

To Leave no Stone Unturned

Annotating Verbal Idioms in the PMB

Rafael Ehren, Kilian Evang and Laura Kallmeyer

Heinrich Heine University Düsseldorf

2024-05-25

MWE-UD @ LREC-COLING 2024

Outline

Introduction

Annotation

Results and Discussion

Conclusions and Future Work

Outline

Introduction

Annotation

Results and Discussion

Conclusions and Future Work

In this Talk

- correcting the annotation of verbal idioms in a sembank
- evaluation in terms of inter-annotator agreement
- challenges and future work

The Parallel Meaning Bank

- Abzianidze et al. (2017, 2020)
- a partially parallel corpus of English, German, Italian, and Dutch text
- annotated with meaning representations following Discourse Representation Theory
- including word senses, semantic roles, discourse connectives and scope, coreference
- used to train and evaluate semantic parsers

The Parallel Meaning Bank, cont.

- initial annotations created through an NLP pipeline, hand-corrected
- document statuses
 - gold = manually checked
 - silver = partially manually checked
 - bronze = not manually checked
- pipeline has strong compositionality assumptions built in (e.g., 1 content word = 1 concept)
- until now, not much focus on idioms in annotation

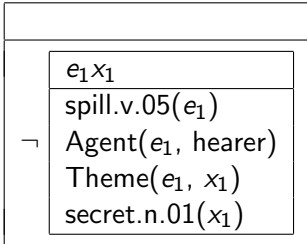
Example 1: No Idiom

- (1) “A jet is flying” (doc. 84/0011, gold)

$x_1 e_1$
jet.n.01(x_1)
fly.v.01(e_1)
Theme(e_1, x_1)

Example 2: Decomposable Idiom

(2) “Don’t spill the beans” (doc. 11/0958, gold)



Example 3: Non-decomposable Idiom

(3) “Are you pulling my leg?” (doc. 01/1871, silver)

$e_1 x_1$
pull.v.01(e_1)
Agent(e_1 , hearer)
Theme(e_1 , x_1)
leg.n.01(x_1)
PartOf(x_1 , speaker)

Example 3: Non-decomposable Idiom

- (4) “Are you pulling my leg?” (doc. 01/1871, our proposed annotation)

x_1, e_1
pull_the_leg_of.v.01(e_1)
Agent(e_1 , hearer)
Theme(e_1 , speaker)

Outline

Introduction

Annotation

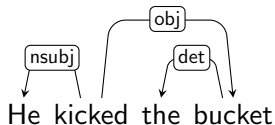
Results and Discussion

Conclusions and Future Work

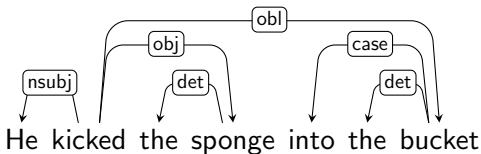
Finding Potentially Idiomatic Expressions (PIEs) in the PMB

- compiled list of 39,521 German verbal idioms from <https://www.redensarten-index.de>
- found 6,187 PIE instances in PMB by dependency pattern matching (Haagsma, 2020)

Finding Potentially Idiomatic Expressions (PIEs) in the PMB, cont.



(a) A PIE instance



(b) Not a PIE instance

Figure: Parsing-based extraction.

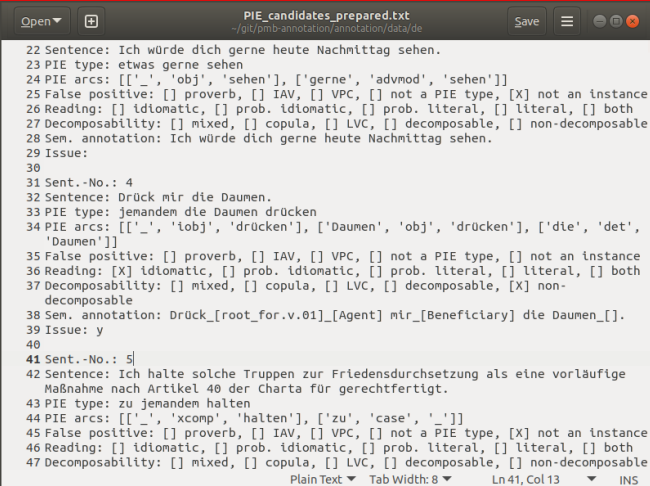
Annotation

- Annotation conducted by three linguistically trained native speakers, every sentence annotated twice
- several cycles of trial annotation, revising guidelines, correction

