



NO MUMIN TROLLS IN THE VALLEY: IDIOMS IN FINNISH AND OTHER LANGUAGES

Sami Itkonen
2022-03-31



AGENDA

- Intro
- Idioms in general
- SemEval 2022
- Finnish idioms
 - Idiomatic minimal pairs



CAREER

- MSc, Engineering Physics, HUT (now Aalto)
- MA, Language Technology, UH
- Doctoral Researcher, UH, 2021-
 - Behind the Words paraphrase project
- Working life before that
 - Project Document Management systems (mainly coding)



IN THE BEGINNING: MUMIN TROLLS

- Started with the Finnish idioms “Ei ole kaikki muumit laaksossa”
 - “Not all of the mumins are in the valley”
 - Kortelainen: construction “ei ole kaikki x:t y:ssä”
 - origin: “ei ole kaikki kotona”
 - i.e. “not everyone/thing is at home”
 - Semantic metaphor “mind is a container”
 - i.e. something is not quite right in the head, something is missing
 - In English:
 - “a few cans short of a six pack”
 - “not playing with the full deck”
- Later: “idiomatic minimal pairs”
 - i.e. “vetää lärviin” vs “vetää lärvit”
 - Pair of (verbal) idioms with the only difference in one surface feature (noun case)
 - In other languages, similar distinctions may be created with e.g. prepositions



MOTIVATIONS

- Work with Finnish language
- Improving the state of the language technology for Finnish
- Cognitively motivated (?)
- “Language is what makes us human”



IDIOMS: BACKGROUND

- Idioms and MWE
 - Properties
 - Variation
- Cognitive background



IDIOMS: DEFINITIONS

- Multiword expressions (MWE)
- Idiom
- Light Verb Construction (LVC)
 - The verb has little meaning on its own (“semantically bleached”)
- Many different definitions from various sources
 - Lots of overlaps between different definitions
- In this talk, “idiomatic” used in the sense “non-literal”



IDIOMS: PROPERTIES

- Idioms are "semantic outliers"
- Morphological / syntactic / lexical fixedness
 - Lack of productivity
- Non-compositionality and semantic opaqueness
- Informality, Figuration, Proverbiality, Metaphors, ...
- Conventionality
 - A linguistic regularity to which "a population has implicitly agreed to conform"
 - "identified as idiomatic within a speech community"; "people know them"
 - Idealized L1 speaker considers an expression idiomatic (Chomskyist)
- Institutionalization
 - "In the dictionary"



IDIOMS: MEASUREMENTS AND VARIATION

More exotic examples

Heittää helmiä sioille ‘Throw pearls to pigs’

- Morphological
- Syntactic
 - Word order
 - Discontinuities / gaps / sparseness
- Lexical
 - Substitution test (with synonyms)
- Non-literal translatability
- Combinations of the above
- Semantic
 - Compositionality
 - Transparency

- Syöttää helmiä sioille ‘feed pearls to pigs’
- Heittää helmiä porsaille ‘throws pearls to piglets’
- Heittää hopealusikoita sioille ‘throw silver spoons to pigs’
- Sellaista helmeä emme sialle heitä ‘not going to throw such a pearl to a pig’
- Tarjota sikoja Helmille ‘offer pigs to Helmi’



IDIOMS: COGNITIVE BACKGROUND

- Idioms are processed faster in the brain
 - Also: “formulaic language”
 - between one third and one half of expressions stored in the long term memory are “formulaic expressions”
 - example: sports commentators
- There is a limit to how much variation there can be (in order to be processed faster)
- Idiom != formulaic expression
 - Only the “canonical” form(s) are (strictly speaking) formulaic



IDIOMS: PREVALENCE

- Fazly et al (2009) analysed 60 idioms
 - Nearly half could be interpreted literally
- OTOH, Savary et al (2019) found that literal occurrences are quite rare
- Taslimipoor et al (2017)
 - Idiom frequency correlates with “fluidity” (MWE can occur idiomatically or literally)
 - Most frequent idioms are generally either/or
- Meanings
 - Multiple non-literal meanings, yes
 - Multiple literal meanings?
 - Not for LVCs, but noun compounds..??



