PHRASALEX II
Phraseological Approaches to Learner's Lexicography
July 22-23, 2021
Institute for Information Science and Natural Language Processing,
University of Hildesheim

Exploring sentence embeddings: a suitable method for a lexicography-oriented analysis of argument structures?

Fritz Kliche
(kliche@uni-hildesheim.de)

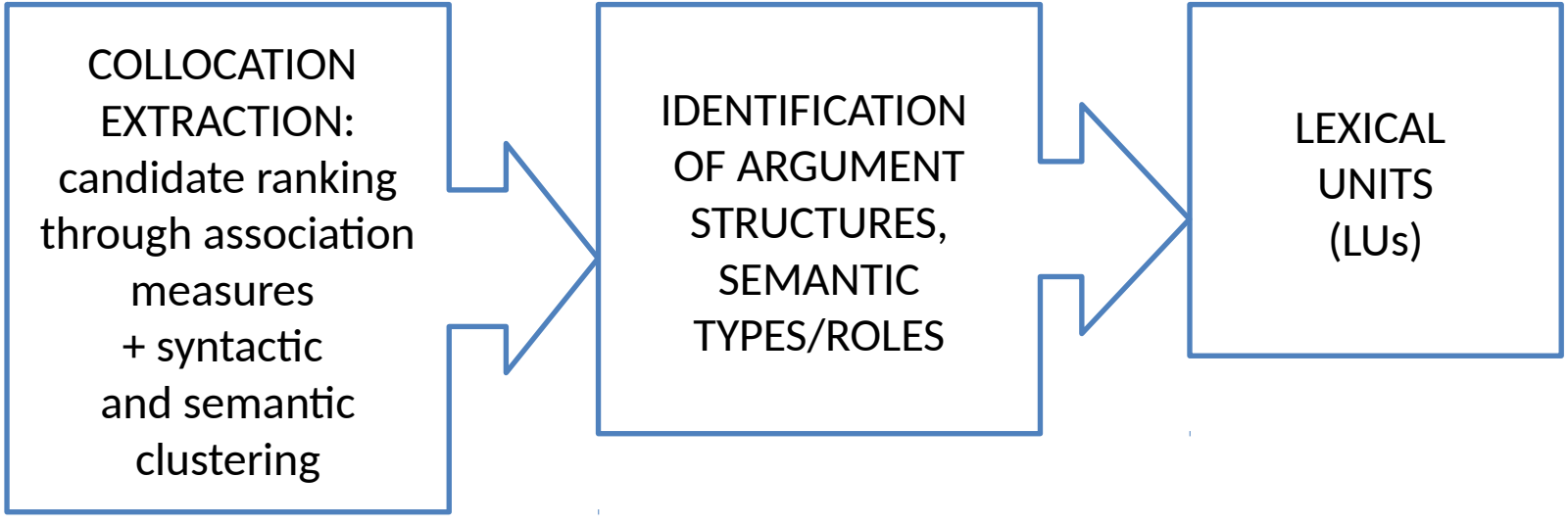
Laura Giacomini
(laura.giacomini@uni-hildesheim.de)

Outline

1. The PhraseBase project: word senses through collocation
2. Word and sentence embeddings for argument structure and lexicography
3. SBERT
4. Experiments:
    a. Clustering
    b. Ranking
5. Conclusions

The PhraseBase project: word senses through collocation

- The creation of Phrase-based Active Dictionaries (PADs) in PhraseBase has a strong phraseological orientation.
- Word senses (Lexical Units, LU) and sense clusters are identified by means of a collocation-based method (Giacomini & DiMuccio-Failla, 2019)



Word and sentence embeddings for argument structure and lexicography

Word and sentence embeddings have recently gained some popularity in lexicography, not least due to the breakthrough of neural machine translation.

- Sørensen, N. H. & Nimb, S. (2018)
- María José Domínguez Vázquez (2021)

Research questions

We would like to test the usefulness of sentence embeddings in the specific context of the PhraseBase project in Learner's Lexicography. The focus on sentence embeddings rather than on word embeddings is motivated by the phraseological orientation of our approach.

Research questions:

- Can the analysis of argument structures through sentence embeddings be employed as
  - a complementary method
    - for detecting sense-related argument structures?
    - for finding, among argument-embedded lexemes, suitable semantic types/roles?
  - a validation method for Lexical Units in the PADs?

