Evaluating Natural Language Processing Tasks with Low Inter-Annotator Agreement: The Case of Corpus Applications

Vojtěch Kovář

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
xkovar3@fi.muni.cz

Abstract. In [1], we have argued that tasks with low inter-annotator agreement are really common in natural language processing (NLP) and they deserve an appropriate attention. We have also outlined a preliminary solution for their evaluation. In [2], we have agitated for extrinsic application-based evaluation of NLP tasks and against the gold standard methodology which is currently almost the only one really used in the NLP field.

This paper brings a synthesis of these two: For three practical tasks, that normally have so low inter-annotator agreement that they are considered almost irrelevant to any scentific evaluation, we introduce an application-based evaluation scenario which illustrates that it is not only possible to evaluate them in a scientific way, but that this type of evaluation is much more telling than the gold standard way.

Key words: NLP, inter-annotator agreement, low inter-annotator agreement, evaluation, application, application-based evaluation, word sketch, thesaurus, terminology

1 Introduction

1.1 Gold standard evaluation methodology

Scientific evaluation of applications in the natural language processing (NLP) field is usually based on so-called *gold standards* – data sets that contain "correct" annotations created mostly by human beings who understand the particular language (and often also the the underlying linguistic theory). In this type of evaluation, we measure the similarity between this gold standard and an output of a particular tool that is being tested.

For example, in case of morphological analysis, such a gold standard is a corpus manually annotated with morphological tags. In case of syntactic analysis, it is a treebank (corpus where each sentence is manually annotated with a syntactic tree). For machine translation, it is a corpus of correct translations. Similarity metrics for these cases usually are:

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2016, pp. 127–134, 2016. © Tribun EU 2016

- percentage of morphological tags that are identical in both gold standard and on the output of a tagger
- various types of tree similarity metrics [3,4,5]
- the famous BLEU score [6] and its modifications

1.2 Problems with gold standards

This methodology, however, has significant drawbacks. In [2], we argued that it often does not measure the important bits of the linguistic information; that the NLP tools often overfit to the gold standards and therefore their output is often not suitable for practical applications; that there is almost no ambiguity allowed in a typical gold standard; or that the particular evaluation results crucially depend on arbitrary decisions taken at the time of building the gold standard.

As we explain in [2], inter-annotator agreement (IAA) is another issue; it is one of the most important and most problematic aspects of gold standard evaluations. Although high IAA is usually considered crucial for the task to be "well-defined", it is rarely officially published. Often, the lack of agreeent is addressed by extensive annotation manuals (one example for all: annotation guide to a tectogrammatical layer of syntactic annotation in the Prague Dependency Treebank [7] with more than 1200 pages!) that are impossible to memorize – which (apart from frequent errors and inconsistencies) leads to the annotations being record of all the arbitrary decisions present in the manual, rather than native speaker language intuition.

However, there is one problem that is even more important: For some tasks, such as collocation extraction, building an automatic thesaurus, or terminology extraction, the IAA is so low that it is almost impossible to build gold standards for them [8,9], and thus they are doomed to be considered ill-defined and not suitable for scientific evaluation. However, these applications are far from being useless, rather the opposite: there are quite strong commercial interests in them, as can be illustrated e.g. by the successful Sketch Engine service [10] – and we need to be able to evaluate them in a scientific way!

1.3 What this paper is about

In [1], we argued that applications with low IAA should not be considered inferior and that we should find a way to evaluate them. We also introduced a preliminary evaluation methodology for these low-IAA tasks, still based on the gold standard methodology. This paper presents a shift from the gold standards to the purely application-based methodology, and presents a concrete evaluation methods for the three already mentioned applications: collocation extraction, as in the word sketches [10], automatic (distributional) thesaurus generation, and terminology extraction.

All of these are commercially interesting applications that are around already for a rather long time, but so far have not been sufficiently evaluated. The idea we present here is basically very simple: for the current users, the output of these applications output is useful as it is – so it is the output itself that should be evaluated, and it must be the users who evaluate it (rather than a combination of a gold standard and a similarity measure).

2 Evaluation methodology

The proposed methodology follows the general idea presented in [2]: We present two different versions of the particular application output to the group of users/evaluators; we highlight differences, and the users (evaluators) will decide which parts of which version are better/worse (and, possibly, how much better/worse).

