# Fantastic Examples and Where to Find Them
## Compiling Czech Dataset for Evaluating Dictionary Examples

Michaela Denisová and Pavel Rychlý

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
{449884,pary}@mail.muni.cz

**Abstract.** Examples are an important part of a dictionary entry, helping users better understand the word and its usage in context. However, selecting good examples is a challenging and time-consuming task due to varying selection criteria and the vast amount of data to choose from. While different tools have been developed to address this, evaluation remains flawed and lacks standardisation. In this paper, we compile an evaluation dataset for the Czech language, using the GDEX tool and manual annotations to classify examples and explain the classification. Based on our findings, we propose general annotation guidelines to improve consistency. This dataset serves as a foundation for the unified evaluation of dictionary example scoring tools and opens discussion on how to annotate examples. Additionally, we make the dataset publicly available. [1]

**Keywords:** Dictionary examples, GDEX, Evaluation.

## 1 Introduction

Examples are one of the most vital components of a dictionary entry since they help the user comprehend the word's frequent and common syntactic and collocative usage patterns. This supports the claim made by Atkins & Rundell (2008) [1] that sometimes, one cannot even understand the word without its example. Furthermore, the examples play an important role in language acquisition and learning. [4,9]

Finding an accurate and representative dictionary example is a non-trivial and time-consuming task. Firstly, because the criteria often vary and require adjustments depending on the target user or language. Secondly, contemporary corpora comprise immense amounts of text, which makes it challenging for lexicographers to efficiently sift through such large data and identify the most suitable examples. [8]

Over the years, various techniques have been proposed to alleviate these problems, ranging from rule-based approaches, such as the commonly used popular and important tool GDEX [2] [3], to machine learning methods [13],

[1] https://github.com/x-mia/czexample
[2] The acronym stands for **G**ood **D**ictionary **EX**amples.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 37–46, 2024. 

as well as hybrid models that combine both strategies. [11,5] Moreover, recent advances focus on utilising generative AI like ChatGPT. [12,9]

However, flawed and non-standardised evaluation is common across all these approaches. They are often evaluated against manual annotations made by lexicographers, lacking general guidelines on what constitutes a good or bad example in practice or explanations of why a given example was or was not selected. On top of that, there are no gold-standard evaluation datasets that would make the evaluations comparable between different approaches.

Therefore, in this paper, we compile an evaluation dataset to assess the good dictionary examples for the Czech language. Using the GDEX scores, we select 40 examples for each of these 6 Czech words: *bandaska*, *lump*, *holdovat*, *očerňovat*, *svalnatý*, and *demotivující*, manually annotate them, with clear explanations provided for why each example was classified as either good or bad, and analyse how our annotations correlate with the GDEX score. On top of that, we define general guidelines which arise from the annotations and our observations.

Our motivation is to take the first step towards the standardised evaluation of dictionary example scoring tools and compile an evaluation dataset for the Czech language. This dataset, along with the provided explanations and guidelines, aims to fine-tune the annotation process and open a discussion on how to annotate examples. Additionally, it serves as a stepping-stone towards a bigger gold-standard dataset and more transparent annotation practices. This enables the comparability across the models and correct interpretation of the results.

Our paper is structured as follows. In Section 2, we introduce the background of the good dictionary examples and the GDEX tool. In Section 3, we present our dataset, its compilation process, annotation and analysis, and the derived criteria. Finally, we offer concluding remarks and outline future work in Section 4.

## 2 Good Dictionary Examples

The function of the examples is to illustrate something already present in the dictionary (e.g., describe the use of a grammatical form) or to add new information about the entry. [1] We distinguish between two types of examples: authentic and invented. In contemporary lexicography, an authentic example is corpus-based and can be either fully authentic or adopted, i.e., when the lexicographer adjusts the sentence from the corpus. An invented example is crafted by a lexicographer. [5] The debate in lexicography over whether to use authentic or invented examples in dictionaries remains ongoing; however, in this paper, we exclusively focus on fully authentic examples.

