Identifying Metadata-Specific Collocations in Text Corpora

Ondřej Herman[1,2], **Miloš Jakubíček**[1,2], **Vojtěch Kovář**[1,2]

[1]Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Brno, Czech Republic
E-mail: xherman1@fi.muni.cz, jak@fi.muni.cz, xkovar3@fi.muni.cz

[2]Lexical Computing
Brno, Czech Republic
E-mail: ondrej.herman@sketchengine.eu, milos.jakubicek@sketchengine.eu, vojtech.kovar@sketchengine.eu

Abstract

Statistical corpus analysis of collocations is one of the important steps in creating a dictionary entry: collocations may distinguish senses, describe typical phrasemes and idioms and outline the whole picture of a word's behaviour. However, some collocations are domain-specific, typical only in particular contexts, and thus far there has been no easy way to distinguish "general" collocations from those that are predominantly typical in particular domains. In this paper, we present a tool which allows lexicographers to see typical domains in which a particular collocation occurs. We introduce a statistical procedure based on corpus metadata to identify domain-specific collocations in an intuitive way, and we also present a user interface connected to the word sketch feature of the Sketch Engine corpus interface (Kilgarriff et al., 2014a).
The new feature can be used in the manual inspection of collocation lists, as well as when using the API or in a semi-automatic post-editing scenario of building a dictionary.

Keywords: collocations; word sketch; meta-data; text types; corpus

1. Introduction

Word sketches (Kilgarriff et al., 2014a) are an intuitive and intelligible summary of a word's collocational behaviour; they have been used in lexicography for nearly 20 years. However, additional information for some of the collocations is sometimes needed.

One of the missing pieces of information is whether a particular collocation is evenly distributed within the corpus, or somehow specific to a particular text type, or even found exclusively in a particular text type. By text type, we understand any type of metadata annotation available within the corpus: web domain, genre, topic, year of publication, author of the text, etc.

This paper addresses the possibilities of adding text type information into lists of collocations such as word sketches. After a discussion of various possible approaches, we select two types of information that may be beneficial for users and show how it can be presented to the users in the Sketch Engine interface and in the API.

We also describe the practical implementation of this new feature within Sketch Engine and discuss some particular advantages and potential problems. Finally, we introduce the compilation of new word sketch indexes that enable this feature and briefly discuss its efficiency.

2. Related Work

Corpus meta-data, as well as collocations, have been used in countless projects and it would make no sense to try to list them all. For example, (Sharoff et al., 2014) used

log-likelihood statistics to extract candidates for multiword dictionary entries. The Word sketch itself, with its default *logDice* score (Rychlý, 2008), has been intensively used since its introduction in 2004 (Kilgarriff et al., 2014a).

Corpus meta-data information has also been used widely. Corpora and subcorpora of different domains have been compared (Kilgarriff, 2009; Kilgarriff et al., 2014b) to obtain domain-specific headword lists suitable for specialised dictionaries, and the automatic generation of dictionary labels using corpus meta-data has been proposed and implemented (Rundell & Kilgarriff, 2011).[1] However, all of this has only been suggested on the word (or term) level. Similar computations have, to the best of our knowledge, never been suggested on the level of collocations, which is what we propose in this paper. The statistics for collocations need to be different from single-word meta-data usage, as the expected usage will be different – we do not need a list of most domain-specific collocations, but we do need to mark all collocations that are likely to be domain-specific.

2.1 Meta-Data and Collocations

To the best of our knowledge, there is no corpus tool capable of adding meta-data information into lists of collocations. However, the statistics presented in the folling sections more or less just play with relative frequencies within particular text types, and specify conditions under which observation of these relative frequencies is interesting.

Of course, finding the frequency distribution of a given collocation across text types was possible before: for example in Sketch Engine it was possible to create a concordance for a specific item in word sketch, and to create a text type frequency distribution for this collocation that contains relative frequencies in particular text types, as illustrated in Figure 1. In that case it reveals that "oil spill" is more than 3x more frequent in *W_misc* and *W_non_ac_polit_law_edu*, than in the rest of the corpus – which may be an interesting item of information.