We are not focusing on finding different words with similar argument structures, we concentrate on single words.

Key data

- Selected words: *follow* (verb); *arm* (noun)
- Input data are retrieved from the British National Corpus (BNC) by means of the Sketch Engine. Kilgarriff et al. (2014)
- BNC: 100 million word corpus of written and spoken British English of the late twentieth century from a wide range of sources
- Methods:
  - **Clustering** lexicalized instances of syntactic patterns of a word
  - **Ranking** occurrences of a word in context by their similarities to an instance of a specific LU of that word

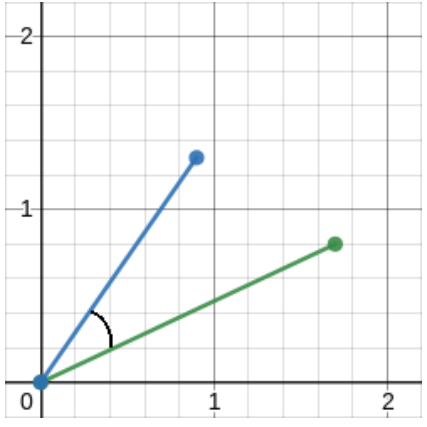
follow a 1958 House
follow a 2 year
follow a 2 year joint
follow a 2-year cycle
follow a band purely
follow a Bensonian way
follow a biological pattern
follow a biorhythmic pattern
follow a braille trail

Method

SBERT (or: Sentence-BERT)

Reimers & Gurevych (2019)

- Sentences (or other sequences) are represented as vectors.
- Vectors which are close in a vector space are semantically similar.
- Similarity is measured by cosine similarity.
- SBERT was used for both clustering and ranking.



Drawn with www.desmos.com/calculator

Clustering method

Input

- List of instances

Parameters

- Threshold for the similarity
- Minimal cluster size

Example

Instances which are clustered

Cluster 31	follows from this necessity
Cluster 31	follows from the need
Cluster 31	follow from the particular requirement

Result

- Instances with cosine similarity > threshold are clustered.
- Clusters with more entries than minimal cluster size are given as output.

Ranking method

Input

- List of instances
- “Seed instance”

Result

- The input instances ranked by their similarity to the seed instance

Example

Instances for “follow” patterns ranked by their similarity to the seed “follows the story”

follows the story	followed the story so	0,8509
follows the story	following the story closely	0,7988
follows the story	following a short story	0,7979
follows the story	followed the story so far	0,7973
follows the story	Following the ongoing story	0,7397
follows the story	followed multiple story lines	0,6645
follows the story	follows the second-person narrative	0,6628
follows the story	following the news two	0,6496
follows the story	following first the news	0,6479
follows the story	follow the the the	0,6468

Collecting data

- Goal: semantically clustering/ranking instances of specific **syntactic patterns** of a word

- Input example:

  LU: *to follow a person/vehicle*

  syntactic pattern of the LU
  [lemma="follow"] [tag="(DT|N|J|PP|CD).*"]+

Applying SBERT: Clustering

Input pattern	Examples	Threshold	Clusters	Cluster size
[lemma="follow"] [tag="(DT|N|J|PP|CD).*"]+	to follow a vehicle, to follow a rule, to follow a discussion, …	0.6	88	6-2479
		0.6	223	2-2479
		0.55	35	2-4562
		0.5	4	2-6543
		0.7	25	10-3181
		0.8	83	10-571
		0.8	196	6-572
		0.9	140	6-35

Observations: Clustering

- Each different N appears in only one cluster.
- Most of the results are 2- or 3-grams, but also longer combinations have been detected (*followed an intelligence assessment*, *follow the bank robbers' plans*, *following an interest rate change*).
- There are some syntactically complex combinations: *followed the short-back-and-sides hair-cut*

Results are extremely heterogeneous:

- syntactic variation of longer combinations (cf. C. 78)

followed audience taste
follow audience taste
followed taste
followed audience
followed the taste
follows a taste
follow audience

- tense variation (cf. C. 78)
- duplicates (cf. C. 59)

follow the abolition
followed the abolition
following abolition
following the abolition
following the abolition

- paradigmatic variation with N belonging to the same semantic field (cf. C. 79)

followed a torrent
followed a stream
followed the storm
followed the stream
followed an outpouring
followed the flood

