PHRASALEX III

PHRASEOLOGICAL APPROACHES TO LEXICOGRAPHY

PHRASEBASE TALK

Data acquisition in PhraseBase: core method and exploratory case studies

Laura Giacomini
Fritz Kliche
Jannis Nolte

universität innsbruck

Universität Hildesheim

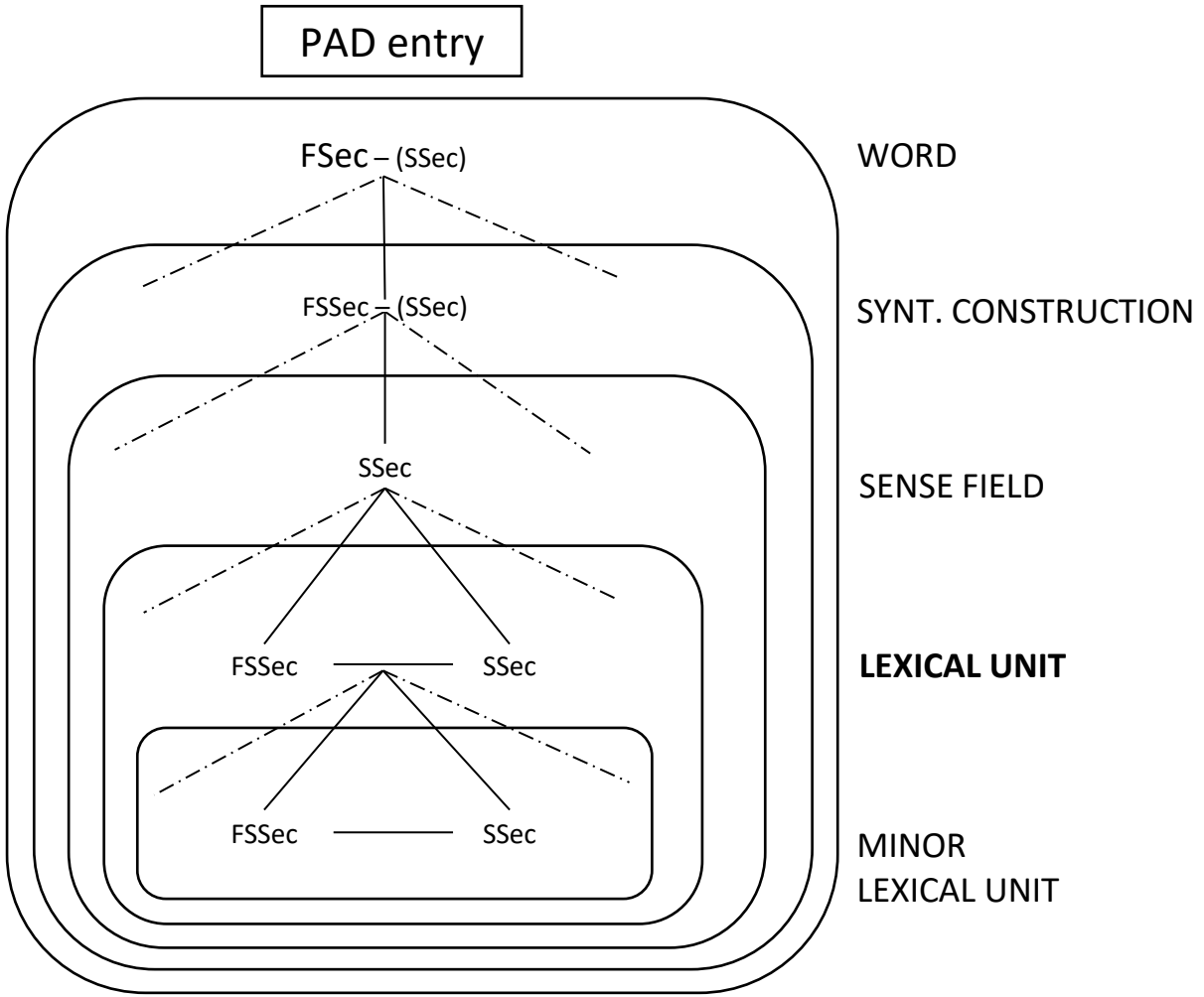
Overview

- The core method:
  - The PAD microstructure
  - The beginnings: the CPA technique
  - A change in perspective: interpreting collocations and examples
- Broadening the scope:
  - Experiment 1: Clustering sense variants using BERT’s contextualized word embeddings
  - Experiment 2: Bootstrapping additional sense variants using Zero-Shot Classification
- Conclusions

The PAD microstructure

DiMuccio-Failla & Giacomini (2022)

FSec = Formal Section
FSSec = Formal-Semantic Section
SSec = Semantic Section



The PAD microstructure

agree \ə ˈgriː\

VERB, REGULAR

to AGREE ‹WITH sb. / s. opinion› ‹ON/ABOUT s.e.› [opinion]

/ABSOLUTELY/TOTALLY/STRONGLY/CERTAINLY / NOT NECESSARILY / NOT QUITE

▶ to think that sb. is right ‹on/about s.e.›

~

▶ to share s. opinion ‹on/about *›

to AGREE (with each other) ON s. decision ‹THAT...›/‹TO do sth.› [expr. of decision]

|TOGETHER | BETWEEN ONESELVES | MUTUALLY [at a particular moment]