However, this process is very time-consuming and we cannot expect anyone to investigate such a frequency distribution for all collocations in a word sketch. Instead, we let the computer do it, and we set conditions under which a collocation is highlighted as specific for a particular text type. That will give lexicographers easy access to information they probably did not access previously.

3. The Evolution of the Idea

In the following text, let us think about a particular collocation **C** (e.g. *good news*), and a particular text type **T** (e.g. *genre: newspaper*). Let us suppose that **C** occurs **N** times in the whole corpus, and **M** times in the text type **T**.

3.1 Initial Idea

We started with a very rough simple idea: if a substantial majority of collocation **C** occurs in **T**, we should report it to the user. For example, if 70% of **C** falls into **T** (or $M/N \geq 0.7$),

[1] However, the automation of dictionary labels does not seem to be intensively used, perhaps due to the lack of useful corpus meta-data, no clear general conception of dictionary labels, or the low accessibility of the related features in the corpus tools.

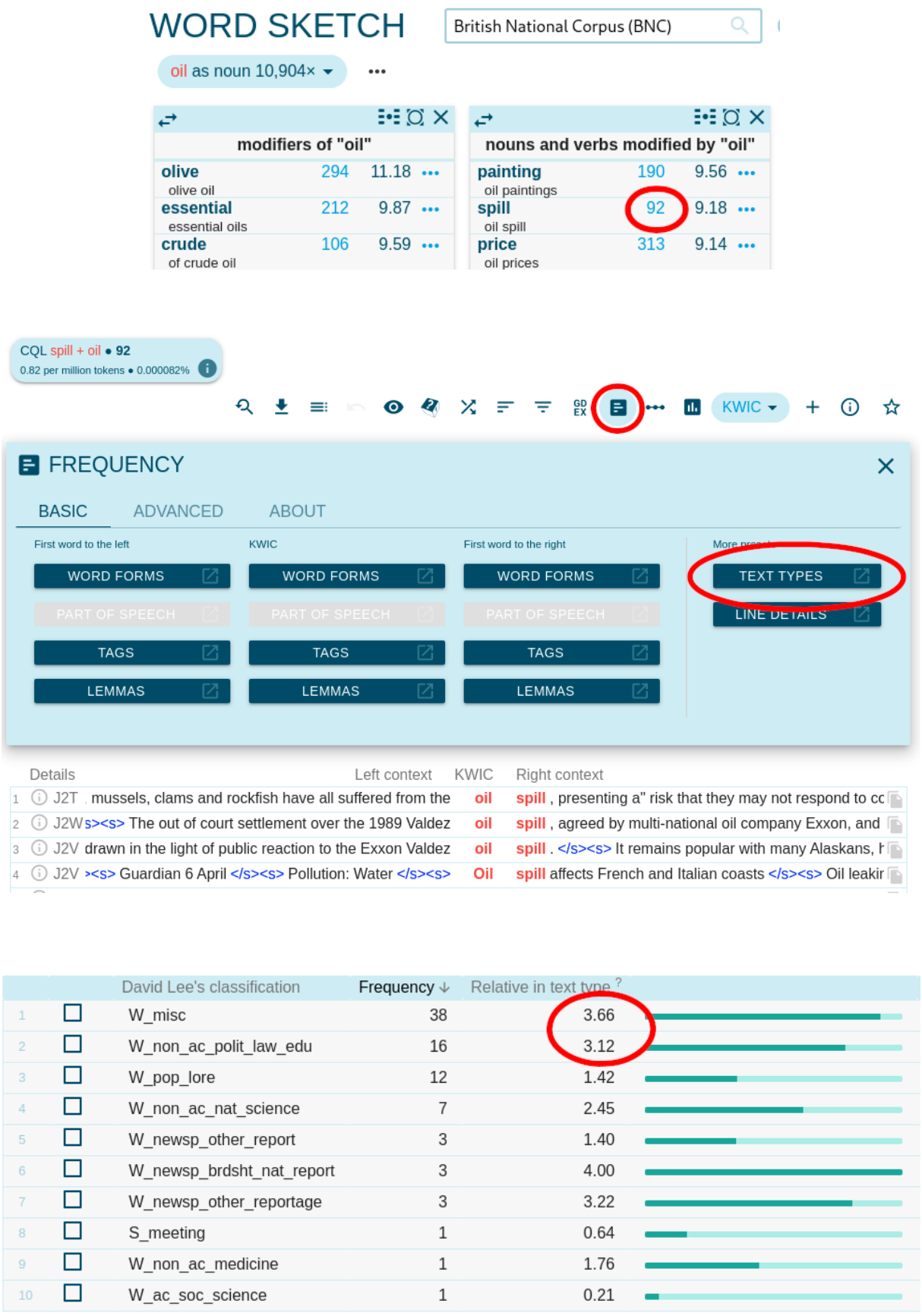

Figure 1: Finding the relative meta-data frequencies of a collocation.

we would say that **C** "usually occurs in" **T**. Or, if 99% of **C** belongs to **T** ($M/N \geq 0.99$), we would tell the user that **C** "only occurs in" **T**. Actually, this method has been built into Sketch Engine for years, it was just not directly visible in the interface.

However, there are significant problems with this simple approach.

It would work well if all of the text types in the corpus were the same size. But if **T** covers a substantial part of the corpus — e.g. 90%, like *Publication date: 1985-1993* in the British National Corpus, BNC (Leech, 1992) — then it is absolutely normal and expectable that the majority of the occurrences of **C** will fall into this text type. The vast majority of all the collocations would probably exceed some 70% threshold and we would report that almost all the collocations "usually occur in **T**". Such information is more or less useless.

On the other hand, if, e.g. half of the occurrences of **C** fall into a small text type (such as *Publication date: 1960-1974* in the BNC, covering only 1.2% of the corpus), it is definitely something interesting and users will want to know. However, our simple method would miss it.

3.2 Including the Text Type Size

It is clear that we need to include the text type size into the computation. Let us suppose that text type **T** covers **P** percent of the corpus text.