Then we sum the overall results from all the evaluators – this will give us one number that expresses which version is better. Note that it does not matter if the annotators agree with each other or not; the solution with more votes is winning, no matter how different the evaluator's opinions are (this may seem unfair but it simulates the real-world situation).

This evaluation methodology allows a lot of options in number of evaluators, the exact evaluation method, testing sample etc. – all of these aspects will influence the quality and the soundness of the evaluation. On the other hand, this variability also enables evaluation of the applications in different usage scenarios.

Also, as we discussed in [2], this method has its drawbacks – it may be more expensive, less sensitive, more suitable for cheating etc. – but its main feature is priceless: The application is evaluated by real users in real usage scenarios, and it is directly the application that is being evaluated, not an artificial "middleware" which may or may not be important (such as syntactic analysis according to a particular treebank annotation).

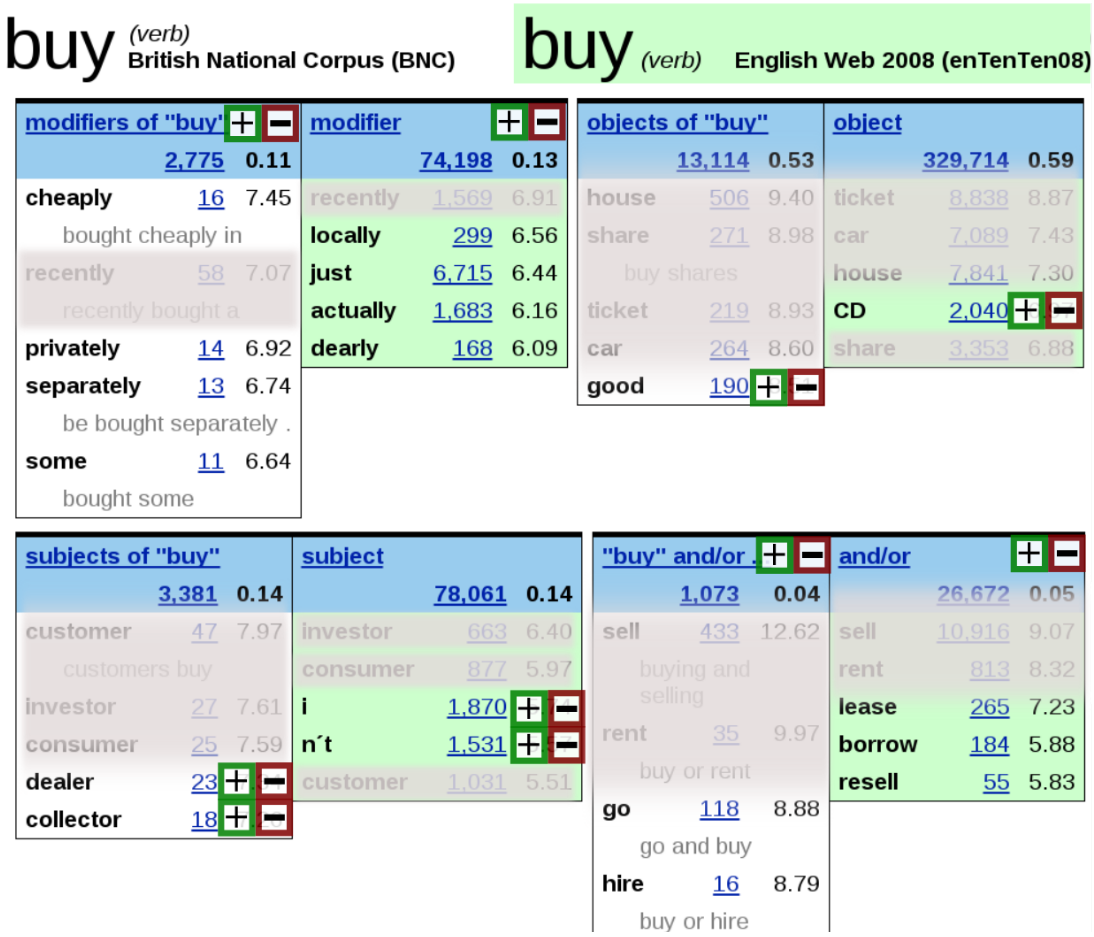
In the following sections, we propose the particular evaluation set-ups for the three already mentioned applications.

3 Collocation extraction

Word sketch [10] is the state of the art application for collocation extraction from corpora. Therefore, we take it as the base for our evaluation. The evaluation will compare two different settings of the word sketch application.

