MUNI FI

Fantastic Examples and Where to Find Them

Compiling Czech Dataset for Evaluating Dictionary Examples

Michaela Denisová, Pavel Rychlý

449884@mail.muni.cz, pary@fi.muni.cz

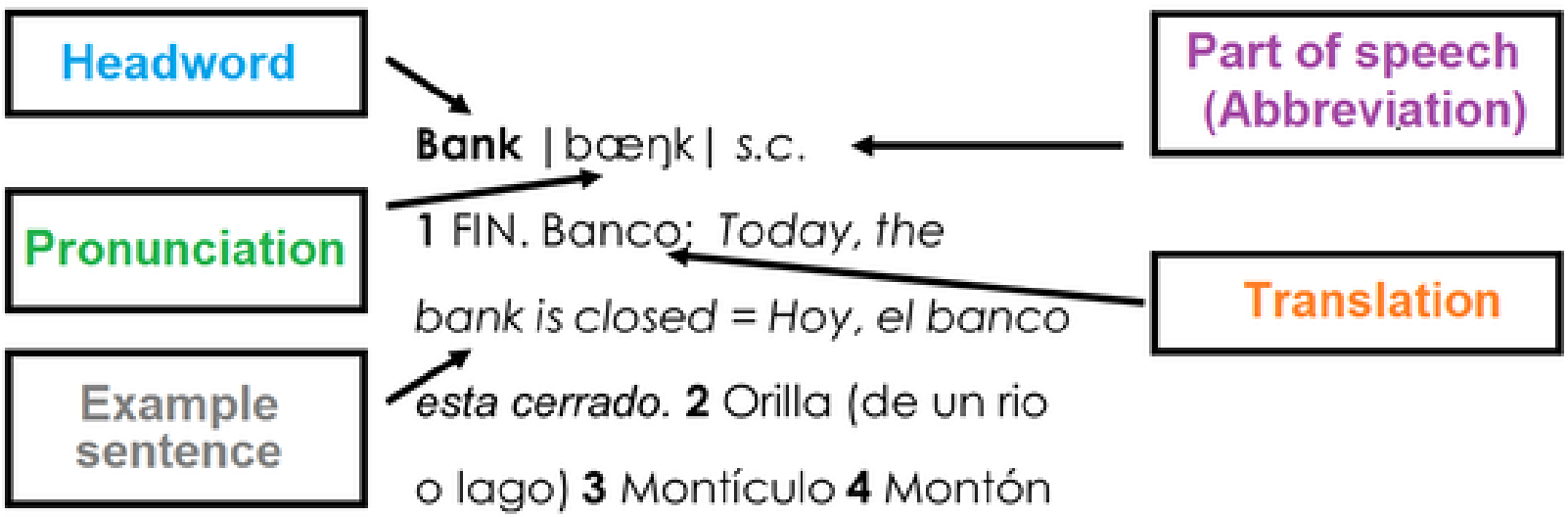
Natural Language Processing Centre

Faculty of Informatics, Masaryk University

December 6-8, 2024

Introduction

- Vital part of a dictionary entry
- Non-trivial and challenging task
- Flawed evaluation:
  - In practice unclear
  - Missing explanations



Motivation

- Compile a small dataset for Czech with clear explanations
- Fine-tune the annotation process
- Open a discussion about the evaluation and annotation
- First step towards the standardised evaluation and clearer guidelines

About Dictionary Examples

- Properties [1]:
  - Typicality
  - Naturalness
  - Informativeness
  - Intelligibility
- GDEX [2]
  - Rule-based tool part of Sketch Engine
  - Length, occurrences, part-of-speech, etc.
  - Adaptions to various languages: Slovene, Estonian, Portuguese, Dutch

About Dataset

- 40 examples for each of these 6 Czech words: *bandaska*, *lump*, *holdovat*, *očerňovat*, *svalnatý*, *demotivující*
- GDEX tool for selection
- Labels **a**, **b**, **c**, **d**

Table 1: Czech keywords used for compiling the evaluation dataset.

Word	Frequency	Part-of-speech	English
lump	9,876	noun	*rascal, crook*
bandaska	1,991	noun	*can, canister*
svalnatý	11,983	adjective	*muscular*
demotivující	1,965	adjective	*demotivating*
holdovat	11,906	verb	*take pleasure in, to wallow*
očerňovat	1,778	verb	*defame*

Annotations

D

- Sentence did not contain the word
- 5 sentences

Example

Heinrich **Lumpe** se narodil jako šesté ze dvanácti dětí obchodníka se dřevem.
Heinrich **Lumpe** was born the sixth of twelve children of a timber merchant.