As the naive approach from the previous section works well if all the text types are the same size, we thought about a statistical correction that would use a weighting of the occurrences within particular text types, in order to virtually make all of them the same size. We normalised the raw number of hits using the percentage of the corpus covered by the text type, and compared these normalised numbers with their sum. In other words, we used **M/P** for all the text types instead of **M**, and the sum of all these fractions instead of **N**. Let us call this sum $\mathbf{N}_{corrected}$.

This approach, however, is problematic in another set of cases, as we noticed shortly. If **T** is small (such as regarding the *Publication date: 1960-1974* in the BNC, $P = 1.2\%$), the normalisation will end up with an unwanted result: imagine two text types **T1** and **T2**, the first covering 99% of the text and containing 45 out of 50 occurrences of **C**. Then $P1 = 99\%, P2 = 1\%, M1 = 45, M2 = 5$. The normalised frequencies are $M1/P1 = 45, M2/P2 = 500$. $N_{corrected} = 545$, so **T2** contains $500/545 = 92\%$ of the corrected occurrences and we would report that **C** "usually occurs in **T2**". But this does not correspond to the real distribution; **T2** contains only 5 of 50 occurrences and "**C** usually occurs in **T2**" is very misleading information.

Another problematic case is when we have two small text types, **T1** and **T2**, both covering e.g. 5% of the corpus ($P1 = P2 = 5\%$). Collocation **C** occurs in both of them with the same frequency (e.g. 30), and never outside these two text types — i.e. $M1 = M2 = 30, N = 60$. Then $M1/P1 = M2/P2 = 30/0.05 = 600, N_{corrected} = 1{,}200$. Neither of the two text types will be mentioned because the corrected ratio for both of them is 50%, which will not exceed the threshold. We will not say anything but that the initial situation is very interesting — **C** only occurs in 10% of the corpus! — so not saying anything is clearly wrong.

3.3 Expected vs. Observed

The last mentioned situation made us rethink the idea of saying "usually in **T**" or "only in **T**": sometimes we have two or more significant text types to report, and none of these messages describes the situation correctly. We came to the conclusion that, in specified cases, we need to say "especially in **T**" which would mean that the collocation is *more often found in this text type than in the others.*

What does this mean? To avoid the problematic results mentioned in the previous section, we used the concept of *expected* and *observed* occurrences of collocation **C**. The expected number of occurrences means, how many hits we would expect in this text type, according to the number of hits in the whole corpus. In other words, $M_{exp} = N * P$. Then we contrast this number with the *observed* **M**. If the observed **M** is significantly higher, we would say "**C** occurs especially in **T**".