▶ ⌜to decide to do sth.⌝ together

~

▶ to choose sth. together

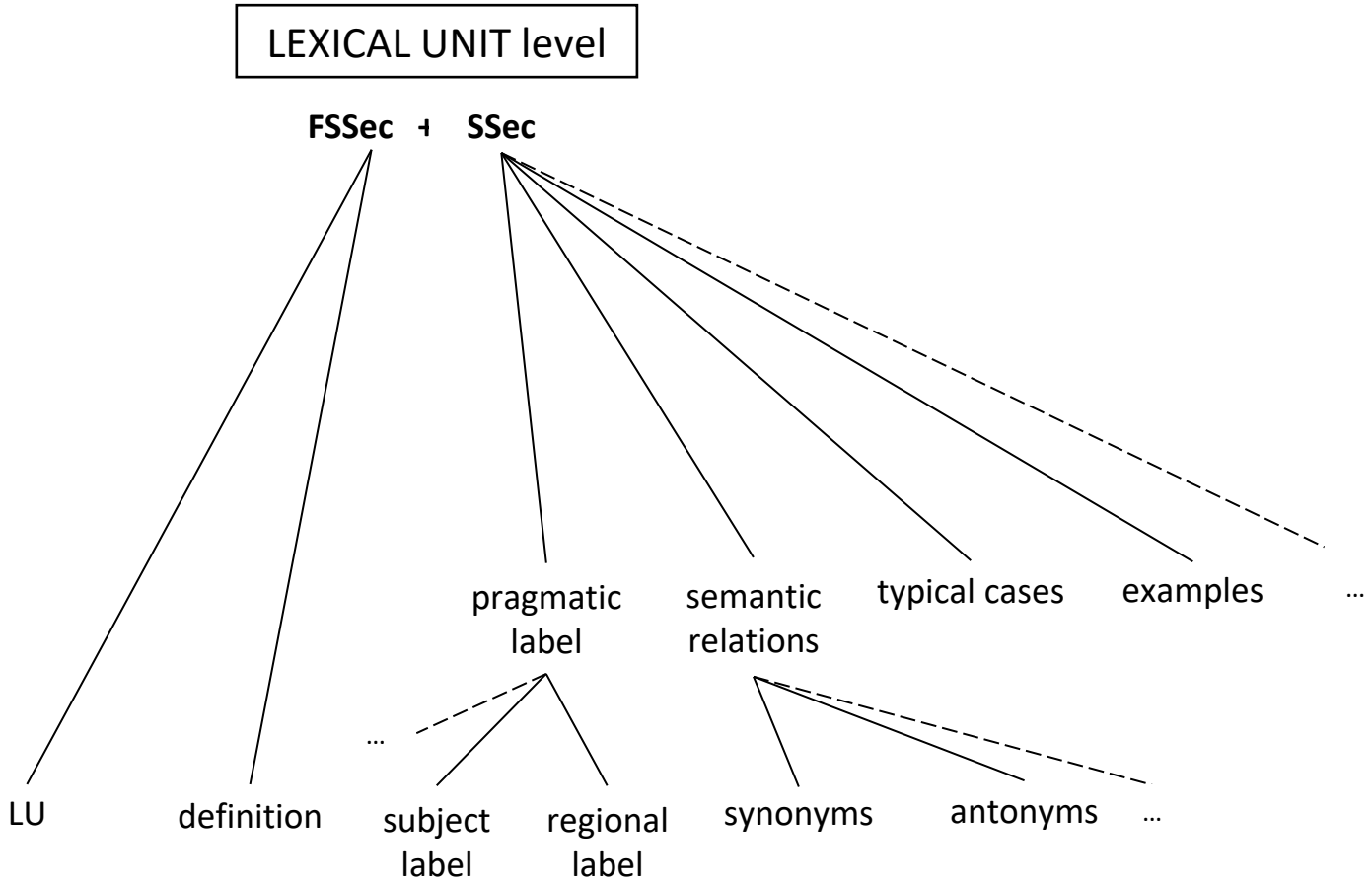
1 to AGREE ‹WITH a given person› ...

... ■ ‹ON a given SUBJECT/TOPIC/ISSUE› OR ‹ABOUT a certain entity› ○ to think that a given person's opinion/assessment ‹on ·› OR ‹about ·› is right

• EXAMPLES: ① ... • SYNONYMS: ① [FML.] to CONCUR ‹WITH ·› ‹ON/ABOUT ·› ② to SHARE a given person's OPINION ‹ON/ABOUT ·› ③ [FML.] to BE IN AGREEMENT ‹WITH ·› ‹ON/ABOUT ·› ④ to THINK THE SAME ‹AS a given person› ‹ON/ABOUT ·› ⑤ [FML.] to BE OF THE SAME OPINION=MIND ‹AS a given person› ‹ON/ABOUT ·› ⑥ to HAVE THE SAME OPINION ‹AS a given person› ‹ON/ABOUT ·› • ANTONYMS: ① to DISAGREE ‹WITH ·› ‹ON/ABOUT ·› ② [FML.] to NOT SEE EYE TO EYE ‹WITH ·› ‹ON/ABOUT

The PAD microstructure

DiMuccio-Failla & Giacomini (2022)



The beginnings: the CPA technique

Data for the Phrase-Based Active Dictionary (PAD), with focus on how to identify Lexical Units (LUs), e.g.

to AGREE ‹WITH a given person› ‹ON a given SUBJECT/TOPIC/ISSUE› OR ‹ABOUT a certain entity›

The initial procedure involved Corpus Pattern Analysis (CPA), a “technique for mapping meaning onto words in texts” (Hanks 2004).

Query abandon-v **250** > Random sample **250** (4.6 per million)

Page 1 of 13 Go Next | Last

the high-minded Virginsky who ` will never , never	abandon	2	these bright hopes ' (my italics) , and another
Labour did not agree that Britain could or should	abandon	1	development , either for itself or for the developing
minority tribes into the governance of the state . He has	abandoned	4.f	much of his Marxist baggage and , so far , set his
morning for a briefing session . They say they will	abandon	1	the dispute over the Government 's refusal to raise
review . After discreet soundings , they prudently	abandoned	2	the idea , which would have involved a major encroachment
assure Janet Daley that these children have not been	abandoned	6	to a type of educational apartheid . It is views
and Excise rates by 1993 , the ministers agreed to	abandon	1	key provisions for revising VAT collection arrangements
quality of British filmmaking nosedived . UK filmmakers	abandoned	1	their innovations with film narrative , producing
oneself , and not selling out to Hollywood , really mean	abandoning	u	melodrama for realism , showmanship for seriousness
had come to demand . Dr Clark was now seen hastily	abandoning	4	all those notes , containing all the furious denunciations
former scourge of the dozy British motorist , has not	abandoned	1	his life 's work just because Her Indoors shunted
£11BILLION a year Lloyd 's of London insurance market is	abandoning	1	rules forcing underwriters to specialise in particular
another to impose controls on everything . Mr Gonzalez	abandoned	1	another long-running tradition by which Argentine
. Only a fortnight ago , the Lithuanian parliament	abandoned	2	the constitutional guarantee of the party 's leading
1987 that Gen Noriega was negotiating with the US to	abandon	1.a	his command for a comfortable exile , sent him a
1989 in protest against the government 's decision to	abandon	1	the British PWR programme . The IAEA and safeguards
weapons if the Soviet Union broke up , was abruptly	abandoned	1	last year . Too political . The distinction between
mean taking up an idea first mooted last year , but	abandoned	1	as too risky : a full coalition between more cautious
which will radically affect their lives . It means	abandoning	6	the working class to sectarian politics and violence
fight against fascism , and eventually the need to	abandon	2	its emphasis upon political independence . It was

Page 1 of 13 Go Next | Last

Baisa et al. (2015)

The beginnings: the CPA technique

CPA was employed for creating the entries of the *Pattern Dictionary of English Verbs* (PDEV, pdev.org.uk).

abandon

1 [[Human | Institution]] abandon [[Activity | Plan]]
[[Human | Institution]] stops doing [[Activity]] or does not begin to do [[Plan]]

2 [[Human | Institution]] abandon [[Concept]]
[[Human | Institution]] ceases to believe in or be influenced by [[Concept]]

3 [[Human_Group = *Military* | Human]] abandon [[Location]]
[[Human | {Human_Group = Military}]] goes away from [[Location]] and does not live there any more, or (in military contexts) ceases to defend it

4 [[Human]] abandon [[Artifact]]
[[Human]] ceases to use or have possession of [[Artifact]] and leaves it somewhere at random