Annotation

- Annotation objectives
 1. Filter out false positives
 2. Annotate the degree of idiomaticity
 3. Judging the degree of decomposability
 4. Sense and role annotation

Annotation



```
PIE_candidates_prepared.txt
~/git/pmb-annotation/annotation/data/de

22 Sentence: Ich würde dich gerne heute Nachmittag sehen.
23 PIE type: etwas gerne sehen
24 PIE arcs: [['_', 'obj', 'sehen'], ['gerne', 'advmod', 'sehen']]
25 False positive: [] proverb, [] IAV, [] VPC, [] not a PIE type, [X] not an instance
26 Reading: [] idiomatic, [] prob. idiomatic, [] prob. literal, [] literal, [] both
27 Decomposability: [] mixed, [] copula, [] LVC, [] decomposable, [] non-decomposable
28 Sem. annotation: Ich würde dich gerne heute Nachmittag sehen.
29 Issue:
30
31 Sent.-No.: 4
32 Sentence: Drück mir die Daumen.
33 PIE type: jemandem die Daumen drücken
34 PIE arcs: [['_', 'iobj', 'drücken'], ['Daumen', 'obj', 'drücken'], ['die', 'det',
'Daumen']]
35 False positive: [] proverb, [] IAV, [] VPC, [] not a PIE type, [] not an instance
36 Reading: [X] idiomatic, [] prob. idiomatic, [] prob. literal, [] literal, [] both
37 Decomposability: [] mixed, [] copula, [] LVC, [] decomposable, [X] non-
decomposable
38 Sem. annotation: Drück_[root_for.v.01]_[Agent] mir_[Beneficiary] die Daumen_[].
39 Issue: y
40
41 Sent.-No.: 5]
42 Sentence: Ich halte solche Truppen zur Friedensdurchsetzung als eine vorläufige
Maßnahme nach Artikel 40 der Charta für gerechtfertigt.
43 PIE type: zu jemandem halten
44 PIE arcs: [['_', 'xcomp', 'halten'], ['zu', 'case', '_']]
45 False positive: [] proverb, [] IAV, [] VPC, [] not a PIE type, [X] not an instance
46 Reading: [] idiomatic, [] prob. idiomatic, [] prob. literal, [] literal, [] both
47 Decomposability: [] mixed, [] copula, [] LVC, [] decomposable, [] non-decomposable

Plain Text ▾ Tab Width: 8 ▾ Ln 41, Col 13 ▾ INS
```

Figure: Text-based annotation interface

Annotation

Decomposable idiom:

- (5) Er_[Experiencer] **schwimmt**_[buck.v.02] **gegen**_[] **den**_[]
He swims against the
Strom_[Stimulus]_[trend.n.01].
tide.
'He bucks the trend.'

Annotation

Decomposable idiom:

- (6) [...] **den**_[] **inneren**_[]
[...] the inner
Schweinehund_[Co-Theme]_[weakness.n.01] zu
pig-hound to
besiegen_[overcome.v.02]_[Theme].
defeat.
'[...] to overcome a weakness.'

Annotation

Non-decomposable idiom:

- (7) **Stecke**_[despair.v.01]_[Experiencer] nicht **den Kopf**_[] **in**
Bury not the head in
den Sand_[]!
the sand!
'Don't despair!'

Annotation

Typical LVC:

- (8) Die Generation_[Theme] der Zeitzeugen
The generation of contemporary witnesses
geht_[end.v.01] **zu**_[] **Ende**_[] [...]
goes to end [...]
'The Generation of contemporary witnesses is ending.'

Outline

Introduction

Annotation

Results and Discussion

Conclusions and Future Work

Computing Inter-annotator Agreement

total annotated instances	6,187
discussed in meetings	341
≠ 2 annotations	18
multiple idiom instances	7
remaining	5,821

Unanimously Classified PIEs

not an idiom	3,448
not an instance	1,968
IAV	194
VPC	149
proverb	142
literal	121
not a verbal PIE type	90

idiom	1,945
non-decomposable	1,335
decomposable	186
LVC	24
copula	19
mixed	2

PIE Classification: Most Frequent Disagreements

IAV, not an instance	349
literal, not an instance	195
decomposable, non-decomposable	181
non-decomposable, not an instance	136
LVC, non-decomposable	108
not a verbal PIE type, not an instance	91
IAV, non-decomposable	73
non-decomposable, not a verbal PIE type	43
IAV, literal	41
literal, non-decomposable	39