In lexicography, the common properties of good dictionary examples are typicality, naturalness, informativeness, and intelligibility. [1,3,8,5]. Typicality means that the example represents the most frequent word forms, syntactic occurrences, collocations, etc., according to the corpus. Naturalness is a more subjective property, and it refers to how authentic the example feels; for example, it is likely to appear in the language usage.

Moreover, the example is informative when it helps the user clarify the use case and complements the definition. Finally, intelligibility indicates that the example does not contain complex constructions, specialised vocabulary, references or other expressions requiring further context to understand. [1,8]

### 2.1 GDEX

GDEX is the most popular rule-based tool in lexicography, which helps sort examples from a corpus based on the assigned score. It was implemented as a part of the Sketch Engine tool and was first used in the electronic version of the Macmillan English Dictionary. [14,3] The rules were established based on Atkins & Rundell (2008) [1] and brought into measurable features. [3]

Specifically, the GDEX tool consists of several classifiers that award or penalise sentences based on their features, resulting in a particular score. Then, the scores are sorted in descending order, which can be limited using a threshold. [8]

The features used to measure the GDEX score stated in Kilgarriff et al. (2008) [3] are the following: the length of the example sentence should be between 10 to 25 words, so the example would not be too long nor too short and incomprehensible; the words in the example sentence should each appear at least 17,000 times in the corpus; example should not contain proper nouns or demonstrative pronouns; example sentences with the strong collocation in the main clause were preferred; whole example sentences starting with a capital letter and ending with punctuation mark were preferred; example sentences that contain other collocations with high occurrence with the main collocation were preferred.

GDEX was further developed for various languages, such as Slovene [7,6], Dutch [15], Estonian [2], and academic Portuguese. [10] Each language-specific GDEX tool has its classifier settings tailored to the unique needs of that language. Among these settings were, for example, blacklisted words and characters, the penalty for sentences with too many words from different classes, awarding the words from a particular subcorpus, and many more. [8]

## 3 Dataset

We began by selecting Czech keywords, which would later be used to generate example sentences. For both tasks, we opted for Czech Web Corpus 2023 (cstenten23) incorporated into the Sketch Engine tool. [3] To ensure variety, we determined that two words were chosen from each of the three part-of-speech groups, i.e., nouns, adjectives, and verbs.

Furthermore, one word in each pair had to have a low frequency, occurring fewer than 2,000 times in the corpus, while the other was required to have a higher frequency, appearing between 9,000 and 12,000 times. Table 1 outlines the information about the selected keywords.

[3] https://www.sketchengine.eu/

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* |

Afterwards, we let the GDEX tool sort 300 lines of sentences from the cstenten23 corpus for each keyword from Table 1 according to the obtained score in descending order. From these 300 sorted sentences, we collected the top 40 with their GDEX scores for each keyword, resulting in 240 sentences in total for the entire dataset. Fig. 1 shows how the procedure looks like in the Sketch Engine application.

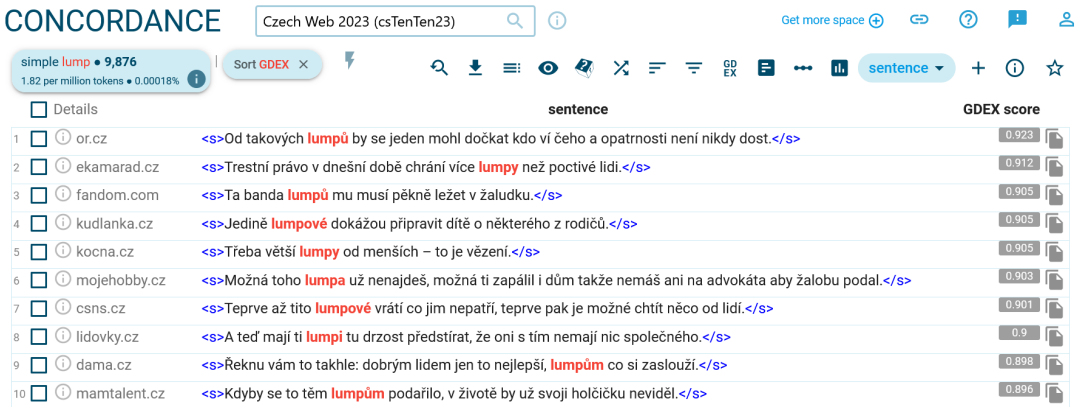


Fig. 1: The top ten dictionary examples of the Czech word *lump* with the highest GDEX score in Sketch Engine.