- most variation is of syntactic nature (cf. C. 51)

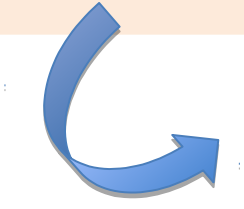
following the patient's stroke
following a stroke
following a stroke
following the stroke
Following a stroke
following the patient's
following the patient
follows the patient
followed the patient

following the promptings	following the examples	follows the tales
following the signs	following the instructions	following the goods
Following the directions	following the whims	following the headlines
followed these events	followed the actions	following the departures
following these events	Following the moves	Following these sections
following the instructions	following the arms	Following the publications
following the reactions	followed all sorts	follow the footsteps
following the results	following the Greens	following these forays
following the Ports	follow these developments	Following these conditions
following the changes	Following the experiences	Following these giants
following the activities	followed the verderers	following the women's
Following these moves	following some comments	following the sales
following those events	following a Governors'	follows the adventures
following the proceedings	following the tails	following the directions
Following the remarks	Follow the footsteps	following the pieces
Following the events	follows the words	follow the others
following the developments	following the others	following the movements
following the implications	Following the implications	following those things
follow the words	followed the activities	following these steps
Following the approaches	following the sounds	following the cattle
following the paths	following the intimations	following the creation
following the words	Following the events	following the inclusion
followed the others	following the actions	following the article
following the ladies	following the reports	following the lead
following the guards	Following the lines	following the process
followed the words	follows	following the publication
following the twists	following the footsteps	follows that
follow the leads	following the incidents	Following that episode
following the announcements	following the names	following the advice
Following the instructions	followed the events	Following the publication

- Large clusters (e.g. C. 0) cover all these phenomena, other than smaller clusters, however, they contain several different semantic fields as well as semantically independent Ns, which makes the classification of Ns a very demanding task.

- From a semantic perspective, the attribution of words and semantic fields to specific clusters is not predictable/coherent.

Input pattern	Examples	Threshold	Clusters	Cluster size
[lemma="follow"] [word="on"]? [word="from"] [tag="dt"]? [tag="N.*"]+	the proposal follows on from the discussion	0.9	-	≥ 6
		0.7	24	≥ 2
[lemma="follow"][word="up"] [tag!="SENT"]{1,5} [word="with"]?	follow up your mail-out with a phone call follow up with a phone call follow up the letters with phone calls		…	…
	follow up with a personal call		…	…
[lemma="follow"] [word="up"] []{0,7} [word="with"] [tag="(DT|N|J|PP|CD).*"]{0,7}	follow up a tape with a phone call follow up with in-depth interviews followed up by in-depth interviews with managers followed up in qualitative interviews with a panel follow up some responses with face-to-face interviews follow up the agreement on constitutional amendments with a detailed agreement followed up their constitutional provision with other reforms		69	≥ 3



Longer sequences can give insights into complex collocations.

Applying SBERT: Ranking

- Goal: ranking occurrences of a word in context by their similarities to an **instance** of a specific LU of that word

- Input example:

LU: *to follow a person/vehicle*

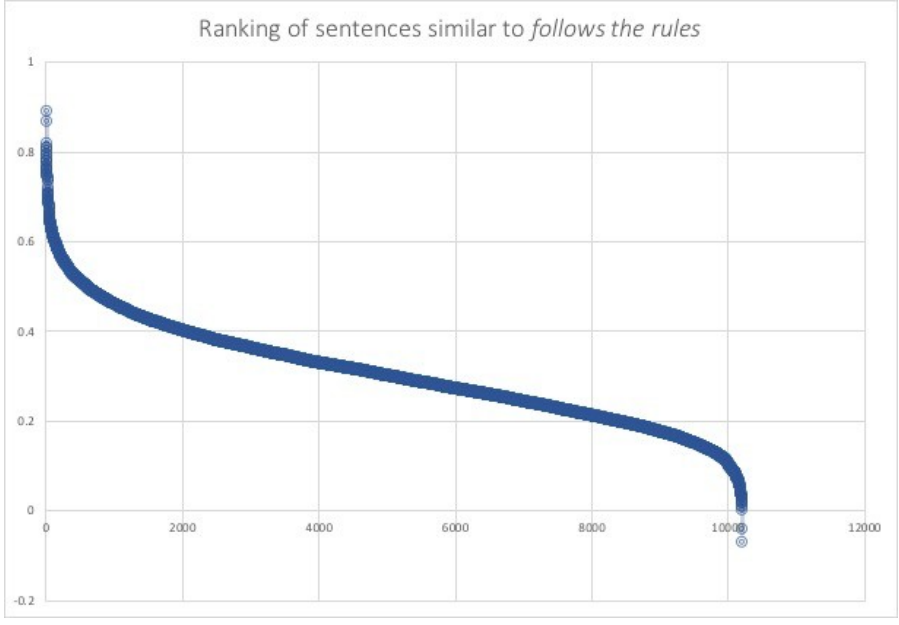
instance of the LU in the BNC (seed sentence): …*follows the police car*