We propose the following evaluation setup:

- we select a set of sample words (may represent a general language use, or can be more specialized)
- for each of the sample words, we display two word sketches on one page, particular relations aligned to each other
- when the relations are very different, we put +/- buttons to the relations
- when the relations only differ in 1 or 2 (maybe 3) words, we put +/- buttons to the particular words
- we hide the common parts



buy (verb) British National Corpus (BNC)

buy (verb) English Web 2008 (enTenTen08)

modifiers of "buy" + −		
	2,775	0.11
cheaply	16	7.45
bought cheaply in		
recently	58	7.07
recently bought a		
privately	14	6.92
separately	13	6.74
be bought separately .		
some	11	6.64
bought some		

modifier + −		
	74,198	0.13
recently	1,569	6.91
locally	299	6.56
just	6,715	6.44
actually	1,683	6.16
dearly	168	6.09

objects of "buy"		
	13,114	0.53
house	506	9.40
share	271	8.98
buy shares		
ticket	219	8.93
car	264	8.60
good	190	+ −

object		
	329,714	0.59
ticket	8,838	8.87
car	7,089	7.43
house	7,841	7.30
CD	2,040	+ −
share	3,353	6.88

subjects of "buy"		
	3,381	0.14
customer	47	7.97
customers buy		
investor	27	7.61
consumer	25	7.59
dealer	23	+ −
collector	18	+ −

subject		
	78,061	0.14
investor	663	6.40
consumer	877	5.97
i	1,870	+ −
n´t	1,531	+ −
customer	1,031	5.51

"buy" and/or + −		
	1,073	0.04
sell	433	12.62
buying and selling		
rent	35	9.97
buy or rent		
go	118	8.88
go and buy		
hire	16	8.79
buy or hire		

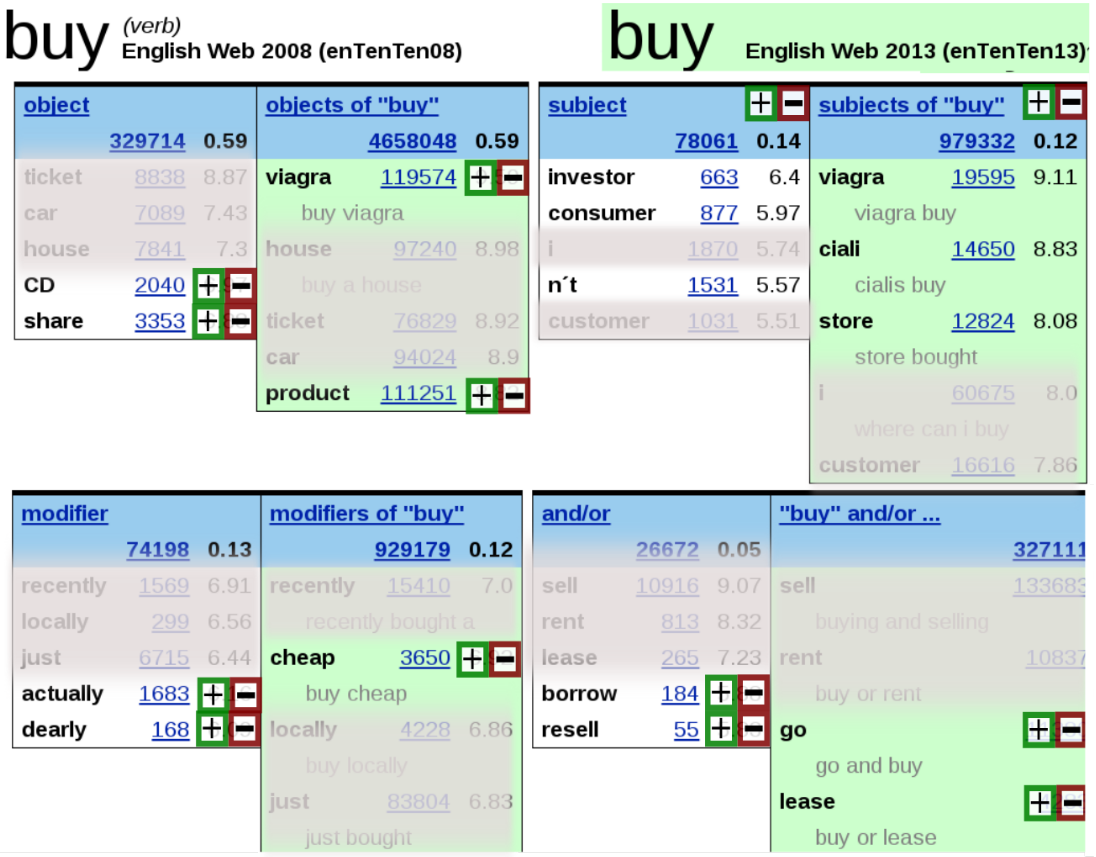
and/or + −		
	26,672	0.05
sell	10,916	9.07
rent	813	8.32
lease	265	7.23
borrow	184	5.88
resell	55	5.83