5 idiom [[Human]] abandon {ship}
[[Human]] leaves {ship} at sea, by jumping into a life boat, life raft, or into the sea

6 [[Human 1 | Animal 1]] abandon [[Human 2 | Animal 2]] [to [[Anything = *Bad*]]]
[[Human 1 | Animal 1]] goes away from and ceases to care for or look after [[Human 2 | Animal 2]], with the result that [[Anything=Bad]] may get them or happen to them

7 [[Human]] abandon [[Self]] [to [[Activity | Attitude]]]
[[Human]] does [[Activity]] or adopts [[Attitude]] without thinking or caring about what is right, proper, or required by duty

8 [[Human]] abandon [[Self]] [to [[Deity]]]
[[Human]] gives up free will and allows [[Deity]] to make all decisions for him/her

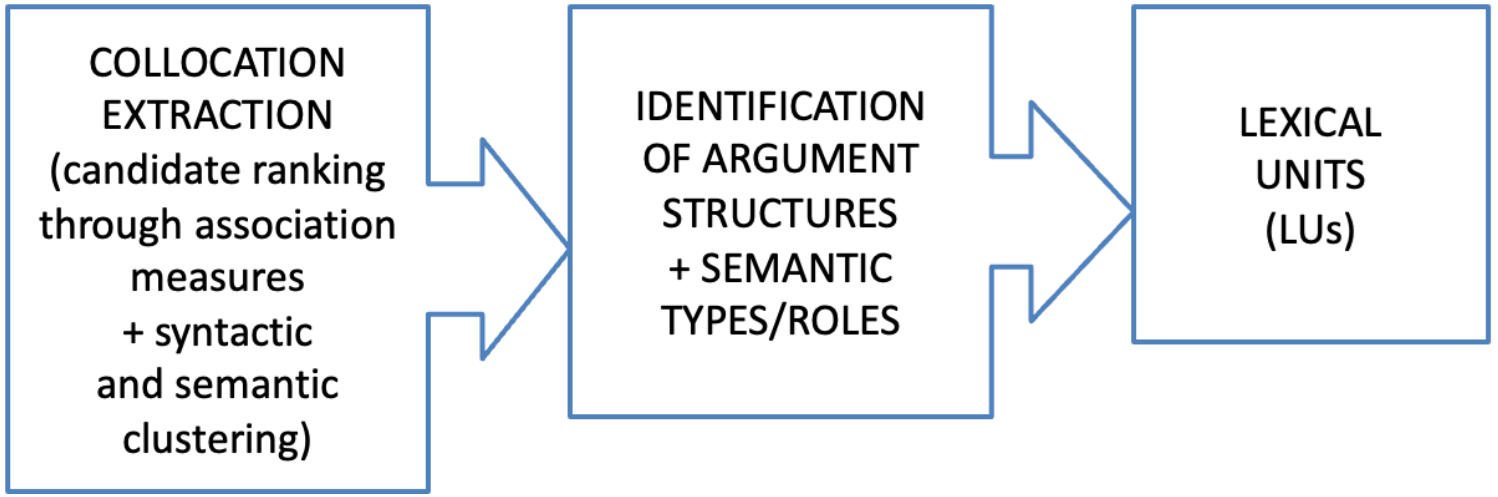
→ Main issues:

- heterogeneous data (free word combinations, collocations, idioms, errors; creative use of language)
- difficult to pinpoint meanings

A change in perspective: interpreting collocations and examples

New procedure: from concordance analysis to **collocation analysis** (Giacomini & DiMuccio-Failla 2019, Giacomini et al. 2020).

Core method:



Collocates of a word are used

- for semantic type/role identification,
- as such in LUs,
- as Recurrent Cases of LUs,
- for Sense Field disambiguation.

A change in perspective: interpreting collocations and examples

The core method **also** includes the analysis of **examples**:

- corpus concordances,
- examples from other language-specific sources (e.g. dictionaries, newspaper archives, electronic books, …)

to paint

	A	B	C	D	E	F	G
1	D. OBJECT		…	…	…	…	EX
2	(a piece of) furniture						They're painting the building white , with gray-green highlights in the inset areas under the windows.
3	the ceiling						
4	the walls						
5	the house						
6	the exterior						
7	the interior						
8	the door						
9	the living room						
10							
11							

→ Examples are useful to complement/verify collocation clusters.

→ However: common issues with examples as microstructural items:

- incompatible with a dictionary for learners since incomplete, vague, ambiguous, non-conventional
- guidelines for adjusting/creating examples are necessary

Broadening the scope

Most of the work for creating PAD entries is manual!

The most difficult task is to identify (and formulate) Lexical Units in such a way that

- they are user-friendly enough to be interpreted by the PAD user immediately after short training,
- they are expressed in a semi-formalized manner.

Formulating Lexical Units implies using suitable semantic types/roles.

> Patrick Hanks (2004: 87–88) stressed that **identifying the right semantic types for selectional preferences**, in particular **not leaving out normal usage** on one side and **not generalizing into abnormal usage** on the other side, **requires linguistic and ontological expertise**: "Among the most difficult of all lexicographic decisions is the selection of an appropriate level of generalization on the basis of which senses are to be distinguished" (ibid.: 88).

This statement is crucial for guaranteeing a high quality standard in lexicography.

However:

→ The creation of an entry involves a long series of steps → at least for some of them it is interesting to look for new NLP approaches → **goal: partial workflow automation**

Broadening the scope

Together with Fritz Kliche (University of Hildesheim): case studies aimed at disambiguating word senses have been carried out based on

- Contextualized Word Embeddings (Experiment 1) and
- Zero-Shot classification (Experiment 2, with Jannis Nolte).

The first case study is presented in:

➢ Fritz Kliche & Laura Giacomini (forthcoming, end 2024): “Exploring BERT’s contextualized word embeddings: a suitable method for a lexicography-oriented analysis of argument structures?”. In: Giacomini, L. & Piunno, V. (eds.), *Patterns of meaning in lexicography and lexicology*, Berlin/Boston: de Gruyter.

Experiment 1: BERT's contextualized word embeddings

- Goal: Finding PAD word meanings for *follow* - using NLP methods

---

- Start:
  We collect instances from the British National Corpus (BNC; a general language corpus of ~100M tokens), matching for the pattern:

  `[lemma:follow]`
  `+ up to 10 tokens with POS tags for`
  `nouns, determiners, adjectives, prepositions, ...`
  (but: no verbs)

- After removal of duplicates
  we work with 48,347 instances.

- Can we **cluster the instances semantically**?

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

Filtering

- We filter the instances:
  - We delete all tokens except for nouns, prepositions and conjunctions.
  - We lemmatize the texts.

- Example:
  - **From BNC:** *following the country 's first general election under universal suffrage on April*
  - **Filtered:** *follow country election under suffrage on april*

- The filter creates again duplicates, which we remove.