In the next phase, we manually annotated the sentences using labels: *a* when the sentence was correct in terms of content and grammar, and it was suitable as a dictionary example, *b* when the sentence was correct in terms of content and grammar, but it was not suitable as a dictionary example, *c* when the sentence was incorrect, incomplete, contained inappropriate language or emoticons, and *d* when the sentence did not contain the keyword. Fig. 2 visualises the number of sentences in each label-category.

Given Fig. 2, the sentences labelled as *b* formed the biggest group, while only five sentences obtained label *d*. Assigning label *d* was straightforward, as it applied only to sentences containing the keywords *lump* and *bandaska*, where the keyword was used as an orthographically identical proper noun, as shown in Examples 1 and 2.

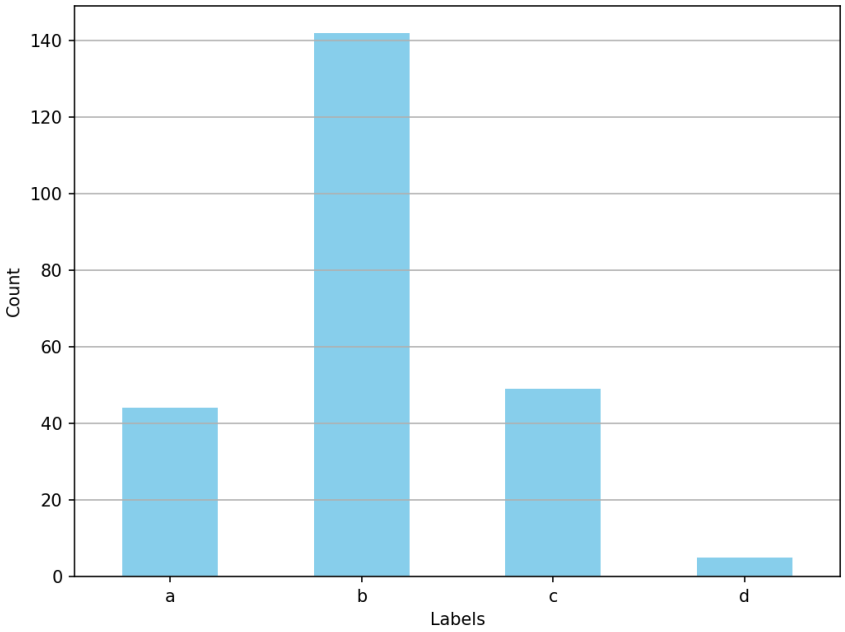


Fig. 2: The number of sentences with obtained labels *a-d*.

*Example 1.* Heinrich **Lumpe** se narodil jako šesté ze dvanácti dětí obchodníka se dřevem. (eng. *Heinrich **Lumpe** was born the sixth of twelve children of a timber merchant.*)

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

The same applied to label *c* as the sentences that obtained this label contained some type of error. Most of these errors were emoticons (see Example 3), grammatical errors (see Example 4), inappropriate language, or the sentence was incomplete (see Example 3) or in a different language (e.g., Slovak language, see Example 5).

*Example 3.* ně **demotivující** ;)

*Example 4.* Okamžitě zruště (*zrušte*) ten legislativní "paskvil" resp. z.č.361-je **demotivující** atd., advokát Janeček nevrhuje (*navrhuje*) jeho zrušení!!!!!!!!!