follow (verb)

LU	**Instance in BNC (input)**
to follow a vehicle	follows the police car
to follow a path	follows the path
to follow someone's example	follows his example
to follow a rule	follows the rules
to follow a course of development	follows a pattern
to follow with one's eyes a person/ object moving	followes him with her eyes
to follow (on) from sth.	follows on from research in the US
…	…

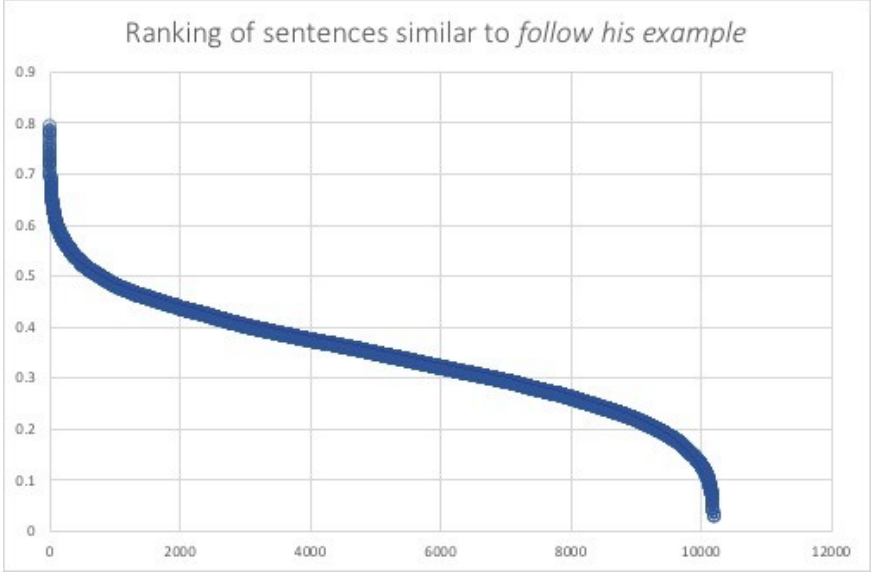
Similarity between sentences from the BNC and a specific input sentence (excerpts):

follows the rules	follows the normal rules	0.8942
follows the rules	following the rules well	0.8692
follows the rules	follow the rules you	821
follows the rules	follow the normal rules	0.8137
follows the rules	following the essential rules	0.8106
follows the rules	follow the above rules	0.8045
follows the rules	follow the rules precisely	0.7994
follows the rules	follow these rules they	0.7944
follows the rules	follows his own rules	0.7904
follows the rules	Follow the basic rules	0.7843
follows the rules	following these basic rules	0.7836
follows the rules	follow the simple rules	0.7826
follows the rules	following the same rules	0.7734
follows the rules	followed these basic rules	0.7702
follows the rules	follows the same general rules	0.7662
follows the rules	following the local rules	0.7639
follows the rules	follow the same rules	0.7619
follows the rules	follow these simple rules the	0.76
follows the rules	follow the above rules you	0.7521
follows the rules	following the rule it	0.7517
follows the rules	follows the same rule	0.7493
follows the rules	following a few simple rules	0.7466