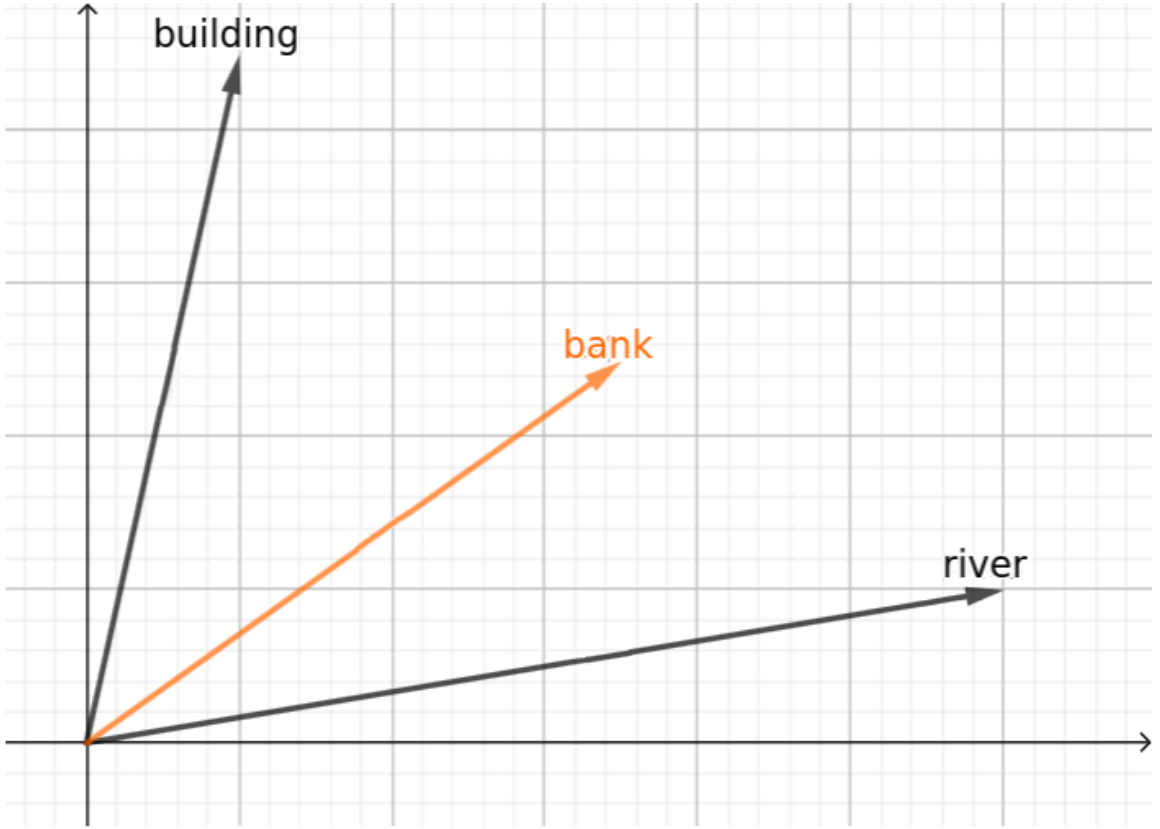
- 30,559 instances remain.

Underlying idea

Word Embeddings

- "Computing meaning"

  Current NLP methods represent the meaning of a word (or of a group of words, or of a sentence, or of a text, ...) with a **vector in a high-dimensional coordinate system**.
- Can we cluster the BNC instances for finding PAD word meanings?



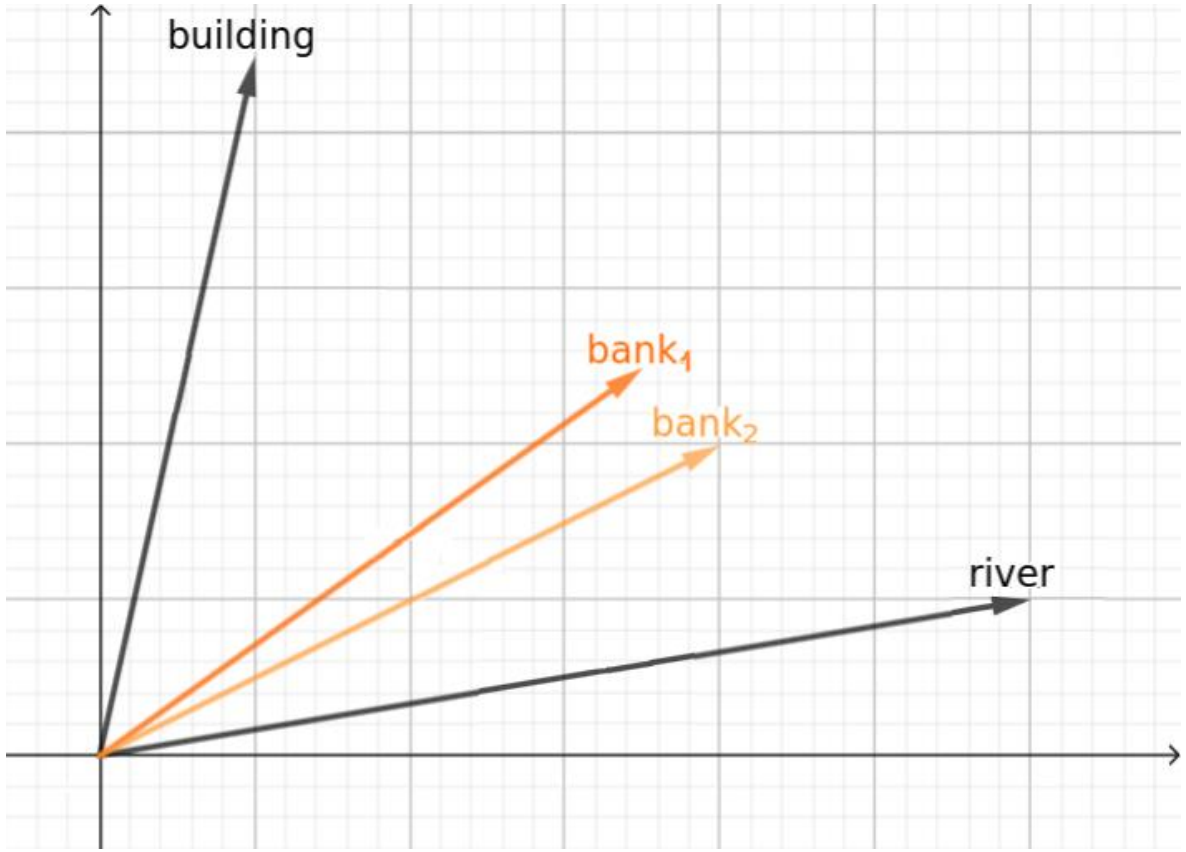
Source of the plots: www.geogebra.org, author of the tool: Fei Chung, link to the licence: https://www.gnu.org/licenses/gpl.html

Underlying idea

<u>Contextualized</u> Word Embeddings
in BERT (Devlin et al., 2019)

- The vector representation of a word is moved towards similar words found in the same context.