*Example 5.* Západ je pripravený vyvolať svetový konflikt, kde masku svetového nepriateľa demokracie nasadili Putinovi tí najhorší **lumpi**, akých táto zem nosí.

On the other hand, annotation using the labels *a* and *b* presented greater challenges due to a thin line between these two categories, where some sentences

were somewhat ambiguous. When looking at Example 6, we assigned to the sentence label *b* because the word order at the beginning sounds unnatural, and the sentence refers to an unknown machine. However, the keyword is central to the sentence, and its meaning is evident from the context.

*Example 6.* Pryč je původní malinká nádrž a místo ní má tenhle stroj pěknou **bandasku** na 17 litrů. (eng. *The original small tank is gone, and instead, this machine now has a nice 17-liter* ***canister****.*)

Moreover, some sentences which obtained label *a* would require minor post-editing before being suitable for use as dictionary examples, as in Examples 7 and 8, where removing the redundant particles or shortening the sentence would improve the quality and make it more suitable as an example for a dictionary. Also, in some cases, the sentences containing proper nouns, which are usually undesired, received label *a* (see Example 9).

*Example 7.* No a proto, že dotace skutečně nemohou být rozdělovány absolutně objektivně, jsou ve svém důsledku **demotivující**. (eng. *Well and that's why subsidies really cannot be distributed absolutely objectively, they are ultimately* ***demotivating****.*)

*Example 8.* Tradiční plechové **bandasky**, s nimiž naše prarodiče nebo rodiče chodívali v dětském věku pro mléko nebo třeba i vodu do studánky, mají dnes místo na kuchyňských poličkách jako retro ozdoba. (eng. *Traditional metal* ***cans****, which our grandparents or parents used to carry as children to get milk or even water from a spring, now have a place on kitchen shelves as retro decorations.*)

*Example 9.* Paní Jungová, Češka, vdova po padlém německém vojákovi, si otevřela mlékárnu s mlékem nalévaným do **bandasek**. (eng. *Mrs. Jungová, a Czech woman and widow of a fallen German soldier, opened a milk shop where milk was poured into* ***cans****.*)

Generally, the label *b* was assigned to the sentences which required more context and cultural knowledge or referred to something that was not a part of the sentence (see Example 10). In other cases, the keyword was not central to the sentence, or it would be difficult to understand the meaning from the context (e.g., from a language learner's perspective). When we compare Examples 11 and 12, we can see that the latter one is more descriptive, and it is more obvious what *muscular* means; therefore, it received label *a*.

*Example 10.* Tím netvrdím, že Kájínek není **lump**. (eng. *I'm not saying that Kájínek isn't a* ***crook****.*)

*Example 11.* Krk je dobré délky, čistý a **svalnatý**. (eng. *The neck is of good length, clean, and* ***muscular****.*)

*Example 12.* Trenéři navíc varují zejména ženy, že při fyzicky příliš náročné jízdě se jim vytvoří silná **svalnatá** stehna. (eng. *Trainers also warn, especially women, that overly strenuous cycling can result in the development of* ***muscular*** *thighs.*)

In conclusion, several criteria for assigning labels *a* or *b* arose from our observations. These are outlined in the following section.

### 3.1 Guidelines

The guidelines for annotating dictionary examples are:

1. The keyword is central to the sentence, and the sentence captures its meaning well (see Example 13 vs Example 14)

   *Example 13.* Je už jenom na vás, jakým oříškům dáte přednost, zda více **holdujete** vlašským nebo třeba lískovým. (eng. *It's entirely up to you which nuts you prefer, whether you **take more pleasure in** walnuts or perhaps hazelnuts.*)

   *Example 14.* Když mě někdo přepadne na ulici, určitě jich tolik nikdy nepřijede," rozčilovala se žena, jež marihuaně údajně **neholduje**. (eng. *When someone attacks me on the street, so many of them never show up," complained the woman, who allegedly **does not indulge** in marijuana.*)