SEMEVAL 2022

- Task 2: Multilingual Idiomaticity Detection
- Three languages: English, Portuguese, Galician (blind test only)
- Data
 - N+N / A+N compounds (MWE) in a sentence + surrounding sentences
 - Example:
 - He was not a *blue blood* jurist issuing judicial decisions that nobody understood affecting people and corporations that nobody knew.
- Subtask A
 - Determine whether the MWE is used literally or not
 - Zero shot: MWEs in the dev / test / eval set not found in training data
 - One-shot: One literal and one idiomatic example of MWE sentence (if available)



METHODS

- Feature model
 - HuggingFace pipelines
 - Lexical substitution with Mask filling
 - MWE replaced with mask token
 - Does predicted token match with MWE component? => Boolean feature “Hassub”
 - Backtranslation
 - Is the MWE found from the backtranslation? => Boolean feature “Trans”
 - Sentiment analysis
 - Idiomatic phrases more likely to be affective (positive/negative)
 - Semantic outlier detection
 - Sentence embedding difference between target sentence and previous/next sentence
 - Boolean surface features: Caps, Quotes
 - Is the MWE capitalised? *Blue Blood* or in quotes? “*blue blood*”
- Feature model combined with BERT model



RESULTS

- Placement: 15th out of 20
- Beat the BERT baseline

Configuration	F1
Hassub	0.551
Trans	0.597
Sentiment	0.542
Majority class	0.545
Sentence transformers	0.614
+ feature	0.713
BERT baseline	0.694
+ multilingual 1: PT from full model	0.721
+ multilingual 2: PT from separate model	0.725
+ feature	0.715
+ multilingual 1 + feature	0.733
+ multilingual 1 + feature, agree	0.733
+ multilingual 1 + feature + trans	0.751
+ multilingual 1 + feature + trans, agree	0.751
+ multilingual 2 + feature + trans, agree	0.743
Majority voting + trans	0.724
Majority voting + trans, agree	0.724

Table 2: Results for the development set.

Configuration	F1
Sentence transformers	0.558
+ feature	0.646
BERT baseline	0.702
+ multilingual 1	0.723
+ feature	0.714
+ feature + trans	0.671
+ feature, agree	0.723
+ multilingual 1 + feature	0.720
+ multilingual 1 + feature, agree	0.725

Table 3: Results for the evaluation set.

Language	F1
English	0.752
Portuguese	0.694
Galician	0.499
Total	0.6625

Table 4: Official results for the test set.