Example

Přesto Proseč podala velice kvalitní výkony proti **Bandaskám** z Brodu, se kterými hrála ve skupině 2:2 a v semifinále 1:1.

Example

English: Nevertheless, Proseč delivered very strong performances against the **Bandasky** from Brod, with whom they played 2:2 in the group stage and 1:1 in the semifinals.

Annotations
c

- Sentence was incorrect, incomplete, in a different language than Czech, contained inappropriate language or emojis

Example

ně **demotivující** ;)

Example

Slovak: Západ je pripravený vyvolať svetový konflikt, kde masku svetového nepriateľa demokracie nasadili Putinovi tí najhorší **lumpi**, akých táto zem nosí.

Example

English: The West is ready to provoke a global conflict, where the mask of the world's enemy of democracy has been placed on Putin by the worst **scoundrels** this earth has ever borne.

Annotations

A and B

- *a*: appropriate as dictionary example
- *b*: sentence is correct but is not suitable as a dictionary example

Example

B: Pryč je původní malinká nádrž a místo ní má tenhle stroj pěknou **bandasku** na 17 litrů.
The original small tank is gone, and instead, this machine now has a nice 17-liter **canister**.

Example

A: Paní Jungová, Češka, vdova po padlém německém vojákovi, si otevřela mlékárnu s mlékem nalévaným do **bandasek**.
Mrs. Jung, a Czech woman and widow of a fallen German soldier, opened a milk shop where milk was poured into **cans**.

Example

B: Krk je dobré délky, čistý a **svalnatý**.
The neck is of good length, clean, and **muscular**.

Example

A: Trenéři navíc varují zejména ženy, že při fyzicky příliš náročné jízdě se jim vytvoří silná **svalnatá** stehna.
Trainers also warn, especially women, that overly strenuous cycling can result in the development of **muscular** thighs.

Guidelines

1. The keyword is central to the sentence, and the sentence captures its meaning well

Example

A: Je už jenom na vás, jakým oříškům dáte přednost, zda více **holdujete** vlašským nebo třeba lískovým.
It's entirely up to you which nuts you prefer, whether you **take more pleasure in** walnuts or perhaps hazelnuts.

Example

B: "Když mě někdo přepadne na ulici, určitě jich tolik nikdy nepřijede," rozčilovala se žena, jež marihuaně údajně **neholduje**.
When someone attacks me on the street, so many of them never show up," complained the woman, who allegedly **does not indulge** in marijuana.

Guidelines

2. The sentence should be clear and fitting

Example

A: Ta banda **lumpů** mu musí pěkně ležet v žaludku.
That gang of **crooks** must really be weighing on his mind.

Example

B: Moje pravé jméno je Aquila z Wenytry, ale doma mi říkají: Arčí, Arinko, zlato, draku, **lumpe**, obludo, malá, pipi, princezno… Slyším vlastně na všechno, ale pro pořádek jsem a vždycky budu ARINKA, přesněji řečeno Áji Arinka.
My real name is Aquila of Wenytra, but at home, they call me: Archy, Arinka, sweetheart, dragon, **rascal**, monster, little one, pipi, princess… I actually respond to anything, but for the record, I am and always will be ARINKA, more precisely Áji Arinka.

Guidelines

3. The sentence should not need more context to be understood, such as cultural knowledge, traditions, or history, or it should not reference something that is missing.

Example

Tím netvrdím, že Kájínek není **lump**.
By that, I'm not saying that Kájínek isn't a **crook**.