  Example: ***bank*** *of the river*
- Idea:
  Can we use these vector movements for finding word meanings?



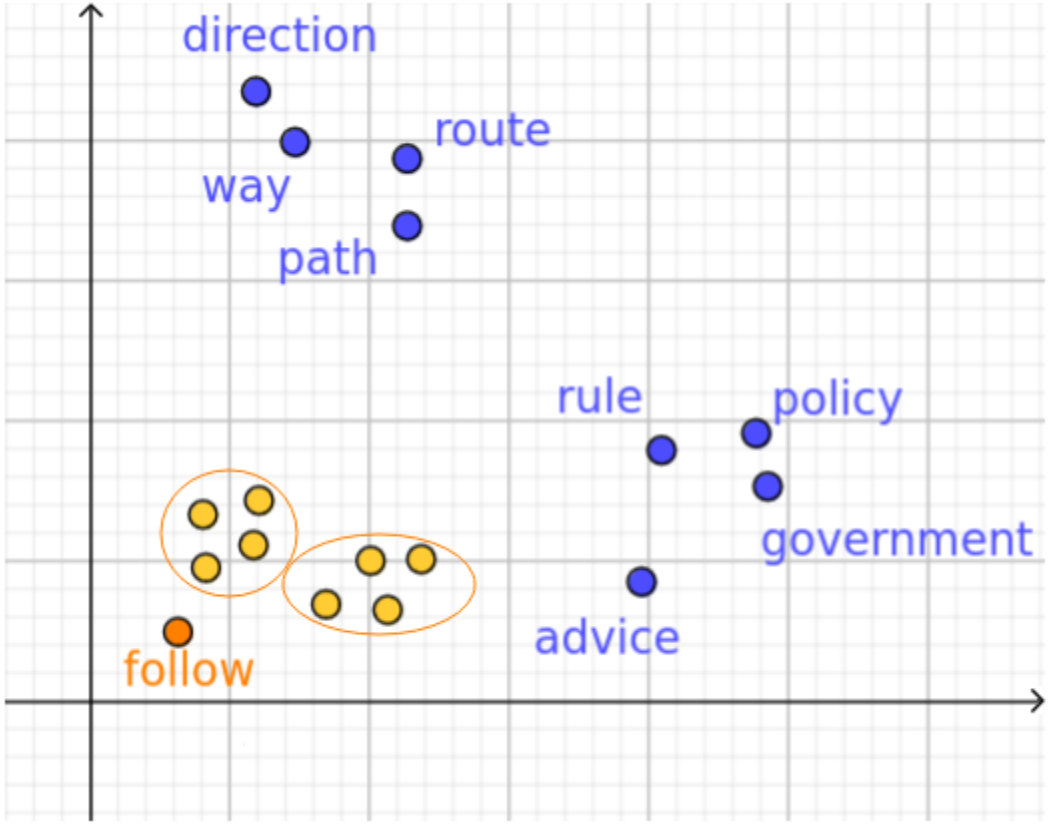
Example: In the sentence "...bank of the river...",
the vector of **"bank"** is deviated in the direction of **"river"**.

Underlying idea

- We process the 30.559 instances from the BNC with BERT.
- We are **only interested in the deviated vectors of *follow***.
- Do we then find PAD word meanings when we cluster these vectors?

→ We use ***k*-means for clustering** the deviated vectors, in three runs:

- 500 clusters
- 1000 clusters
- 2000 clusters



Clustering

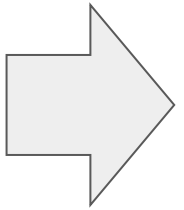
We sort the resulting clusters by size
(= by number of instances in the cluster).

***k*-means cluster size**	**500**	**1000**	**2000**
Largest cluster	142 instances	107 instances	51 instances
Cluster of median size	56 instances	29 instances	14 instances
Smallest cluster	13 instances	2 instances	1 instance

Examples from a cluster

For *k*=1000:
Examples for cluster #500 (total: 29 instances)

followed the butler into the private rooms
followed Culley ‘s gaze round the bleak room
followed her daughter through to the kitchen
followed Edgar a few paces across the room
followed their father into the kitchen



Semantic interpretation

follow a person spatially to somewhere

Syntactic interpretation

follow
+ dir obj.
+ PP

Evaluation of the clusters

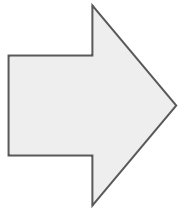
For each run (*k*=500; 1000; 2000), **we evaluate 40 clusters**:

Can we give both a **semantic** and a **syntactic** label for the majority of the instances in the cluster?

We label as “unclear” otherwise.

Results:

- For *k*=500: 26 clusters are unclear.
- For *k*=1000: **12 clusters are unclear**.
- For *k*=2000: 17 clusters are unclear.



k=1000 best suited for our goal?

Or are *k*=500 and *k*=2000 not “worse”

but rather: different?

Examples for *k*=500

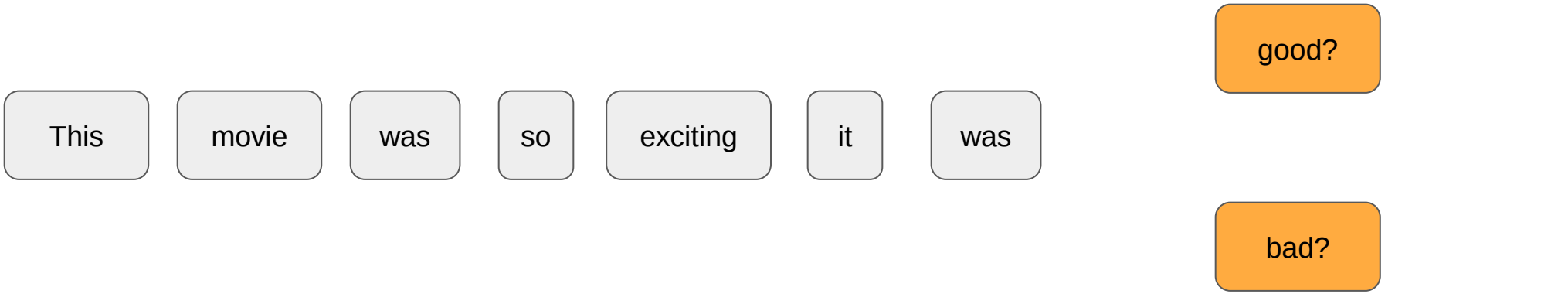
- k=500 results in large clusters.
- Only 4 of the 10 clusters of median length can be given a semantic-syntactic interpretation,
- but some clusters show interesting results.
- E.g., a cluster with the label **“negatively come to an end”**

- *following its abandonment of a nuclear fuel*
- *following any accident damage to your caravan*
- *following a breakdown in health*
- *following the breakdown of her marriage*
- *followed the breakup of his parents ' marriage*
- *following the loss on disposal of businesses*