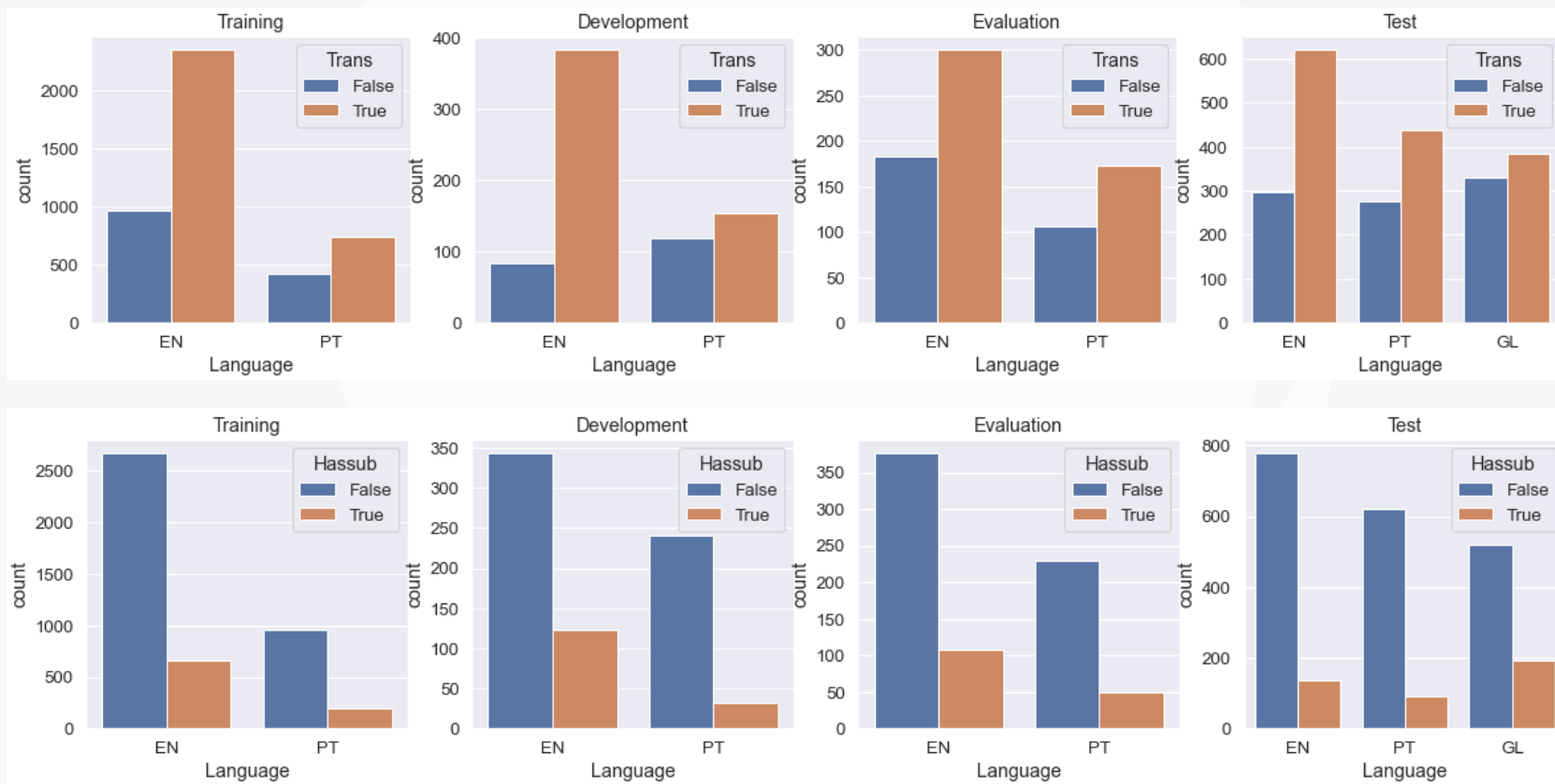
BOOLEAN FEATURES: TRAINING SET



- Literal: 1
- Idiomatic: 0

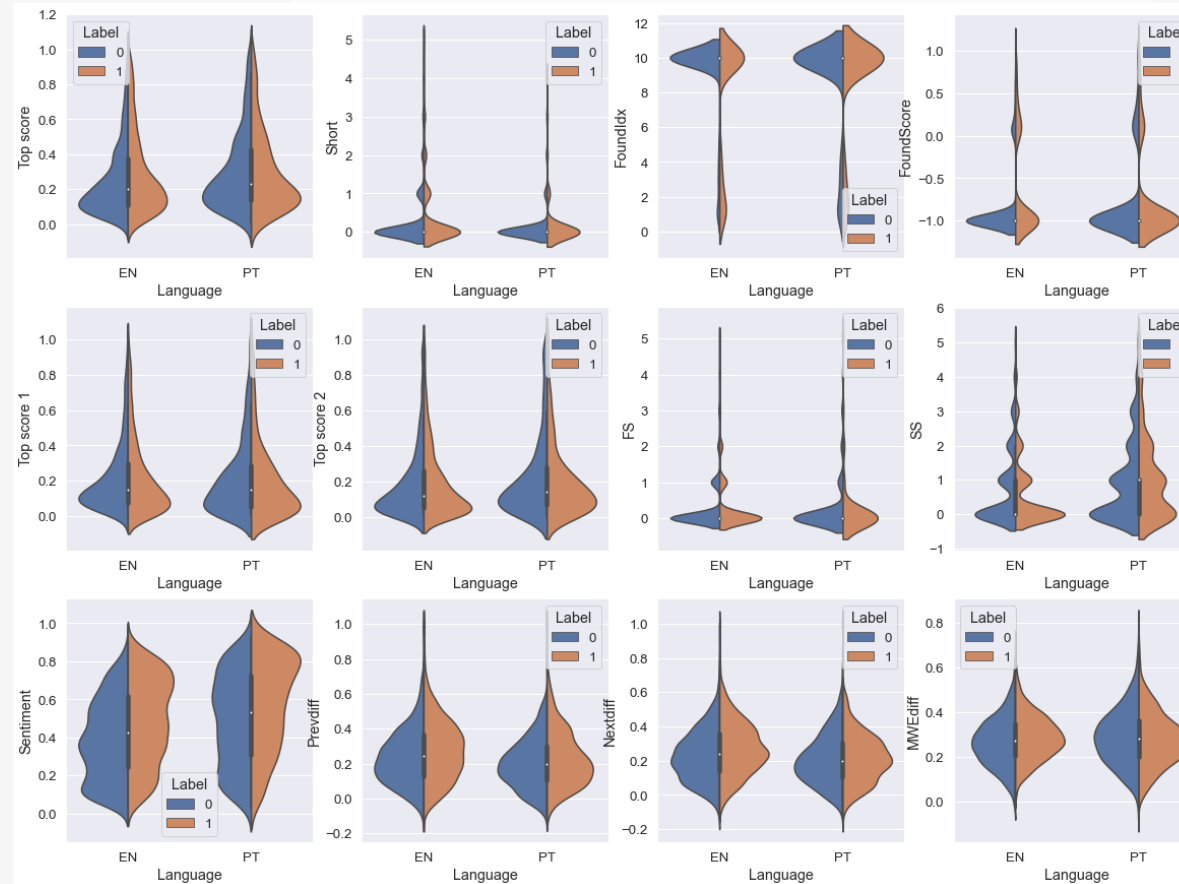


COUNTS PER FEATURE, SET AND LANGUAGE





FEATURES AND LANGUAGE



SUBSTITUTIONS

Target	Label	Top terms	Top score	Hassub	Short	Found Idx	Found Score
There are several theories behind the origin of the term “ <i>Double Dutch</i> .”	0	., -, ..., s, <i>man</i>	0.008	False	3	10	-1.000
<i>Double Dutch</i> also derives from the same era, Dutch seeming a strange and convoluted language hence Double Dutch meaning indiscernible, mad and generally all round not on foreign speak.	0	It, <i>English</i> , This, Dutch , German	0.385	True	1	4	0.062
No vídeo divulgado nas redes sociais, é possível perceber que um som faz o casal olhar para o prédio da frente e ver o <i>efeito especial</i> da fumaça.	0	<i>som</i> , efeito , tamanho, aumento, ar	0.214	True	1	2	0.120
Os <i>efeitos especiais</i> são necessários em cenas de batalha, porém, a DC costuma abusar da técnica.	0	efeitos , equipamentos, personagens, dados, filmes	0.158	True	0	1	0.158



CONCLUSIONS

- Feature model combined with BERT classifier is competitive with the baseline
 - OTOH it takes a lot of effort
- Most features work best for English
 - None of the features are great by themselves
- A classification system utilizing idiomatic properties can be implemented with readily available transformer APIs
- Future ideas
 - Combine “good” translation model with a “bad” one
 - Better method for detecting semantic outliers