Guidelines

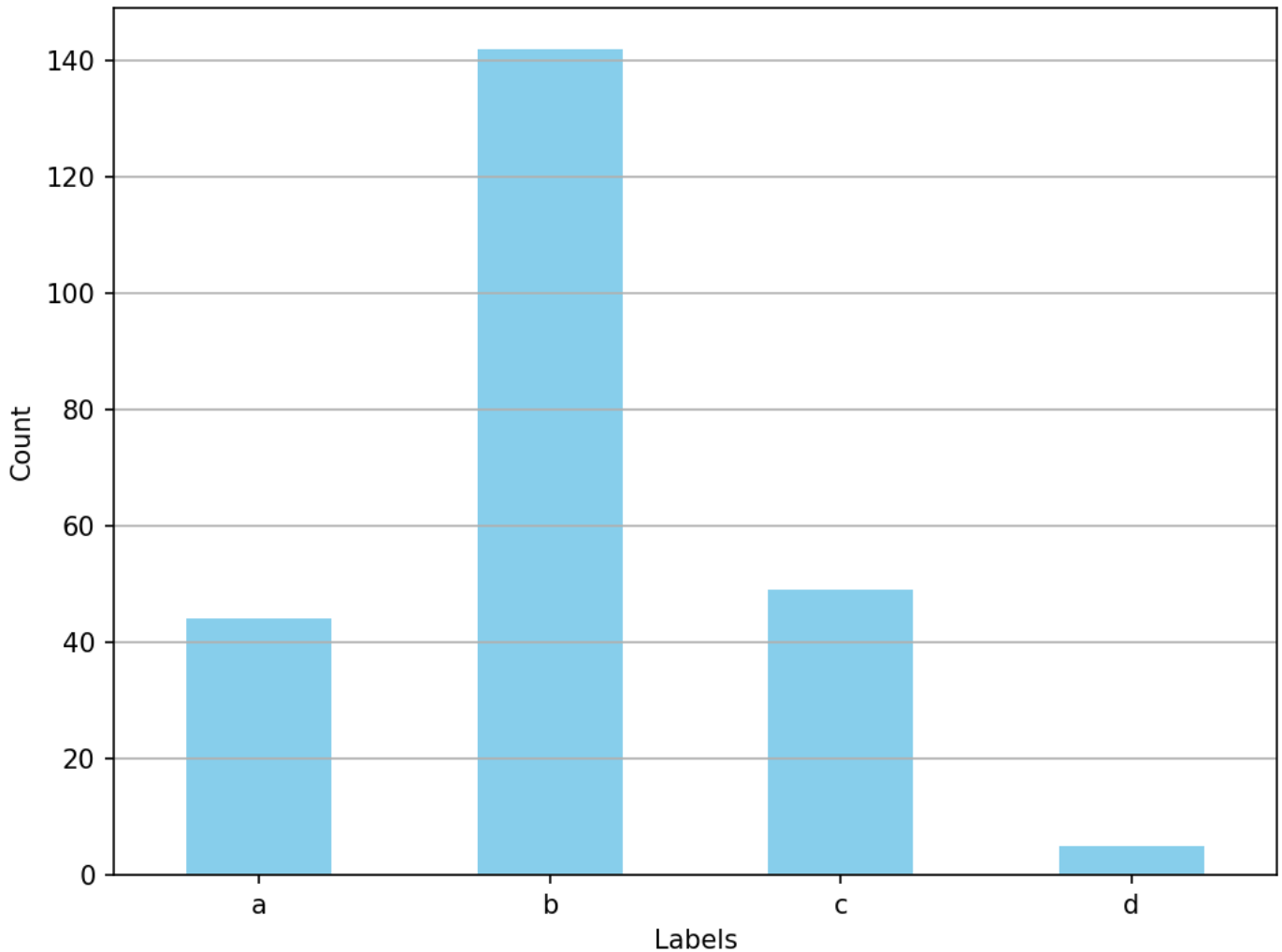
4. The sentence should not contain a controversial topic, such as PARSNIPs [1], irony, or abstract or symbolic meaning