3.4 Statistical Significance

Significantly higher in the previous sentence should definitely incorporate statistical significance. For our purposes, however, it is crucial that the information provided to users can be explained easily. And in pure hypothesis testing, we usually do not get easily explainable numbers: How to communicate to the user that e.g. an increase 1,000→1,100 (i.e. 10%) is statistically significant, whereas 40→60 (i.e. 50%) may not be? Especially when we only want to provide an extremely simple message "**C** occurs especially in **T**" – we want users to have some clear idea behind this message.

In addition to that, it has recently been argued (Kilgarriff, 2005; Koplenig, 2019) that statistical significance is not the right measure in corpus linguistics, because

- language is not random and therefore does not fulfil the assumptions of statistical hypotheses testing,
- therefore, if we have enough data, almost everything becomes statistically significant,
- therefore measuring statistical significance means only measuring if we have enough data, and it is not a good base for estimating what is linguistically interesting.

For these two reasons, we decided to employ a simple, explainable criterion: if observed **M** is at least twice as big as the expected $\mathbf{M}_{exp}$, we will show that "**C** occurs especially in **T**". To avoid reporting random noise, we added the following thresholds that must be met in order to display the message:

- the minimum total frequency of the collocation (**N**) is 20
- the minimum $\mathbf{M}_{exp}$ is 5

The minimum thresholds still ensure statistical significance with $p < 0.05$, using the binomial test.

3.5 Usually and Only

In the previous two sections, we specified some notable criteria and decided to mark them by telling the user "**C** occurs especially in **T**". However, we did not abandon the idea of marking "usually" and "only" along with "especially". We just returned back to their original, naive meaning.

For "usually" and "only", we use absolute frequencies, the uncorrected number of hits, to ensure that the words really mean the same to the system and to the user. If absolute frequency in text type **T** stands for more than 70% of the occurrences of the collocation's overall frequency, we indicate "**C** occurs usually in **T**". If it is more than 97%, we show "**C** occurs only in **T**". (These two thresholds are arbitrary, as agreed with initial users of this new feature.)

However, we will show the message under this condition only if **T** is not a dominant text type, i.e. only if it covers less than 50% of the corpus – this is to avoid the problematic scenario with *Publication date: 1985-1993* described above. For dominant text types (covering more than 50% of the corpus), we can still show "usually" and "only" but the conditions are different:

- absolute frequency in text type **T** stands for more than 70% (97%) of the occurrences of the collocation's overall frequency,
- the minimum expected frequency $\mathbf{M}_{exp}$ in the rest of the corpus is 20,
- the observed frequency in the rest of the corpus is less than 20% of $\mathbf{M}_{exp}$.

In other words, we report "usually" and "only" for the dominant text type only if the frequency in the rest of the corpus is much lower than expected.

4. Specification

In less detail, we want to inform word sketch users about three types of the collocation's specificity:

1. The collocation is *only* present in a particular text type, and (nearly) not at all in the others. We show "only **T**" if more than 97% of the collocation's occurrences (in absolute numbers) falls into text type **T**.
2. Most of the collocation occurrences fall into a particular text type, i.e. the text type is dominant for the collocation but not for the whole corpus. We show "usually **T**" if more than 70% (but less than 97%) of the collocation's occurrences falls into text type **T**. (There are separate rules for the dominant text type, see the previous section.)
3. The relative frequency of a particular collocation in a particular text type is much higher than the relative frequency of that collocation in the whole corpus. We show "especially **T**" if the collocation's *relative* frequency in text type **T** is at least twice as high as its relative frequency in the whole corpus.

These three characteristics are now part of the word sketch interface, if compiled. We describe the compilation procedure and the user interface in the following sections.

5. Implementation

5.1 Compilation

The statistics are computed at the time of corpus compilation and are instantly available in the word sketch database indexes. To save the numbers for each collocation, we had to change the format of the word sketch indexes. The resulting data are slightly larger, for the BNC with 3 different text types ("Text type", "Publication date" and "David Lee's classification") the increase was 22% (1.03GB→1.25GB). The additional compilation time was 13 minutes.

