

Approximating Compound Compositionality based on Word Alignments

Fabienne Cap



**UPPSALA
UNIVERSITY
SWEDEN**

March 9th, 2017

Motivation

Methodology

Results

Background

Several previous works have used alignments to identify MWEs:

- Medeiros de Caseli et al. (2010) used **alignment asymmetries** to identify MWEs in Brazilian Portuguese.
- Salehi and Cook (2013) compared the translations of English MWEs with the **translations of their parts**.
- Salehi et al. (2014) measured **distributional similarity** of English and German MWEs and their translations.
- Villada Moirón and Tiedemann (2006) used the **variance** of MWE alignments to identify idiomatic MWEs in Dutch.

Background

Several previous works have used alignments to identify MWEs:

- Medeiros de Caseli et al. (2010) used **alignment asymmetries** to identify MWEs in Brazilian Portuguese.
- Salehi and Cook (2013) compared the translations of English MWEs with the **translations of their parts**.
- Salehi et al. (2014) measured **distributional similarity** of English and German MWEs and their translations.
- Villada Moirón and Tiedemann (2006) used the **variance** of MWE alignments to identify idiomatic MWEs in Dutch.

Background

Several previous works have used alignments to identify MWEs:

- Medeiros de Caseli et al. (2010) used **alignment asymmetries** to identify MWEs in Brazilian Portuguese.
- Salehi and Cook (2013) compared the translations of English MWEs with the **translations of their parts**.
- Salehi et al. (2014) measured **distributional similarity** of English and German MWEs and their translations.
- Villada Moirón and Tiedemann (2006) used the **variance** of MWE alignments to identify idiomatic MWEs in Dutch.

Background

Several previous works have used alignments to identify MWEs:

- Medeiros de Caseli et al. (2010) used **alignment asymmetries** to identify MWEs in Brazilian Portuguese.
- Salehi and Cook (2013) compared the translations of English MWEs with the **translations of their parts**.
- Salehi et al. (2014) measured **distributional similarity** of English and German MWEs and their translations.
- Villada Moirón and Tiedemann (2006) used the **variance** of MWE alignments to identify idiomatic MWEs in Dutch.

Background

Several previous works have used alignments to identify MWEs:

- Medeiros de Caseli et al. (2010) used **alignment asymmetries** to identify MWEs in Brazilian Portuguese.
- Salehi and Cook (2013) compared the translations of English MWEs with the **translations of their parts**.
- Salehi et al. (2014) measured **distributional similarity** of English and German MWEs and their translations.
- Villada Moirón and Tiedemann (2006) used the **variance** of MWE alignments to identify idiomatic MWEs in Dutch.

Background

Several previous works have used alignments to identify MWEs:

- Medeiros de Caseli et al. (2010) used **alignment asymmetries** to identify MWEs in Brazilian Portuguese.
- Salehi and Cook (2013) compared the translations of English MWEs with the **translations of their parts**. **vpart, nn**
- Salehi et al. (2014) measured **distributional similarity** of English and German MWEs and their translations. **vpart, nn**
- Villada Moirón and Tiedemann (2006) used the **variance** of MWE alignments to identify idiomatic MWEs in Dutch. **v+pp**

Background

Several previous works have used alignments to identify MWEs:

- Medeiros de Caseli et al. (2010) used **alignment asymmetries** to identify MWEs in Brazilian Portuguese.
- Salehi and Cook (2013) compared the translations of English MWEs with the **translations of their parts**. **vpart, nn**
- Salehi et al. (2014) measured **distributional similarity** of English and German MWEs and their translations. **vpart, nn**
- **Villada Moirón and Tiedemann (2006)** used the **variance** of MWE alignments to identify idiomatic MWEs in Dutch. **v+pp**

We used **this** approach and applied it to determine the compositionality of **German noun-noun compounds**.