2. The sentence should be clear and fitting (see Example 15 vs Example 16).

   *Example 15.* Ta banda **lumpů** mu musí pěkně ležet v žaludku. (eng. *That gang of **crooks** must really be weighing on his mind.*)

   *Example 16.* 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. (eng. *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.*)

3. The sentence should be simple and not contain complicated sentence constructions.
4. 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 (see Examples 10, 11).
5. The sentence should not contain demonstrative pronouns (e.g., *that*, *this*, *these*, *those*, etc.) or numbers.
6. The sentence is whole, starting with a capital letter and ending with a dot; it should not begin with a subordinate clause or contain direct speech or three dots.
7. The sentence should not contain grammatical errors, foreign words, abbreviations, emoticons, or inappropriate language, such as vulgarism, racist or sexual content.
8. The sentence should not contain a controversial topic, such as PARSNIPs [4], subliminal meaning, irony, or have abstract or symbolic meaning (see Example 17).

   *Example 17.* Otázka: Jak by se měli dívat věřící rodiče na to, kdy jejich děti **holdují** počítačovým hrám? (eng. *Question: How should religious parents view the fact that their children **indulge** in computer games?*)

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

### 3.2 Correlation with GDEX

In this section, we analyse how our annotations correspond with the scores assigned by the GDEX tool. Fig. 3 shows the distribution of the GDEX scores across the label categories.

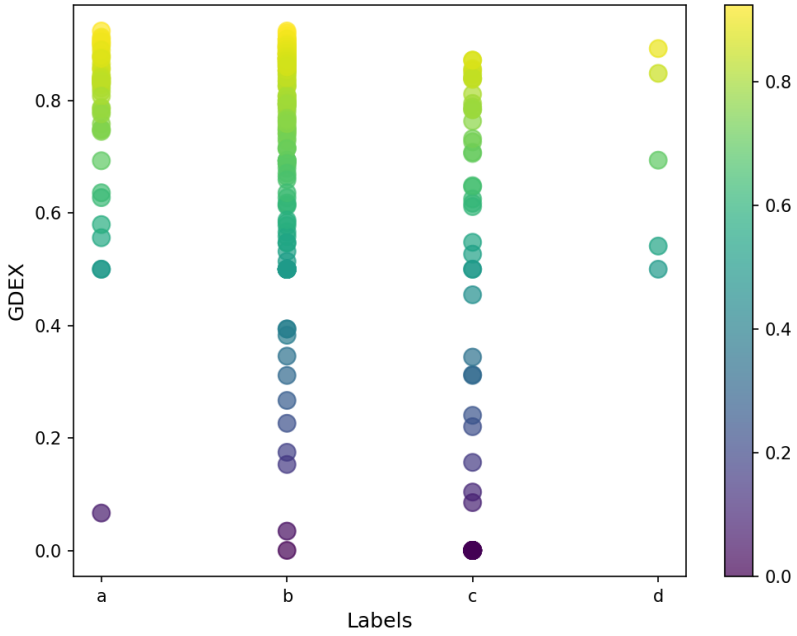


Fig. 3: The distributions of the GDEX scores across the labels.

Given Fig. 3, we can see that the sentences labelled as *a* tend to have their GDEX scores higher, except Example 18. Although this example was classified as *a*, it appears to be ambiguous; while somewhat lengthy, it captures the meaning of the keyword well.

*Example 18.* 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. (eng. *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.*)

Moreover, the scores with the labels *b* and *c* were evenly spread out through the whole range. Example 19 shows the sentence labelled as *b* with the highest GDEX score. The sentence is missing some information, and the keyword's meaning is unclear from the context. The sentence labelled as *c* with the highest GDEX score was nonsensical (see Example 20).