Of course, these numbers depend on various details (sketch grammar, the number of text types included, the distribution of text types within the corpus etc.) and cannot be generalised; they are rather illustrative.

The compilation program is written in the Go programming language.

5.2 User Interface

The notes "only", "usually", and "especially" are displayed in the standard word sketch interface under the particular collocations. Depending on the sketch grammar, the number of text types and their distribution in the corpus, they can take up a lot of space on user's screen – therefore they can be turned off. We have also considered an option where they are displayed on mouseover or after clicking a small icon, but this is so far only a matter for future development.

Another idea for future development is the option to filter the word sketch by the metadata labels, or by *always/usually/especially*. This is likely to appear in the interface soon.

The notes are also available in the Sketch Engine REST API, so that external tools can benefit from this new feature.

6. Lexicographic Potential

Of course, the new feature can be used in lexicographical work – the text types in the corpus may provide useful insights leading to dictionary labels for particular collocations, or even for whole entries:

- **Revealing metadata-specific senses.** Collocations are often used to describe different senses of the headword. If we notify the lexicographer that a particular collocation is domain-specific, it may lead to a useful dictionary label for the particular sense (e.g. *American English* or *legal texts*, depending on the available meta-data).
- **Richer information on collocations.** Dictionaries often include typical collocations and examples of the headword. Now it is easy to add more information to these particular collocations, e.g. *black hole (astronomy)*.
- **Pre-generating label candidates.** In post-editing lexicography, which is becoming increasingly popular, it can be used directly for suggesting the labels. The collocations can be exported from the corpus into a dictionary writing system, together with the meta-data information, and a lexicographer can only edit the collocations and the labels – which will result in richer dictionaries with less work.

7. Examples

Figure 2 shows two examples of metadata-specific collocations, as can be newly identified in word sketches. Both examples use the British National Corpus and David Lee's classification (Lee, 2002).