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
 - A lexical counterpart might be missing in the other language
 - Translators have to work around it
 - → highly likely that these "work-arounds" will differ
 - **Example:** *Herzblut*: passion, commitment, dedication
-
- Ad-hoc created compounds might also lack a counterpart
 - But: due to their compositional meaning, the translator is likely to create the same compound in the other language
 - **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- **Example:** *Herzblut*: passion, commitment, dedication

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- Example: *Herzblut*: passion, commitment, dedication

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- Example: *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- **Example:** *Herzblut*: passion, commitment, dedication

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- Example: *Herzblut*: passion, commitment, dedication
- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- Example: *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- **Example:** *Herzblut*: passion, commitment, dedication

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- **Example:** *Herzblut*: passion, commitment, dedication

How about compositional constructions?

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- **Example:** *Herzblut*: passion, commitment, dedication

How about compositional constructions?

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- **Example:** *Herzblut*: passion, commitment, dedication

How about compositional constructions?

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 1: missing counterparts

- The meaning of non-compositional compounds is lexicalised
- A lexical counterpart might be missing in the other language
- Translators have to work around it
- → highly likely that these "work-arounds" will differ
- **Example:** *Herzblut*: passion, commitment, dedication

How about compositional constructions?

- Ad-hoc created compounds might also lack a counterpart
- But: due to their compositional meaning, the translator is likely to create the same compound in the other language
- **Example:** *Blutbus*: blood bus

Why alignment variance?

Motivation 2: contexts

- Some compounds have both, a compositional and a non-compositional meaning, depending on their context.
- **compositional:** *Die Blütezeit der Kirschbäume.*
“The flowering period of the cherry trees.”
- **non-compositional:** *Die Blütezeit der Dampfmaschine.*
“The heyday of the steam machine.”
- Translations thus differ considerably, which adds variance

Why alignment variance?

Motivation 2: contexts

- Some compounds have both, a compositional and a non-compositional meaning, depending on their context.
- **compositional:** *Die Blütezeit der Kirschbäume.*
“The flowering period of the cherry trees.”
- **non-compositional:** *Die Blütezeit der Dampfmaschine.*
“The heyday of the steam machine.”
- Translations thus differ considerably, which adds variance

Why alignment variance?

Motivation 2: contexts

- Some compounds have both, a compositional and a non-compositional meaning, depending on their context.
- **compositional:** *Die Blütezeit der Kirschbäume.*
“The flowering period of the cherry trees.”
- **non-compositional:** *Die Blütezeit der Dampfmaschine.*
“The heyday of the steam machine.”
- Translations thus differ considerably, which adds variance

Why alignment variance?

Motivation 2: contexts

- Some compounds have both, a compositional and a non-compositional meaning, depending on their context.
- **compositional:** *Die Blütezeit der Kirschbäume.*
“The flowering period of the cherry trees.”
- **non-compositional:** *Die Blütezeit der Dampfmaschine.*
“The heyday of the steam machine.”
- Translations thus differ considerably, which adds variance

Why alignment variance?

Motivation 2: contexts

- Some compounds have both, a compositional and a non-compositional meaning, depending on their context.
- **compositional:** *Die Blütezeit der Kirschbäume.*
“The flowering period of the cherry trees.”
- **non-compositional:** *Die Blütezeit der Dampfmaschine.*
“The heyday of the steam machine.”
- Translations thus differ considerably, which adds variance

Why alignment variance?

Motivation 2: contexts

- Some compounds have both, a compositional and a non-compositional meaning, depending on their context.
- **compositional:** *Die Blütezeit der Kirschbäume.*
“The flowering period of the cherry trees.”
- **non-compositional:** *Die Blütezeit der Dampfmaschine.*
“The heyday of the steam machine.”
- Translations thus differ considerably, which adds variance

How about compositional constructions?

→ less variation of contexts

Why alignment variance?

Motivation 3: part of larger idioms

- Some non-compositional compounds occur only/mostly within larger idioms
- Translation variations shown by previous works on MWEs
- **Example 1:** *von der Bildfläche verschwinden*
“disappear”, “vanishing into thin air”
- **Example 2:** *Dreh- und Angelpunkt sein*
“the crux of the matter”, “the key element”,

Why alignment variance?

Motivation 3: part of larger idioms

- Some non-compositional compounds occur only/mostly within larger idioms
- Translation variations shown by previous works on MWEs
- **Example 1:** *von der Bildfläche verschwinden*
“disappear”, “vanishing into thin air”
- **Example 2:** *Dreh- und Angelpunkt sein*
“the crux of the matter”, “the key element”,

Why alignment variance?