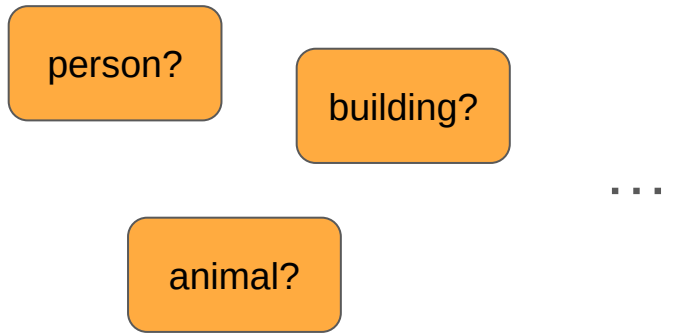
- *follow the disappearance or non-appearance of a rash , for instance if*
- *followed the crash of 1987 and a period of lacklustre performance*
- *follow their European Cup collapse less than a week ago*
- *following the defeat of his buy-out plan*

Experiment 2: Zero-Shot Classification

Underlying idea (Yin et al., 2019)



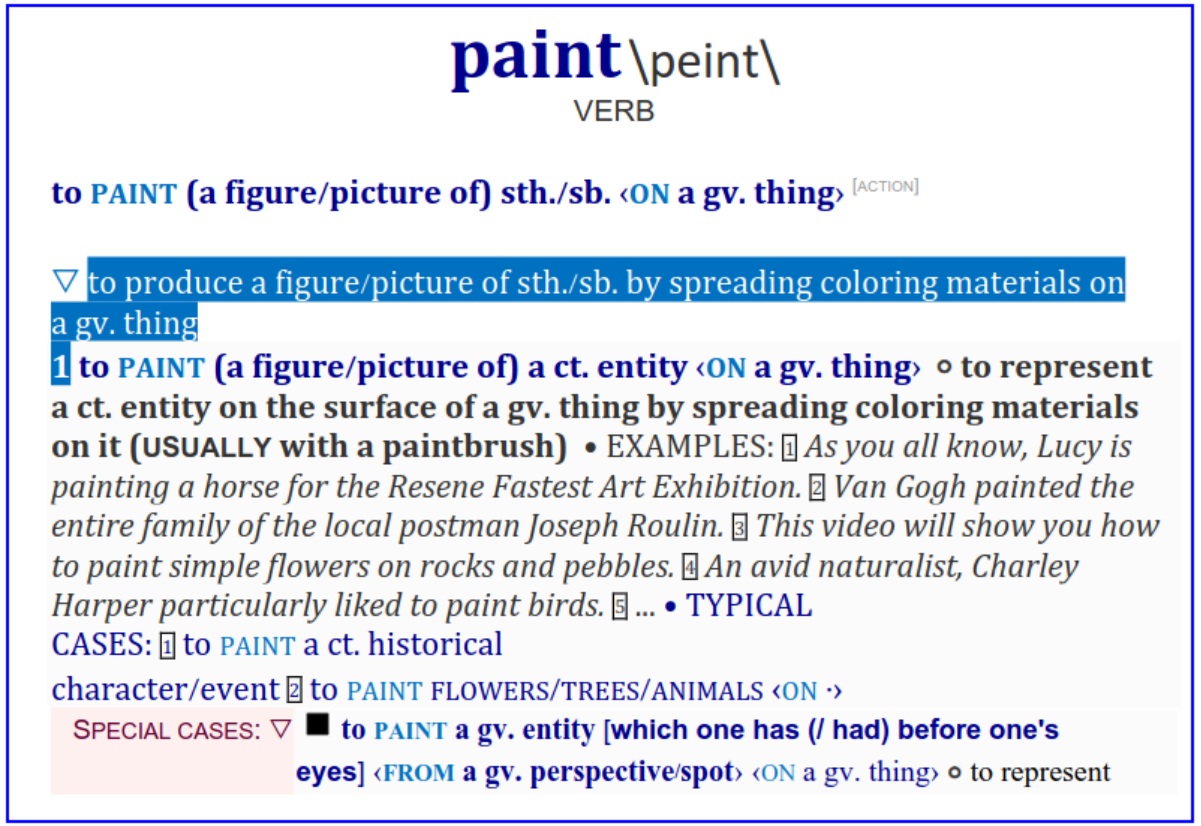
Can we use Zero-Shot Classification for semantically clustering BNC instances?



Finding PAD entries for “paint”

We set out from the already existing PAD entry for *paint*.

Did we miss in the PAD entry
an important sense variant or
important collocations?



paint \peint\
VERB

to PAINT (a figure/picture of) sth./sb. ‹ON a gv. thing› [ACTION]

▽ to produce a figure/picture of sth./sb. by spreading coloring materials on a gv. thing

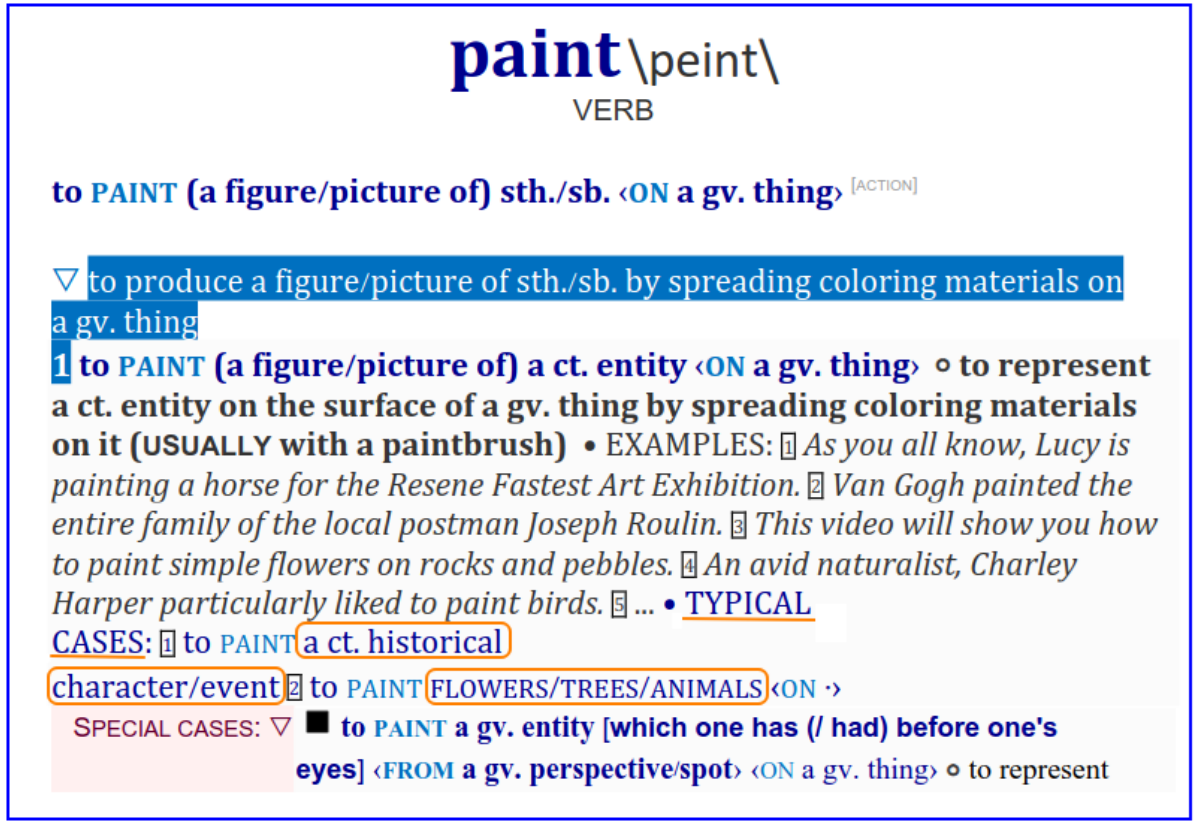
1 to PAINT (a figure/picture of) a ct. entity ‹ON a gv. thing› ∘ to represent a ct. entity on the surface of a gv. thing by spreading coloring materials on it (USUALLY with a paintbrush) • EXAMPLES: 1 *As you all know, Lucy is painting a horse for the Resene Fastest Art Exhibition.* 2 *Van Gogh painted the entire family of the local postman Joseph Roulin.* 3 *This video will show you how to paint simple flowers on rocks and pebbles.* 4 *An avid naturalist, Charley Harper particularly liked to paint birds.* 5 ... • TYPICAL CASES: 1 to PAINT a ct. historical character/event 2 to PAINT FLOWERS/TREES/ANIMALS ‹ON ·›