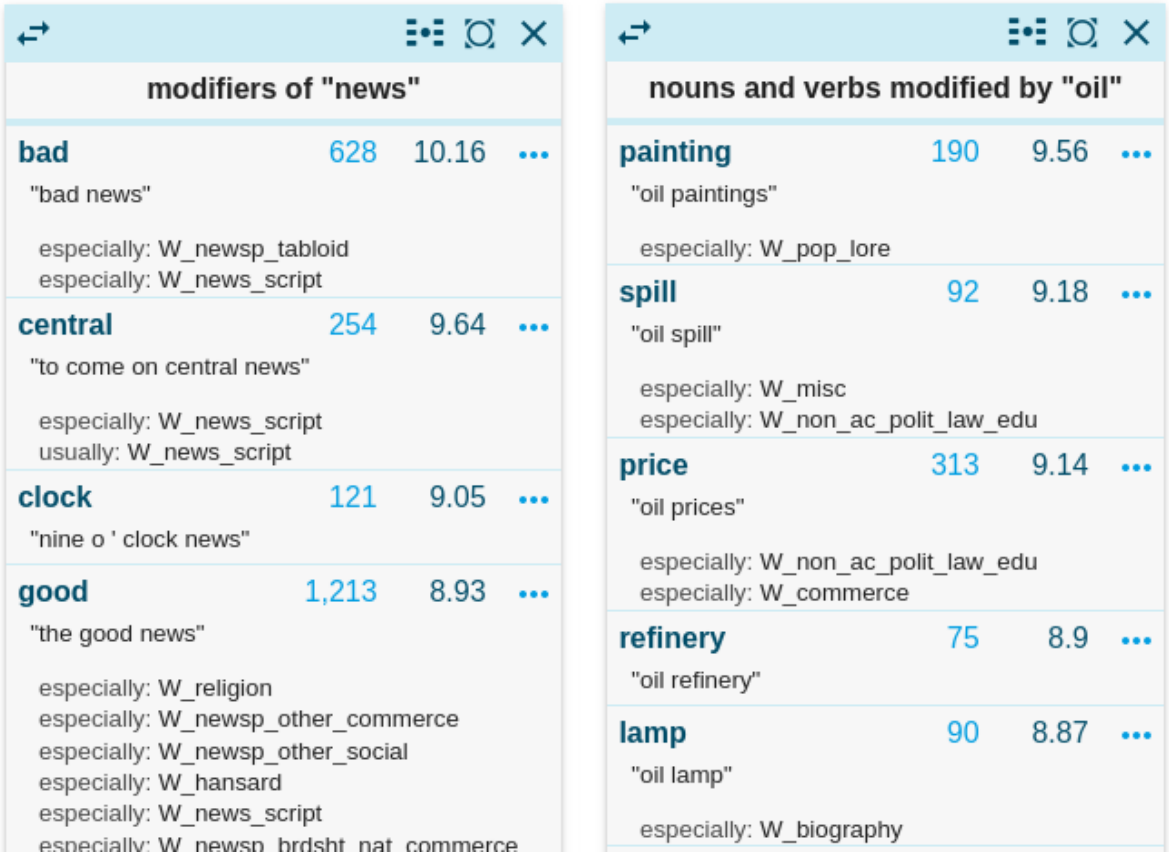

Figure 2: Examples of metadata-specific collocations in the British National Corpus

The first one is a fragment of a word sketch for "news" and shows that *bad news* is specific to tabloid newspapers and TV autocue scripts, whereas *good news* occurs mostly in religious and commercial texts and a variety of other genres.

The second fragment shows the genre-specific collocations of the word "oil": *oil paintings* occurs most frequently in popular magazines, *oil lamps* in biographies, *oil prices* and *oil spills* are political topic,s and *oil prices* is also important in financial texts (*oil spills* is not). *Oil refineries* is covered evenly within all the text types.

Figure 3 shows another example and different text types in the Estonian National Corpus. The example is a fragment of a word sketch for "*kass*" (cat) and shows, for example "*koerte ja kasside pidamise eeskiri*" (rules for keeping dogs and cats) being typical in *Politics, Government & Law*, "*kassi silmad*" (cat eyes) being typical in *Culture & Entertainment* or "*julgem kass*" (braver cat) being predominantly present in *Pets & Animals*.

Figure 3: Examples of metadata-specific collocations in the Estonian National Corpus (Estonian NC 2019)

8. Conclusion

In this paper, we introduced a procedure for including text type information into collocation summaries, such as word sketches. We explained the mental process that ended up with the current specification, then we outlined the implementation, described the user interface and illustrated the output with examples.

The newly introduced functionality is still in its early stage of existence; so far it has only limited production use and has not yet been tested on a large scale. Therefore, some of the parameters may change slightly in the future.

However, we can say that – as in most of the cases concerning corpus data – the future usability of the new feature depends on the quality of the data: the text type annotation, the selection of the right text types to be shown in the word sketch, the corpus having a decent size, as well as the size of particular text types. The quality of the language data in general is one of the biggest challenges for computational linguistics and semi-automatic lexicography in the coming years.

9. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

This work has been partly supported by the Ministry of Education of CR within the LINDAT-CLARIAH-CZ project LM2018101.

10. References

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, 1(2), pp. 263–276.

Kilgarriff, A. (2009). Simple maths for keywords. In *Proceedings of the Corpus Linguistics Conference. Liverpool, UK*.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014a). The Sketch Engine: ten years on. *Lexicography*, 1. URL http://dx.doi.org/10.1007/s40607-014-0009-9.

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P. & Suchomel, V. (2014b). Finding terms in corpora for many languages with the Sketch Engine. *EACL 2014*, p. 53.

Koplenig, A. (2019). Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory*, 15(2), pp. 321–346. URL https://doi.org/10.1515/cllt-2016-0036.

Lee, D. (2002). Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the BNC Jungle. *Language Learning and Technology*, 5.

Leech, G. (1992). 100 million words of English: the British National Corpus (BNC). *Language Research*, 28(1), pp. 1–13.

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries: Where will it all end? *A Taste for Corpora. In Honour of Sylviane Granger*, pp. 257–282.

Rychlý, P. (2008). A lexicographer-friendly association score. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pp. 6–9.

Sharoff, S., Umanskaya, E. & Wilson, J. (2014). *A frequency dictionary of Russian: Core vocabulary for learners*. Routledge.