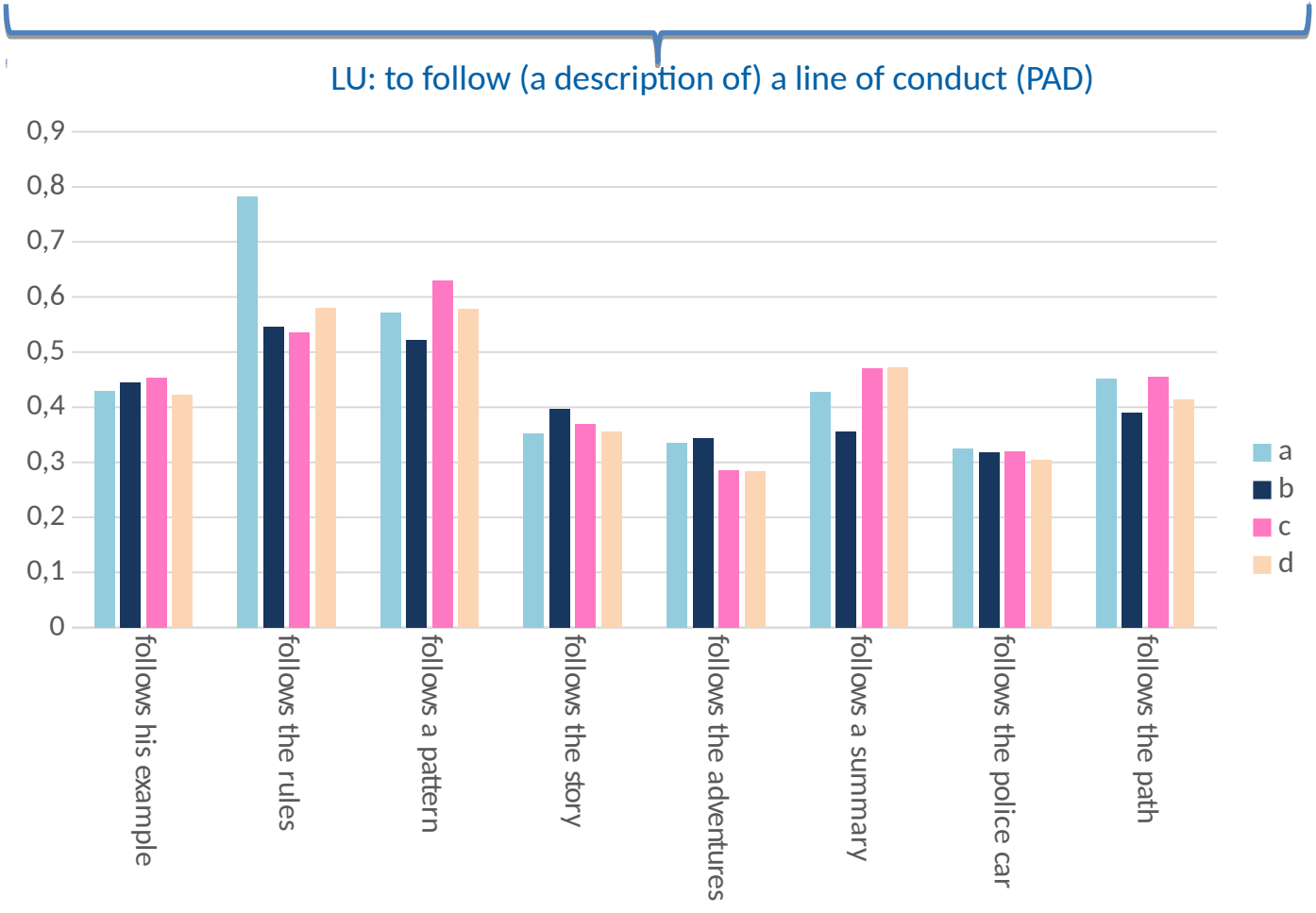
Similarity between sentences from the BNC and a specific input sentence (excerpts):

follows his example	follows a simple example	0.7902
follows his example	followed his bad example	0.7796
follows his example	followed the excellent example	0.7796
follows his example	follow the successful example	0.7776
follows his example	following a parents example	0.7722
follows his example	follow your good example	766
follows his example	following the German example	0.7598
follows his example	followed her example rather	0.7528
follows his example	Following your example I	0.7495
follows his example	following the English example	0.7473
follows his example	following two examples the	0.7426
follows his example	follow her fathers example	0.7351
follows his example	follow her brothers example	0.7321
follows his example	Following the successful example	0.7292
follows his example	followed the Wolfsons example	0.7271
follows his example	follow this good example	0.7221
follows his example	follow his own lead	0.72
follows his example	follow the good example	0.7192
follows his example	follows this bold example	0.7156
follows his example	follow the German example	0.7149
follows his example	Following the earlier example	0.7137
follows his example	followed the Russian example	0.7004



Similarity between sentences from the BNC (a-d) and different input sentences (excerpts):

a) follow the simple **rules** b) followed **instructions** c) following the same **procedure** d) following the same **procedures**



Combining both methods

Clustering on ranked data

Within sets of ranked data, we cluster sentences which cosine similarity ≥ 0,4 to the input.