SPECIAL CASES: ▽ ■ **to PAINT a gv. entity [which one has (/ had) before one's eyes] ‹FROM a gv. perspective/spot› ‹ON a gv. thing› ∘ to represent**

PAD entry for “paint”

The PAD entry contains several “typical cases” from which we derive **49 single-word or multi-word collocates:**

historical character – historical event – flower – tree – animal – …



paint \peint\
VERB

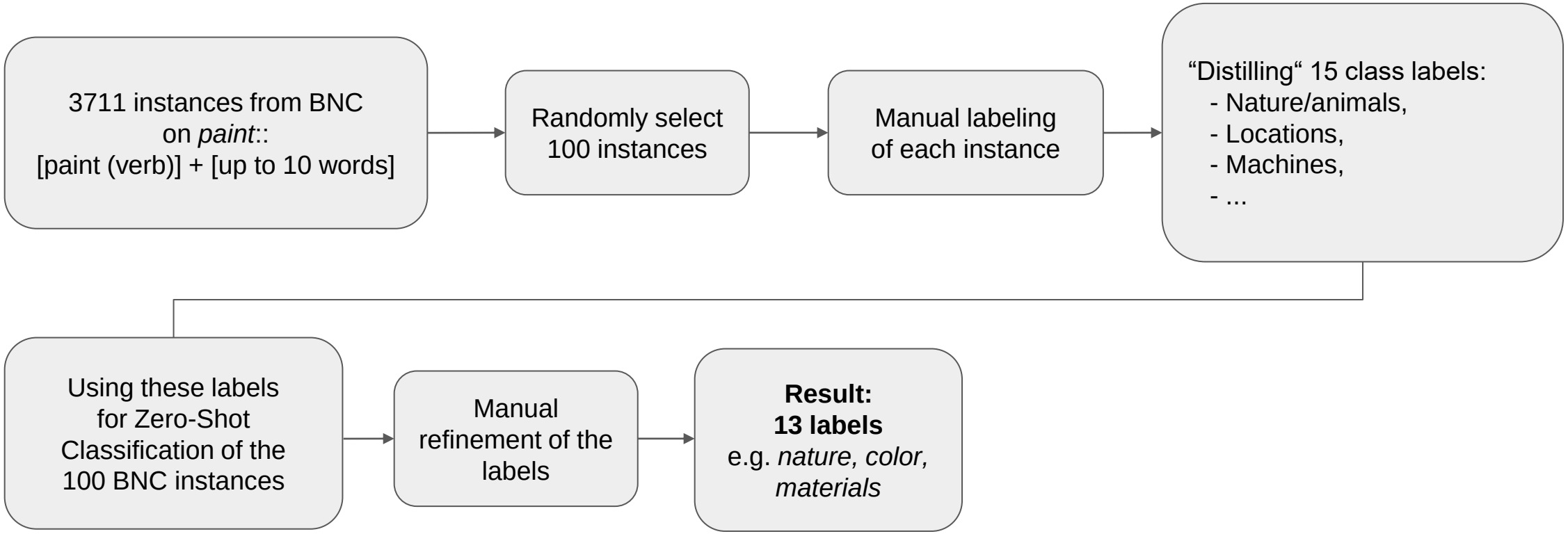
to PAINT (a figure/picture of) sth./sb. ‹ON a gv. thing› [ACTION]

▽ to produce a figure/picture of sth./sb. by spreading coloring materials on a gv. thing

1 **to PAINT (a figure/picture of) a ct. entity ‹ON a gv. thing› ∘ to represent a ct. entity on the surface of a gv. thing by spreading coloring materials on it (USUALLY with a paintbrush)** • EXAMPLES: 1 *As you all know, Lucy is painting a horse for the Resene Fastest Art Exhibition.* 2 *Van Gogh painted the entire family of the local postman Joseph Roulin.* 3 *This video will show you how to paint simple flowers on rocks and pebbles.* 4 *An avid naturalist, Charley Harper particularly liked to paint birds.* 5 … • TYPICAL CASES: 1 to PAINT a ct. historical character/event 2 to PAINT FLOWERS/TREES/ANIMALS ‹ON ·›

SPECIAL CASES: ▽ ■ **to PAINT a gv. entity [which one has (/ had) before one's eyes] ‹FROM a gv. perspective/spot› ‹ON** a gv. thing› ∘ to represent

Bootstrapping labels from *frequent* sense variants

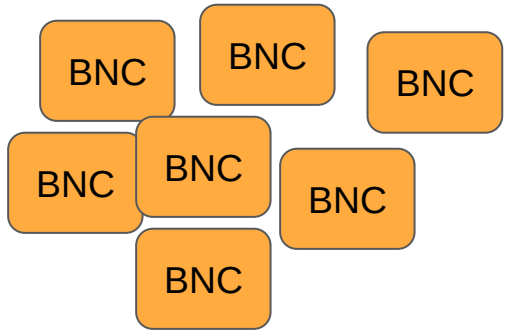


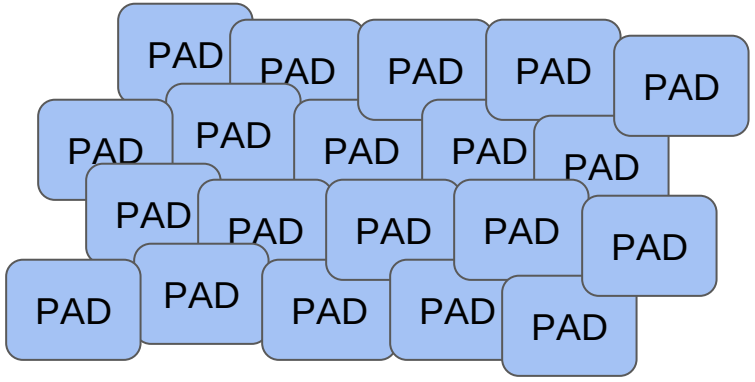
painted Van Gogh painting sunflowers	artist, design, natur, flower
painting most of the morning	time, daytime
painted with swirls of henna , like a Hindu bride	structure, design, shape, technique (bodyart)
paint erm in relation to this of course Mr	?
painted white in the early 1900	color, time, year
painted trucks , one of them with La Resurrección elaborately	car