Fig. 1: Word sketch evaluation proposal: British National Corpus vs. English web corpus enTenTen08

- each evaluator can click each button several times (to express different importance of the differences) but they are not obliged to click anything (to be able to express that something is not really important)
- at the end of the evaluation, we count +1/-1 point for every +/- click on a collocate, +2/-2 points for every click on a relation; the overall sum is the result

Figures 1 and 2 illustrate the particular examples of what the evaluators would see – in some cases, it is perfectly clear which side is better (e.g. the “n’t” collocate is a result of a processing error, “viagra” etc. is a result of web spam present in the corpus), in other cases the opinions may differ.

The figures contain the names of the two corpora but this is only for illustration purposes. In reality we would not show the different settings to the evaluators; firstly because their opinions could be biased by the corpus names, and also because we can measure a wide range of different settings (e.g. different minimum frequency), not just a different corpus.



buy *(verb)* **English Web 2008 (enTenTen08)**

buy **English Web 2013 (enTenTen13)**

object		
	329714	0.59
ticket	8838	8.87
car	7089	7.43
house	7841	7.3
CD	2040	+ −
share	3353	+ −

objects of "buy"		
	4658048	0.59
viagra	119574	+ −
buy viagra		
house	97240	8.98
buy a house		
ticket	76829	8.92
car	94024	8.9
product	111251	+ −

subject	+ −	
	78061	0.14
investor	663	6.4
consumer	877	5.97
i	1870	5.74
n´t	1531	5.57
customer	1031	5.51

subjects of "buy"	+ −	
	979332	0.12
viagra	19595	9.11
viagra buy		
ciali	14650	8.83
cialis buy		
store	12824	8.08
store bought		
i	60675	8.0
where can i buy		
customer	16616	7.86

modifier		
	74198	0.13
recently	1569	6.91
locally	299	6.56
just	6715	6.44
actually	1683	+ −
dearly	168	+ −

modifiers of "buy"		
	929179	0.12
recently	15410	7.0
recently bought a		
cheap	3650	+ −
buy cheap		
locally	4228	6.86
buy locally		
just	83804	6.83
just bought		

and/or		
	26672	0.05
sell	10916	9.07
rent	813	8.32
lease	265	7.23
borrow	184	+ −
resell	55	+ −

"buy" and/or ...	
	327111
sell	133683
buying and selling	
rent	10837
buy or rent	
go	+ −
go and buy	
lease	+ −
buy or lease	

Fig. 2: Word sketch evaluation proposal: Older English web corpus enTenTen08 vs. newer English web corpus enTenTen13

4 Thesaurus

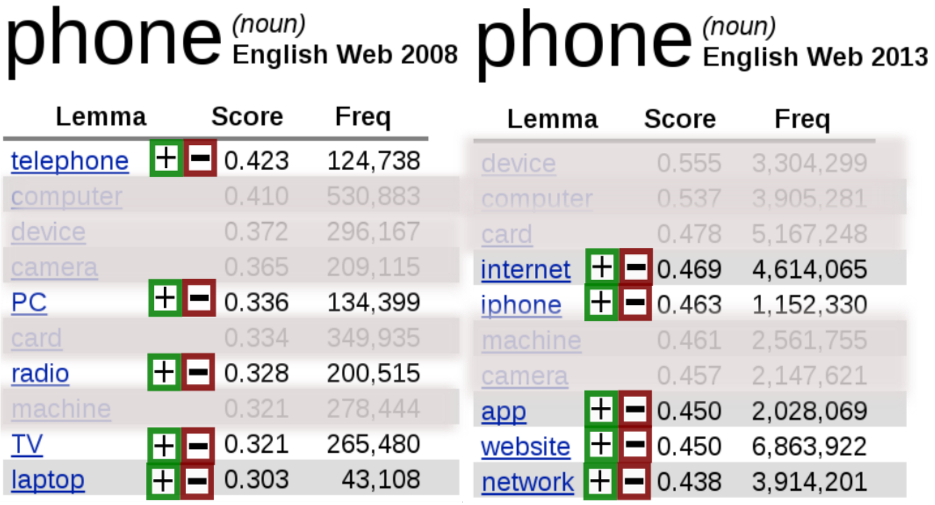
Thesaurus is basically just a list of similar words, so the task reduces to comparison of two lists, again for a given sample of words.