Agreement on PIE Classification

	Cohen's κ
coarse-grained	.8433
fine-grained	.6311

Semantic Annotation Agreement

Head selection	.9769
Head sense classification	.5862

Internal argument identification	.9914
Internal argument role classification	.7296
Internal argument sense classification	.6824

External argument identification	.9845
External argument role classification	.8352

Decomposability

- (9) Tom_[Agent] **legte**_[reveal.v.02] **die**_
Tom laid the
Karten_[Topic]_[intention.n.01] **auf**_
cards on the table.
'Tom revealed his intentions'.

Decomposability

- (10) Der Gouverneur_[Agent] **setzte**_[set.v.05] die
The governor set the
Häftlinge_[Patient] **auf freien Fuß**_[Result]_[free.a.01].
prisoners on free foot.
'The governor set the prisoners free'.

Missing Senses

- (11) Dichter_[AttributeOf] wie Milton **sind**_[rare.a.03] **dünn**
Poets like Milton are thinly
gesät_[].
sowed.
'Poets like Milton are few and far between.'

Missing Senses

- (12) Tom **hat nichts zu verlieren**.
Tom has nothing to lose.
'Tom has nothing to lose.'

Missing Senses

- (13) er_[Agent] **gab**_[give.v.20] ihm_[Patient] einen tüchtigen
he gave him a hearty
Fußtritt_[Theme] **mit**_[] **auf**_[] **den**_[] **Weg**_[]
kick with on the way
'he gave him a good kick (as he was leaving)'

Missing Senses

- (14) Tom_[AttributeOf] **schwimmt**_[rich.a.01] **im** **Geld**_[].
Tom swims in the money.
'Tom is rolling in money.'

Missing Senses

- (15) Mir-[Experiencer] **fällt**-[cabin_fever.n.00] **die**-[] **Decke**-[]
Me falls the ceiling
auf-[] **den**-[] **Kopf**-[].
on the head.
'I'm starting to get cabin fever'.

Collocations

- Idioms proper = idioms of decoding: hearer must know idioms to understand it. E.g., *pull s.o.'s leg*
- Collocations = idioms of encoding: speaker must know idiom to speak idiomatically, but they can nevertheless be understood compositionally with the correct meaning, e.g., *brush teeth* vs. *make teeth clean*
- Fillmore et al. (1988)
- Mere collocation out of scope for us, but difference sometimes hard to tell

Collocations

- (16) Endlich **zeigte** er **sein wahres Gesicht**.
Finally shows he his true face.
'Finally he reveals his real personality.'

Collocations

- (17) Wir sollten das wohl **unter vier Augen**
We should that probably among four eyes
besprechen.
talk about.
'We should probably discuss this in private.'

Outline

Introduction

Annotation

Results and Discussion

Conclusions and Future Work

Summary

- idioms present many challenges to semantic annotation
- result: underrepresented or inadequately annotated in sembanks
- this work: targeted annotation of German PIEs in PMB for idiomatic status and meaning
- IAA encouraging considering task complexity
- biggest challenges: word senses, subtle difficulties classifying PIEs

Next Steps (Ongoing)

- integrate annotations into PMB
- how to resolve disagreements?
- how do we get documents with idioms to gold status?
- testing semantic parsers on idiom-aware annotation

Thank you!

Bibliography I

- Abzianidze, L., Bjerva, J., Evang, K., Haagsma, H., van Noord, R., Ludmann, P., Nguyen, D.-D., and Bos, J. (2017). The Parallel Meaning Bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In Lapata, M., Blunsom, P., and Koller, A., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 242–247, Valencia, Spain. Association for Computational Linguistics.
- Abzianidze, L., van Noord, R., Wang, C., and Bos, J. (2020). The parallel meaning bank: A framework for semantically annotating multiple languages. *Applied mathematics and informatics*, 25(2):45–60.
- Fillmore, C. J., Kay, P., and O'Connor, M. C. (1988). Regularity and idiomaticity in grammatical constructions: the case of *let alone*. *Language*, 64:501–538.
- Haagsma, H. (2020). *A Bigger Fish to Fry: Scaling up the Automatic Understanding of Idiomatic Expressions*. PhD thesis, Rijksuniversiteit Groningen.