Motivation 3: part of larger idioms

- Some non-compositional compounds occur only/mostly within larger idioms
- Translation variations shown by previous works on MWEs
- **Example 1:** *von der Bildfläche verschwinden*
“disappear”, “vanishing into thin air”
- **Example 2:** *Dreh- und Angelpunkt sein*
“the crux of the matter”, “the key element”,

Why alignment variance?

Motivation 3: part of larger idioms

- Some non-compositional compounds occur only/mostly within larger idioms
- Translation variations shown by previous works on MWEs
- **Example 1:** *von der Bildfläche verschwinden*
“disappear”, “vanishing into thin air”
- **Example 2:** *Dreh- und Angelpunkt sein*
“the crux of the matter”, “the key element”,

Why alignment variance?

Motivation 3: part of larger idioms

- Some non-compositional compounds occur only/mostly within larger idioms
- Translation variations shown by previous works on MWEs
- **Example 1:** *von der Bildfläche verschwinden*
“disappear”, “vanishing into thin air”
- **Example 2:** *Dreh- und Angelpunkt sein*
“the crux of the matter”, “the key element”,

Motivation

Methodology

Results

Motivation

Methodology

Results

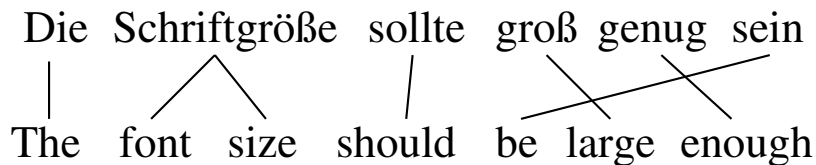
Die Schriftgröße sollte groß genug sein

The font size should be large enough

Die Schriftgröße sollte groß genug sein

The font size should be large enough

- **Parallel Corpus**



- **Parallel Corpus**
- **Statistical Word Alignment**

Die Schriftgröße sollte groß genug sein

The font size should be large enough

- **Parallel Corpus**
- **Statistical Word Alignment**

Die **Schriftgröße** sollte groß genug sein

The font size should be large enough

- **Parallel Corpus**
- **Statistical Word Alignment**
- **Compound Splitting**

Die Schriftgröße sollte groß genug sein
The font size should be large enough

- **Parallel Corpus**
- **Statistical Word Alignment**
- **Compound Splitting**

Word Alignment

(a) compositional: Schriftgröße (102 occurrences)

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 ...)

Word Alignment

(c) compositional: Schriftgröße (102 occurrences)

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 ...)

(d) non-compositional: Schriftzug (89 occurrences)

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 ...)

Word Alignment

(e) compositional: Schriftgröße (102 occurrences)

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 ...)

(f) non-compositional: Schriftzug (89 occurrences)

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 ...)

(g) compositional: Schriftgröße

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 ...)

(h) non-compositional: Schriftzug

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 ...)

(i) compositional: Schriftgröße

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 ...)

(j) non-compositional: Schriftzug

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 ...)

Calculation of **translational entropy**:

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

(k) compositional: Schriftgröße

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 ...)

(l) non-compositional: Schriftzug

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 ...)

Schriftgröße: **1.451**

Schriftzug: **3.827**

Calculation of **translational entropy**:

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

Ranking

Compound	Freq.	TE
Seilbahn	561	3.809
Sonnenschirm	594	3.315
Seemann	76	3.114
Armband	178	3.058
Stereoanlage	119	2.899
Wasserhahn	50	2.778
Kaffeemaschine	333	2.725
Hausboot	34	2.718
Bettwäsche	842	2.670
Telefonzelle	26	2.602
Gewächshaus	165	2.584
Schlauchboot	56	2.524
Mülleimer	61	2.500
Kopfkissen	83	2.481
Handtuch	911	2.463
Mülltonne	34	2.459
Schachbrett	66	2.408
Tintenfisch	75	2.394
Sessellift	134	2.368

Motivation

Methodology

Results

Overview

Motivation

Methodology

Results

Results

- Calculate Spearman rank correlation coefficient (ρ -value) on the von der Heide/Borgwaldt dataset
- Compare to vector-based approach by Schulte im Walde (2016)