The scenario here is very similar to what we propose in case of collocation extraction: take the top of both lists, put the two lists side by side, ignore common items and evaluate individual items on the list by clicking +/-. Sum of positive and negative points is then the score of a particular list.

An example of what the annotators would see is in Figure 3. Again, the corpus names would not be shown.

5 Terminology extraction

Extraction of terminology from domain-specific texts is in fact another list, so the procedure can be very similar to the thesaurus evaluation, as introduced above. There is just one difference: the terminology is not for one particular word, but for a whole corpus, so no sample of words is needed here. Rather



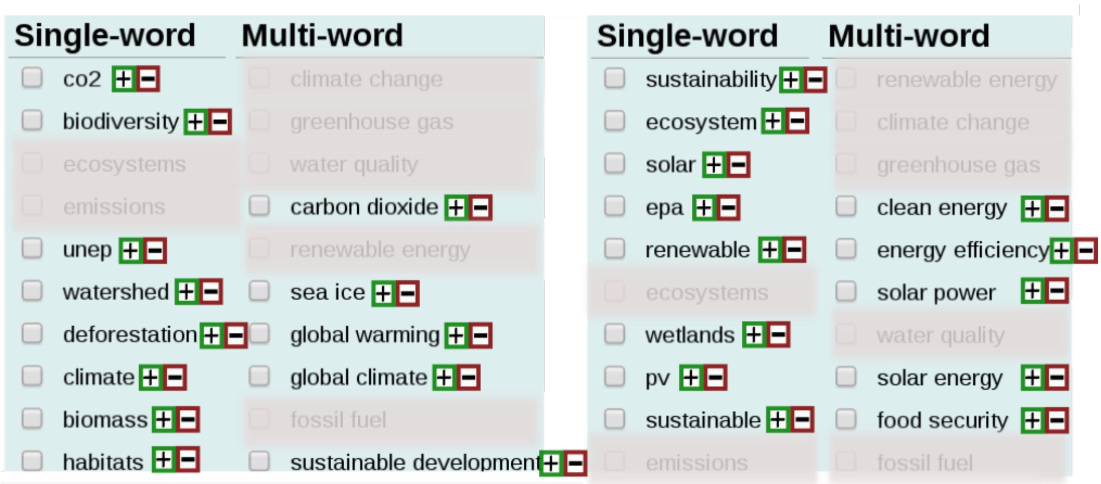
phone *(noun)* English Web 2008

Lemma		Score	Freq
telephone	+ −	0.423	124,738
computer		0.410	530,883
device		0.372	296,167
camera		0.365	209,115
PC	+ −	0.336	134,399
card		0.334	349,935
radio	+ −	0.328	200,515
machine		0.321	278,444
TV	+ −	0.321	265,480
laptop	+ −	0.303	43,108

phone *(noun)* English Web 2013

Lemma		Score	Freq
device		0.555	3,304,299
computer		0.537	3,905,281
card		0.478	5,167,248
internet	+ −	0.469	4,614,065
iphone	+ −	0.463	1,152,330
machine		0.461	2,561,755
camera		0.457	2,147,621
app	+ −	0.450	2,028,069
website	+ −	0.450	6,863,922
network	+ −	0.438	3,914,201

Fig. 3: Thesaurus evaluation proposal: Older English web corpus enTenTen08 vs. newer English web corpus enTenTen13



Single-word	Multi-word
□ co2 + −	□ climate change
□ biodiversity + −	□ greenhouse gas
□ ecosystems	□ water quality
□ emissions	□ carbon dioxide + −
□ unep + −	□ renewable energy
□ watershed + −	□ sea ice + −
□ deforestation + −	□ global warming + −
□ climate + −	□ global climate + −
□ biomass + −	□ fossil fuel
□ habitats + −	□ sustainable development + −