Classification of all 3711 instances

- Third, we apply Zero-Shot Classification on the **3711 *paint* instances** from BNC.
- Each instance is labeled with the most probable category for an argument of *paint* – which either stems from the **13 BNC categories** or from the **49 PAD categories**.
- **First, we focus on the instances which are labeled with a BNC category** – as they don't seem to fit into the PAD examples, but are closer to the additional BNC examples.

Do we find additions to the PAD entry?





Some results

furniture	**vehicle**	**material**	**animal**	**nature**	**color**	**character**
0,287	0,224	0,223	0,219	0,202	0,198	0,173
29	27	149	50	3	443	2
painted furniture Derwent Upholstery Bistro Dining Range , Terranova Designs Interlubke bedroom	Paint Your Wagon	painted Perspex Aluminium tinfoil Formica Blackboard Newspaper Wallpaper Corrugated card Blotting paper	painted on rabbits ' ears , cancerous growths	painting a complete structure picture from nature , so fundamental a tenet	painted green , every toilet door , yellow	painted , a symbol
Painted dressers	painting a car again	Painted steel	painting a dog	paint landscape directly from nature	painted yellow , the colour of the bottle	painted on her scalp , like that of a wooden doll
painted dresser with tins	painted truck	painted Mm , a marble , a marble loaf , there you	paint a cricket	painted those faces of nature of	painted blue , and	
painted dining table and chairs	painted truck	Painting metal	painting cats		painted green	
painted dining room chairs	paint buses and taxis	Painted metal Bow chair Flooring	paint as naturally as possible with a mouse		painted green ,	

person	**surface**	**artist**	**picture**	**style**	**technique**
0,167	0,159	0,157	0,15	0,139	0,115
34	61	169	68	38	32
painted over , but someone	painted over the entire surface of the concrete	painted by Zoffany	paint a picture of an industry with a lot of technology on	painted red and a red and white gingham curtain hanging , cafe-style	painted in blocks of flat colour , cleverly
paints with passion , someone	painted on a plain surface	painted by Leon Bakst	painted picture of something like it in the Rockefeller Centre	paint in this particular style but his earlier work	painted in the same direct , rather violent technique
painted from a human mind	painted sign on the tarmac	painted by Stevenson	paint an example of	painted like the other statues , but still a popish image	paint-conserving method of
paint the room , but a woman in paint	painting the surface of the clean key cap with clear nail varnish	painted by Oswald , court painter to Charles IV	paints a picture of a moderately parallel MVS-run 390 with 50 to	painted in the style of Palmer , he	painted in the requisite shade by hand
painted along each arm , but she	painted surfaces of wood and tin ,	painted by Brandl	painted the picture black , because	painted in simple style	painting your toenails ,

Some results

The end of the ranking:
Words with no proximity to one of the 62 labels

artist	0,072	painted over and	bnc
artist	0,072	paint over it ,	bnc
artist	0,071	painted over the inheritance of his predecessor , but the Oak Leaf	bnc
graffiti	0,07	paint the gashes as soon as possible so that rust	2
person	0,069	painted my arsehole once	bnc
color	0,069	painted over ,	bnc
color	0,069	painted with two or more coats of matt black paint	bnc
surface	0,069	painted white again	bnc
artist	0,069	paint your being , Your body into them and out like blood	bnc
color	0,068	painted all round them , so that they	bnc
technique	0,068	painted twice up for reduction	bnc
style	0,067	paint over with a protective fungicidal paint such as Arbrex	bnc
artist	0,066	painted its waxiness	bnc
picture	0,066	paint thin thinners in this you	bnc
surface	0,065	painted on the other half , nothing	bnc
picture	0,065	paint a certain square amount of area Mm	bnc
badly	0,063	painted over in the Stalinist era	3
surface	0,063	painted on the bottom of the routes	bnc
material	0,063	painted on their roofs , so the helicopters	bnc
room	0,062	painted over the ocelli	7

Conclusions

- Our NLP methods (BERT and Zero-Shot Classification) are based on **semantic similarities of context words**.
- We find clusters with a **semantic and syntactic coherence**.
- The derivation of reading variants is **still a manual task**.
- **The more data we (consistently!) process**, the higher the likelihood of being able to use the processed data to automate the workflow:

  Clustering → integration of **manually** processed data → enhanced **automated** categorization of similar words (e.g. synonyms)

References

- Baisa, Vít, El Maarouf, Ismaïl, Rychlý, Pavel and Rambousek, Adam (2015). Software and data for corpus pattern analysis. In: *Proceedings of the Ninth Workshop on Recent Advances in Slavonic Natural Language Processing*. Brno, Czechia.
- DiMuccio-Failla, Paolo, Giacomini, Laura (2022): A proposed microstructure for a new kind of active learner's dictionary. In: *Lexicographica* 38(1), Berlin and Boston: De Gruyter. 475–499.
- Giacomini, Laura and DiMuccio-Failla, Paolo (2019). Investigating semi-automatic procedures in pattern-based lexicography. In: *Proceedings of the eLex 2019 conference. Electronic lexicography in the 21st century.* Sintra, Portugal.
- Giacomini, Laura, DiMuccio-Failla, Paolo and Lanzi, Eva. (2020). The interaction of argument structures and complex collocations: role and challenges for learner's lexicography. In: *Proceedings of the XIX EuraLex International Congress*. Alexandroupolis, Greece.
- Hanks, Patrick (2004). Corpus pattern analysis. In: *Proceedings of the XI EURALEX International Congress.* Lorient, France.
- Kliche, Fritz and Giacomini, Laura (2024, forthcoming): Exploring BERT's contextualized word embeddings: a suitable method for a lexicography-oriented analysis of argument structures? In: Giacomini, Laura and Piunno, Valentina (eds.): *Patterns of Meaning in Lexicography and Lexicology, Lexicographica Series Maior.* Berlin and Boston: De Gruyter.
- Yin, Wenpeng, Hay, Jamaal, and Roth, Dan (2019): Benchmarking zero-shot text classification: datasets, evaluation and entailment approach. 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)*. Hong Kong, China.