*Example 19.* Ráno jsem se tam tedy vybaven plechovou **bandaskou** po babičce vypravil podívat. (eng. *In the morning, I set out to take a look, equipped with my grandma's old tin **can**.*)

*Example 20.* Dludli to byl překlep - **Bandaska** nebo účel? (eng. *Dludli, was that a typo - **canister** or purpose?*)

## 4 Conclusion and Future Work

In this paper, we have introduced the task of selecting good dictionary examples. We have compiled an evaluation dataset for the Czech language using six diverse words and complemented it with manual annotations and explanations. We have discussed how the annotations correlate with the obtained GDEX scores. On top of that, we thoroughly analysed the examples and derived several selection guidelines, which, together with the compiled dataset, present a first step towards a unified evaluation of good dictionary examples.

Our research suggests that despite the detailed criteria, distinguishing between good and bad examples remains challenging and subjective, as many fall within a grey area. Moreover, our analysis revealed gaps in the GDEX scoring system. We propose that future work prepare more in-depth guidelines and inter-annotator agreements for the evaluation data complemented with explanations of why a given example was annotated as good or bad and explore different scoring alternatives. On top of that, we plan on extending the dataset.

**Acknowledgements.** This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062.

## References

1. Atkins, B.T.S., Rundell, M.: The Oxford Guide to Practical Lexicography. Oxford University Press, New York (2008)
2. Kallas, J., Kilgarriff, A., Koppel, K., Kudritski, E., Langemets, M., Michelfeit, J., Tuulik, M., Viks, U.: Automatic generation of the Estonian collocations dictionary database. In: Proceedings of the eLex 2015 conference. pp. 1–20. Electronic lexicography in the 21st century (2015)
3. Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P.: GDEX: Automatically finding good dictionary examples in a corpus. In: Proceedings of the 13th EURALEX International Congress. pp. 425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra (2008)
4. Kilgarriff, A., Marcowitz, F., Smith, S., Thomas, J.: Corpora and language learning with the Sketch Engine and SKELL. Revue française de linguistique appliquée **XX(1)**, 61–80 (2015). https://doi.org/10.3917/rfla.201.0061
5. Koppel, K.: Näitelausete korpuspõhine automaattuvastus eesti keele õppesõnastikele. Ph.D. thesis, Tartu University (2020)

6. Kosem, I., Gantar, P., Krek, S.: Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. In: Proceedings of eLex 2013 conference. Electronic lexicography in the 21st century (2013)
7. Kosem, I., Husák, M., McCarthy, D.: GDEX for Slovene. In: Proceedings of eLex 2011 conference. pp. 151–159. Electronic lexicography in the 21st century (2011)
8. Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J., Tiberius, C.: Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. International Journal of Lexicography **32**(2), 119–137 (2018). https://doi.org/10.1093/ijl/ecy014
9. Kosem, I., Kuhn-Zingano, T., Arhar-Holdt, S., Koppel, K., Tiberius, C., Zviel-Girshin, R., Wasznik, V., Zgaga, K.: Can AI assist in selecting dictionary examples? A case study in four languages. In: Book of Abstracts of the XXI EURALEX International Congress. pp. 128–130. XXI EURALEX International Congress (2024)
10. Kuhn, T.Z.: A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students. Ph.D. thesis, University of Lisbon (2017)
11. Lemnitzer, L., Pölitz, C., Didakowski, J., Geyken, A.: Combining a rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German. In: Proceedings of the eLex 2015 conference. pp. 21–31. Electronic lexicography in the 21st century (2015)
12. Lew, R.: ChatGPT as a COBUILD lexicographer. Humanities and Social Sciences Communications **10**(704) (2023). https://doi.org/10.1057/s41599-023-02119-6
13. Ljubešić, N., Peronja, M.: Predicting corpus example quality via supervised machine learning. In: Proceedings of the eLex 2015 conference. pp. 427–442. Electronic lexicography in the 21st century (2015)
14. Rundell, M.: Macmillan English Dictionary for Advanced Learners. Macmillan, Oxford, 2nd edn. (2002,2007)
15. Tanneke Schoonheim, R.T.: Dutch lexicography in progress: the Algemeen Nederlands Woordenboek (anw). In: Proceedings of the 14th EURALEX International Congress. pp. 718–725. Fryske Akademy (2010)