FINNISH IDIOMS: THEMES

- Building
- Testing
- Analysis



THEME: BUILDING CORPORA

- Generally: building language resources for Finnish idioms
- Idiomatic-literal classification dataset
- Paraphrase datasets
 - Idiomatic-literal paraphrases
 - e.g. “vetää lärvit” in the sentence replaced with “juoda itsensä humalaan”
 - Idiomatic dataset with “idiomatic minimal pairs”
 - Pair of LVCs where the difference is in noun case
 - Both are idiomatic
- Universal Dependencies dataset with MWE tagging
- Idiom dictionary
 - Generation of idiom definition from data



THEME: TESTING

- Utilize the constructed data on various tasks
 - Identify whether V+N is used idiomatically
 - Does adding the constructed data set help in task X?
 - May have to create new tasks and benchmarks for Finnish data
- Idiomatic paraphrasing
 - Idiomatic to literal
 - Literal to idiomatic



THEME: LINGUISTIC ANALYSIS

- Idiomatic minimal pairs
 - How many can actually be found in data?
 - How minimal is minimal?
- (Non-)compositionality, semantic transparency
- Confirm importance of noun case



DATA SET(S)

- Finnish language books from Project Gutenberg
 - 1963 books, 6 million sentences, vocabulary of 2 million
- Chosen as the initial data set because I worked on it previously with my Master's thesis
- Not the ideal set
 - Spoken data might have been better for formulaic idioms
 - Where / when are idioms born?
 - Very heterogeneous set
 - Both Finnish and translated works
 - Fiction, non-fiction
- Initial focus: V+N verbal idioms