Example

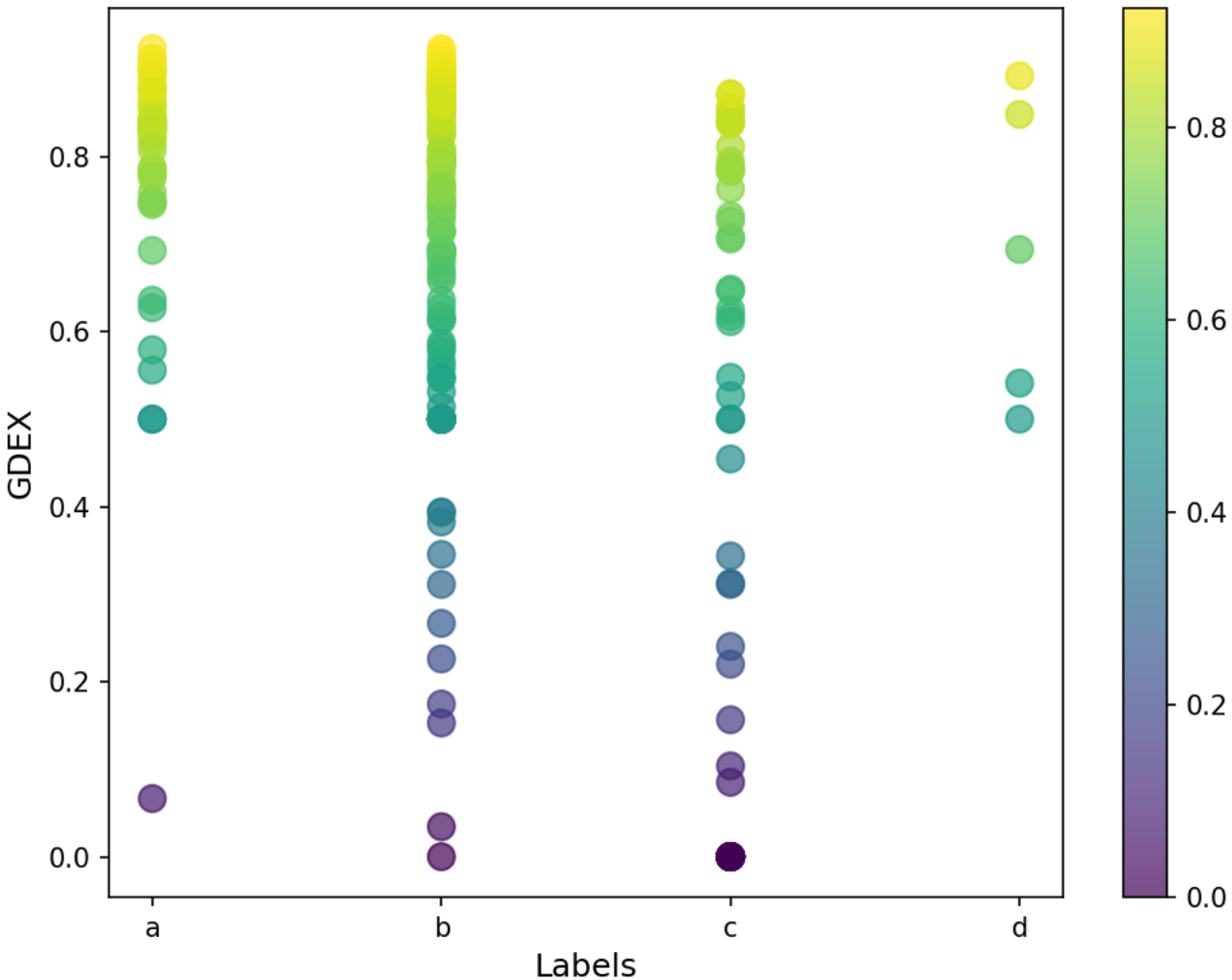
Otázka: Jak by se měli dívat věřící rodiče na to, kdy jejich děti **holdují** počítačovým hrám?
Question: How should religious parents view the fact that their children **indulge** in computer games?

[1] PARSNIPs stands for politics, alcohol, religion, sex, narcotics, -isms, and pork.

Label's counts



Correlation with GDEX



A with lowest GDEX score

Example

Kdo například neúměrně **holduje** pivu či kávě, musí počítat s tím, že se mu kolem očí objeví nehezké temné kruhy, zatímco věrnému konzumentovi ovocných šťáv nic podobného nehrozí.

Example

English: For example, anyone who excessively **indulges** in beer or coffee must expect unsightly dark circles to appear around their eyes, while a loyal consumer of fruit juices faces no such risk.

B and C with highest GDEX score

Example

B: Ráno jsem se tam tedy vybaven plechovou **bandaskou** po babičce vypravil podívat.
In the morning, I set out to take a look, equipped with my grandma's old tin **can**.

Example

C: Dludli to byl překlep - **Bandaska** nebo účel?
Dludli, was that a typo - **canister** or purpose?

Conclusion and Future Work

- Distinguishing between good and bad examples remains challenging and subjective
- Prepare more in-depth guidelines and inter-annotator agreements for the evaluation data complemented with explanations
- Extend the dataset

Bibliography I

[1] Beryl T. Sue Atkins and Michael Rundell. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press, 2008.

[2] Adam Kilgarriff et al. “GDEX: Automatically Finding Good Dictionary Examples in a Corpus”. In: *Proceedings of the 13th EURALEX International Congress*. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 2008, pp. 425–432.

Thank You for Your Attention!

MUNI
FACULTY
OF INFORMATICS