	478	332	236	155
#opaque	91	75	55	41	23
freq	0.176	0.180	0.188	0.194	0.218
mod.te	0.216	0.225	0.228	0.234	0.192
head.te	0.211	0.221	0.233	0.243	0.220
te	0.220	0.229	0.233	0.240	0.198

Table 4: Uap scores for the English dataset.

ings outperform the frequency-based baseline for both compositionality ranges and for minimal frequencies up to 50. In the high-frequent range, the frequency-based ranking slightly outperforms our TE ranking, but note that in this range only 35 non-compositional compounds occur in the compositionality ≥ 50 range occur (and only 23 for ≥ 75). The quality of the rankings improves with a higher minimal frequency of up to 50 and the head scores again seem to be more informative for compositionality.

6 Conclusion and Future Work

We have shown that translational entropy scores calculated from word alignments show a small correlation with compound compositionality. Our results showed that translational entropy scores are most reliable when calculated for compounds which occurred at least 25 times in the parallel corpus. Moreover, for German, we found that the alignment variance of the compound head is a better indicator for non-compositionality than variance observed for compound modifiers. For English the diffference is less clear and should be subject to further investigation in the future.

The major drawback of this approach is its dependence on parallel resources. We found that many compounds of the gold standards do not (or not sufficiently often) occur in the parallel corpus to produce reliable results. Nevertheless we are convinced that translational entropy scores can be used as an informative feature combined with previous (e.g. vector-based) approaches to compositionality identification.

For the future, we plan to compare and combine the translational entropy scores other scoring metrics based on word alignments. One example is to compare the alignments of the components when they occur in the context of the compound vs. when they occur independently similar to (Villada Moirón and Tiedemann, 2006) and (Salehi and Cook, 2013). Moreover, we will take the symmetry of word alignments into account and add a feature that indicates how many alignments were 1:1 vs. 1:n. Finally, we want to experiment with a wider range of languages on which the alignment is calculated, preferably including more contrastive languages.

Acknowledgements

This project has been funded by a VINNMER Marie Curie Incoming Grant within VINNOVAs Mobility for Growth programme. Thanks to the anonymous reviewers for their constructive comments and to Schulte im Walde et al. (2016a) for sharing their results to enable a direct comparison.

References

Stefan Evert. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. University of Stuttgart, PhD dissertation.

Meghdad Farahmand, Aaron Smith, and Joakim Nivre. 2015. A multiword expression dataset: Annotating non-compositionality and conventionalization for english noun compounds. In *MWE-NAACl'15: Proceedings of the 11th Workshop on Multiword Expressions*.

Fabienne Fritzinger and Alexander Fraser. 2010. How to avoid Burning Ducks: Combining Linguistic Analysis and Corpus Statistics for German Compound Processing. In *ACL'10: Proceedings of the 5th Workshop on Statistical Machine Translation and Metrics MATR of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 224–234.

Qin Gao and Stephan Vogel. 2008. Parallel Implementations of Word Alignment Tool. In *ACL'08: Proceedings of the Workshop on Software Engineering, Testing, and Quality Assurance for Natural Language Processing of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 49–57. Association for Computational Linguistics.

Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.

Helena Medeiros de Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language resources and evaluation*, 44(1-2):59–77.

Bahar Salehi and Paul Cook. 2013. Predicting the compositionality of multiword expressions using translations in multiple languages. In **SEM'13: Proceedings of the Second Joint Conference on Lexical and Computational Semantics*, pages 266–275.

Bahar Salehi, Paul Cook, and Timothy Baldwin. 2014. Using distributional similarity of multi-way translations to predict multiword expression compositionality. In *EACL'14: Proceedings of the Annual Meeting of the European Chapter of the Association for Computational Linguistics*, pages 472–481.

Helmut Schmid, Arne Fitschen, and Ulrich Heid. 2004. SMOR: A German Computational Morphology Covering Derivation, Composition and Inflection. In *LREC '04: Proceedings of the 4th Conference on Language Resources and Evaluation*, pages 1263–1266.

Sabine Schulte im Walde, Anna Hätty, and Stefan Bott. 2016a. The role of modifier and head properties in predicting the compositionality of english and german noun-noun compounds: A vector-space perspective. In **SEM'16: Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 148–158, Berlin, Germany. Association for Computational Linguistics.

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016b. GhoSt-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *LREC'16: Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2285–2292.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*.

Begoña Villada Moirón and Jörg Tiedemann. 2006. Identifying idiomatic expressions using automatic word-alignment. In *Proceedings of the EACL 2006 Workshop on multi-word-expressions in a multilingual context*, pages 33–40.

Claudia von der Heide and Susanne Borgwaldt. 2009. Assoziationen zu Unter-, Basis- und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.