ON [FINNISH] IDIOMS

- Most typical idiom is a verb phrase idiom
 - Light Verb Construction (LVC)
- Idiom-prone words
 - basic or nuclear words (semantically neutral)
 - body part nouns
- (Often) sensitive to noun inflection
 - vetää lärviin vs vetää lärvit
 - “to punch in the face” vs “to get drunk”



ON [FINNISH] IDIOMS

■ 50 % of the verbs in verbal idioms:

- olla
- ottaa
- saada
- mennä
- pitää
- vetää
- tulla
- tehdä
- käydä
- panna

Most common nouns:

- silmä
- mieli
- pää
- suu
- naama
- asia
- korva
- aika
- sana
- turpa



METHODOLOGIES

- Dependency parsed data
- Unsupervised
 - Sentiment analysis
- Supervised
- Semantic properties
 - Lexical substitution
 - Fixedness testing
 - (Non-)compositionality
 - Idioms are “semantic outliers”
 - Lexical cohesion
 - Compare edit distance (e.g. Levenshtein) to semantic distance (embeddings)
- Semantic mirroring (translation)
- Lexical methods
 - using dictionaries such as WordNet, Wiktionary



SAMPLE IDIOMS

- iskeä silmää vs iskeä silmänsä
 - ‘to punch the eye’ (wink)
 - ‘to punch [one’s] eye (to become interested in something)
 - Minä en ikinä ole iskenyt silmää hänelle.
 - ‘I have never winked at her’
 - Se mies on iskenyt silmänsä häneen.
 - ‘That man has become interested in her’
- ummistaa silmänsä ‘close one’s eyes’: 3 meanings found (annotation)
 - literal interpretation
 - ‘to choose to not see something (bad/evil)’
 - ‘to die’
- pistää silmään ‘sticks out’
 - mostly idiomatic, but literal interpretations exist



CHALLENGES FOR FINNISH

- Ambiguous forms
 - *silmä* 'eye' possessive form *silmääni* can be..
 - Partitive: silm+ää+ni
 - 'Of my eye'
 - Illative: silm+ään+i
 - 'To my eye'
 - Can be mislabeled by the dependency parser
 - Also *silmänsä*
- Subword tokenization might not split based on case (BERT)



ANNOTATION SCHEMES

- Idiomatic/literal classification
 - Idiomatic (always)
 - Idiomatic in this context
 - Unsure
 - Literal in this context
 - Literal (always)
- Paraphrase set
 - both idiomatic
 - both literal
 - idiomatic/literal
 - Mean same thing or not
 - Pairs with the same noun case



SEMANTIC CHANGE OF IDIOMS

- Speed of change
 - “ei ole kaikki x:t y:ssä”
 - “*Riginoldus: Tosin ei sinulla ole caicki kotona*” (1644)
 - => Most likely older than Finnish written language
- LVCs very resistant to change (?)



THANKS!



REFERENCES

- Idioms in general
 - Miriam Butt and Aditi Lahiri. Diachronic pertinacity of light verbs. *Lingua*, 135:7–29, 2013. doi:10.1016/j.lingua.2012.11.006.
 - Kathy Conklin and Norbert Schmitt. The processing of formulaic language. *Annual Review of Applied Linguistics*, 32:45–61, 2012. doi:10.1017/S0267190512000074.
 - Matthieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. *Multiword Expression Processing: A Survey*. 2017.
- Finnish
 - Kristiina Kortelainen. *Ei ole kaikki muumit laaksossa*. PhD thesis, University of Turku, 2011.
 - Marja Nenonen. *Helppo nakki ? Virittäjä*, 106(3):414–418, 2002. URL: <https://journal.fi/virittaja/issue/view/2418>.
 - Marja Nenonen. *Idiomit ja leksikko: Lausekeidiomien syntaktisia, semanttisia ja morfologisia piirteitä suomen kielessä*. PhD thesis. Joensuun Yliopisto, 2002
 - Marja Nenonen. *Prototypical Idioms: Evidence from Finnish*. *SKY Journal of Linguistics*, 20(20):309–330, 2007.
 - Oksana Petrova. *Of Pearls and Pigs*. PhD thesis, Åbo Akademi, 2011.
- Corpora
 - Hessel Haagsma, Johan Bos, and Malvina Nissim. *MAGPIE: A large corpus of potentially idiomatic expressions*. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, number May, 279–287. 2020.
 - Jenna Kanerva, Filip Ginter, Li-Hsin Chang, Iiro Rastas, Valtteri Skantsi, Jemina Kilpeläinen, Hanna-Mari Kupari, Aurora Piirto, Jenna Saarni, Maija Sevón, and Otto Tarkka. *Annotation Guidelines for the Turku Paraphrase Corpus*. 2021. URL: <https://arxiv.org/abs/2108.07499>, arXiv:2108.07499.