Data to be clustered	Threshold	Clusters	Cluster size
follow a path	0.7	132	≥ 2
follow a rule	0.7	81	≥ 2

Objects of *follow* in the clusters of *follow a path* (clusters 1-24):

1 path route pattern road track trajectory trend line order sequence trail procedure direction	**2** girl sister woman visitors move istructions example lead	**3** move attempt policy committee company campaign teams partners	**4** diet dietary advice schedule	**5** catch run score move victory match	**6** road accident	**7** clinical course recommendations marketing concepts ideas	**8** son brother brothers footsteps nephew brothers example fathers example friend
9 ?	**10** operations stages programme courses strategies steps themes exercises	**11** career campaign route	**12** Christian practice Christian example Christian responsbility religion religious instincts church servce	**13** thickness length amount	**14** ?	**15** south traverse road south track south south coast ditch south-east lane south-west	**16** manufacturers nstructions safety guidelines manufacturers directions
17 step-by-step routine step-by-step programme step-by-step account step-by-step guide	**18** wood forest forest path forest road	**19** bridleway steps	**20** rhine model german example european model	**21** cliff edge cliff path steep edge waters edge	**22** IBM Corp Model business model	**23** ?	**24** parents party line parents example party line family tradition

Observations: Ranking

- Ranking experiments are carried out on instances of *available* LUs.

- These experiments can be useful to check
  - how close a corpus sentence is to each LU,
  - which LUs are most similar to each other (overlapping senses).

- Clustering on ranked data further refines ranking results.
  - The data assigned to the LU *follow a path* in PhraseBase are not matched by all clusters (most of them are in C. 1).
  - Some clusters (grey fields) seem to have no conceptual connection to *follow a path*.
  - Some others provide interesting clues on possible connections with other LUs (metaphorical/ metonymical derivation from *follow a path*) that cannot be quicky discovered with our present method for collocation extraction.

arm (noun)

LU	**Instance in BNC (input)**
the arm of a person	he broke his arm
the arm of a person	muscular arm
the arm of an instrument/ object	arms of the cross
the arm of a machine	mechanical arm
the arm of a piece of seating furniture	arm of the sofa
the arm of an organisation	arm of the company
an arm of land/water	arm of the sea
arms	to bear arms
…	…

Conclusions

Overall applicability to lexicography is difficult to assess.

General issues in the context of PADs:

- Choice of the initial input for the sentence transformer
- Analysis of implicit semantic arguments
- Amount of required manual work to clean, select and order data

Benefits in the context of PAD compilation:

- Most of the well-formed combinations (clustering and ranking methods) match the collocations extracted by our method (different corpus!).
- While clustering alone provides too heterogeneous results, ranking and clustering on ranking data can be a useful method for validating PAD data and gaining more insights into LU relations.
- We need to devise best practices (e.g. parameters for a word class, a language, …).

References

DiMuccio-Failla, P. & Giacomini, L. (2017). In: M. Mitkov (ed.). *Computational and Corpus-Based Phraseology. Second International Conference*, Europhras 2017, LNAI 10596. Springer, 290-305.

Domínguez Vázquez, M. J. (2021). Zur Darstellung eines mehrstufigen Prototypbegriffs in der multilingualen automatischen Sprachgenerierung: Vom Korpus über word embeddings bis hin zum automatischen Wörterbuch. *Lexikos*, *31*(1).

Giacomini, L. & DiMuccio-Failla, P. (2019). Investigating Semi-Automatic Procedures in Pattern-Based Lexicography. In: *Proceedings of the eLex 2019 conference. Electronic lexicography in the 21st century*. Sintra, Portugal.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1, 7-36.

Reimers, N, & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: *Proceedings of EMNLP 2019*.

Sørensen, N. H., & Nimb, S. (2018). Word2dict – lemma selection and dictionary editing assisted by word embeddings. In: *Proceedings of the 18th EURALEX International Congres: Lexicography in Global Contexts*, 819-827.