Single-word	Multi-word
□ sustainability + −	□ renewable energy
□ ecosystem + −	□ climate change
□ solar + −	□ greenhouse gas
□ epa + −	□ clean energy + −
□ renewable + −	□ energy efficiency + −
□ ecosystems	□ solar power + −
□ wetlands + −	□ water quality
□ pv + −	□ solar energy + −
□ sustainable + −	□ food security + −
□ emissions	□ fossil fuel

Fig. 4: Terminology evaluation proposal: 60M Environment domain corpus with two different reference corpora: big (11 billion words) web corpus enTenTen12 on the left, and small (7 million words) manually created corpus Brown Family.

than that, we need a sample of domain specific texts. Again, the results may be very different for different samples, however, this reflects the reality: A terminology extractor can also be very good on one domain and very bad on another one.

The fact that no sample of words is needed means that we can include more items into the list, not just 10 or 20 as in case of word sketches and thesaurus. And we really should do that because terminology extraction has a different use case than the two other applications. In both word sketch and thesaurus, the user typically looks at up to 20 top items; in case of terminology, thousands of items may be extracted (e.g. for the purpose of compiling a specialized dictionary) to be further processed.

We should always be sure that we are testing something that is as close to the real use case, to the real application of the particular tool, as possible.

An example of what the annotators would see is in Figure 4 – but as we've just explained, in reality the lists would be longer (and because of that, they would probably also contain more hidden items).

6 Conclusions

Based on previous theoretical work, we have introduced a concrete scenario of application-based evaluation of three NLP tasks with low inter-annotator agreement. We believe this proposal will be implemented in a short time and used as an evaluation framework for these tasks.

Future work consists mainly in actually *doing* a robust evaluation of these three tasks according to the scenarios introduced in this paper, for various corpora and various settings.

Acknowledgments. The research leading to these results has received funding from the Norwegian Financial Mechanism 2009–2014 and the Ministry of Education, Youth and Sports under Project Contract no. MSMT-28477/2014 within the HaBiT Project 7F14047.

References

1. Kovář, V., Rychlý, P., Jakubíček, M.: Low inter-annotator agreement = an ill-defined problem? In: Eighth Workshop on Recent Advances in Slavonic Natural Language Processing, Brno, Tribun EU (2014) 57–62
2. Kovář, V., Jakubíček, M., Horák, A.: On evaluation of natural language processing tasks: Is gold standard evaluation methodology a good solution? In: Proceedings of the 8th International Conference on Agents and Artificial Intelligence, Rome, SCITEPRESS (2016) 540–545
3. Grishman, R., Macleod, C., Sterling, J.: Evaluating parsing strategies using standardized parse files. In: Proceedings of the third conference on Applied natural language processing, Association for Computational Linguistics (1992) 156–161
4. Sampson, G.: A proposal for improving the measurement of parse accuracy. International Journal of Corpus Linguistics **5**(01) (2000) 53–68
5. Sampson, G., Babarczy, A.: A test of the leaf-ancestor metric for parse accuracy. Natural Language Engineering **9**(04) (2003) 365–380
6. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics (2002) 311–318

7. Mikulová, M., Bémová, A., Hajič, J., Hajičová, E., Havelka, J., Kolářová, V., Kučová, L., Lopatková, M., Pajas, P., Panevová, J., Razímová, M., Sgall, P., Štěpánek, J., Urešová, Z., Veselá, K., Žabokrtský, Z.: Annotation on the tectogrammatical layer in the Prague Dependency Treebank (2005) `http://ufal.mff.cuni.cz/pdt2.0/doc/manuals/en/t-layer/pdf/t-man-en.pdf`.
8. Kilgarriff, A., Kovář, V., Krek, S., Srdanović, I., Tiberius, C.: A quantitative evaluation of Word Sketches. In: Proceedings of the XIV Euralex International Congress, Ljouwert, Netherlands, Fryske Akademy (2010) 372–379
9. Kilgarriff, A., Rychlý, P., Jakubíček, M., Kovář, V., Baisa, V., Kocincová, L.: Extrinsic corpus evaluation with a collocation dictionary task. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, European Language Resources Association (ELRA) (2014) 1–8
10. Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V.: The Sketch Engine: ten years on. Lexicography **1** (2014)