REFERENCES

- Adversarial / contrastive
 - Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend, and Vivek Srikumar. Putting Words in BERT's Mouth: Navigating Contextualized Vector Spaces with Pseudowords. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 10300–10313. 2021. URL: <http://arxiv.org/abs/2109.11491>, arXiv:2109.11491.
 - Rebecca Marvin and Tal Linzen. Targeted syntactic evaluation of language models. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018, pages 1192–1202, 2020. arXiv:1808.09031, doi:10.18653/v1/d18-1151
 - Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng Fu Wang, and Samuel R. Bowman. BLiMP: The Benchmark of Linguistic Minimal Pairs for English. Transactions
- Mask filling
 - Chris Donahue, Mina Lee, and Percy Liang. Enabling Language Models to Fill in the Blanks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2492–2501. 2020. arXiv:2005.05339, doi:10.18653/v1/2020.acl-main.225.
 - William Fedus, Ian Goodfellow, and Andrew M. Dai. MaskGAN: Better Text Generation via Filling in the _____. 2018. URL: <https://arxiv.org/abs/1801.07736v2>.
 - Wanrong Zhu, Zhiting Hu, and Eric Xing. Text Infilling. 2019. URL: <http://arxiv.org/abs/1901.00158>, arXiv:1901.00158.
 - Wanzheng Zhu and Suma Bhat. Euphemistic Phrase Detection by Masked Language Model. Findings of the Association for Computational Linguistics: EMNLP 2021, pages 163–168, 2021. URL: <https://aclanthology.org/2021.findings-emnlp.16>.



REFERENCES

- Meghdad Farahmand and James Henderson. Modeling the Non-Substitutability of Multiword Expressions with Distributional Semantics and a Log-Linear Model. In Proceedings of the 12th Workshop on Multiword Expressions, 61–66. 2016. doi:10.18653/v1/w16-1809.
- Afsaneh Fazly, Paul Cook, and Suzanne Stevenson. Unsupervised type and token identification of idiomatic expressions. Computational Linguistics, 35(1):61–103, 2009. doi:10.1162/coli.08-010-R1-07-048.
- Navnita Nandakumar, Timothy Baldwin, and Bahar Salehi. How Well Do Embedding Models Capture Non-compositionality? A View from Multiword Expressions. In Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP, 27–34. 2019. doi:10.18653/v1/w19-2004.
- Jing Peng and Anna Feldman. Experiments in idiom recognition. COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers, pages 2752–2761, 2016.
- Jing Peng, Anna Feldman, and Hamza Jazmati. Classifying idiomatic and literal expressions using vector space representations. In International Conference Recent Advances in Natural Language Processing, RANLP, volume 2015-Janua, 507–511. 2015.
- Jing Peng, Anna Feldman, and Ekaterina Vylomova. Classifying idiomatic and literal expressions using topic models and intensity of emotions. In EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 2019–2027. 2014. arXiv:1802.09961, doi:10.3115/v1/d14-1216.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Giancarlo D. Salton, Robert J. Ross, and John D. Kelleher. Idiom token classification using sentential distributed semantics. 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, 1:194–204, 2016. doi:10.18653/v1/p16-1019.
- Agata Savary, Silvio Ricardo Cordeiro, Timm Lichte, Carlos Ramisch, Uxoa Iñurrieta, and Voula Giouli. Literal Occurrences of Multiword Expressions: Rare Birds That Cause a Stir. The Prague Bulletin of Mathematical Linguistics, 112(1):5–54, 2019. doi:10.2478/pralin-2019-0001.
- Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. AStitchInLanguageModels: Dataset and Methods for the Exploration of Idiomatcity in Pre-Trained Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2021, 3464–3477. 2021. URL: <http://arxiv.org/abs/2109.04413>, arXiv:2109.04413.
- Thomas Wolf et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 EMNLP (Systems Demonstrations), 38–45. 2020.