Results

- Calculate Spearman rank correlation coefficient (ρ -value) on the von der Heide/Borgwaldt dataset
- Compare to vector-based approach by Schulte im Walde (2016)

Results

- Calculate Spearman rank correlation coefficient (ρ -value) on the von der Heide/Borgwaldt dataset
- Compare to vector-based approach by Schulte im Walde (2016)

Results

- Calculate Spearman rank correlation coefficient (ρ -value) on the **von der Heide/Borgwaldt** dataset
- Compare to **vector-based** approach by Schulte im Walde (2016)

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

Results

- Calculate Spearman rank correlation coefficient (ρ -value) on the **von der Heide/Borgwaldt** dataset
- Compare to **vector-based** approach by Schulte im Walde (2016)

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

A closer look at the results

– Rankings for the 18 highest frequent compounds

VDHB mod ranking	mod.te ranking
Handtuch	Sonnenschirm
Visitenkarte	Seilbahn
Nachttisch	Armband
Haselnuss	Gewächshaus
Sonnenblume	Visitenkarte
Stereoanlage	Bettwäsche
Sessellift	Papierkorb
Kreditkarte	Nachttisch
Armband	Sessellift
Seilbahn	Stereoanlage
Papierkorb	Handtuch
Postkarte	Postkarte
Eisberg	Kaffeemaschine
Sonnenschirm	Wasserfall
Bettwäsche	Haselnuss
Gewächshaus	Sonnenblume
Kaffeemaschine	Eisberg
Wasserfall	Kreditkarte

VDHB head ranking	head.te ranking
Bettwäsche	Seilbahn
Stereoanlage	Sonnenschirm
Seilbahn	Armband
Wasserfall	Stereoanlage
Eisberg	Kaffeemaschine
Armband	Bettwäsche
Papierkorb	Gewächshaus
Kreditkarte	Handtuch
Gewächshaus	Sessellift
Kaffeemaschine	Wasserfall
Nachttisch	Papierkorb
Sessellift	Nachttisch
Postkarte	Postkarte
Sonnenschirm	Visitenkarte
Handtuch	Haselnuss
Visitenkarte	Kreditkarte
Haselnuss	Eisberg
Sonnenblume	Sonnenblume

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
 - Head variance better indicator than modifier variance (!)
 - **Problem:** data sparsity
-
- Weighting of TE scores
 - Combine alignments into different languages
 - Combine with other existing alignment-based scores
 - Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
- Head variance better indicator than modifier variance (!)
- **Problem:** data sparsity

- Weighting of TE scores
- Combine alignments into different languages
- Combine with other existing alignment-based scores
- Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
- Head variance better indicator than modifier variance (!)
- **Problem:** data sparsity

- Weighting of TE scores
- Combine alignments into different languages
- Combine with other existing alignment-based scores
- Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
 - Head variance better indicator than modifier variance (!)
 - **Problem:** data sparsity
-
- Weighting of TE scores
 - Combine alignments into different languages
 - Combine with other existing alignment-based scores
 - Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
- Head variance better indicator than modifier variance (!)
- **Problem:** data sparsity

Future Work:

- Weighting of TE scores
- Combine alignments into different languages
- Combine with other existing alignment-based scores
- Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
- Head variance better indicator than modifier variance (!)
- **Problem:** data sparsity

Future Work:

- Weighting of TE scores
- Combine alignments into different languages
- Combine with other existing alignment-based scores
- Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
- Head variance better indicator than modifier variance (!)
- **Problem:** data sparsity

Future Work:

- Weighting of TE scores
- Combine alignments into different languages
- Combine with other existing alignment-based scores
- Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
- Head variance better indicator than modifier variance (!)
- **Problem:** data sparsity

Future Work:

- Weighting of TE scores
- Combine alignments into different languages
- Combine with other existing alignment-based scores
- Combine with monolingual approaches

Conclusion and Future Work

Conclusions:

- Alignment variance is weakly correlated with compositionality
- Head variance better indicator than modifier variance (!)
- **Problem:** data sparsity

Future Work:

- Weighting of TE scores
- Combine alignments into different languages
- Combine with other existing alignment-based scores
- Combine with monolingual approaches