ShadowSense: a Multi-annotated Dataset for Evaluating Word Sense Induction

Ondřej Herman[1,2], **Miloš Jakubíček**[1,2]
[1] Lexical Computing, Brno, Czech Republic
[2] Faculty of Informatics, Masaryk University, Brno, Czech Republic
ondrej.herman@sketchengine.eu, milos.jakubicek@sketchengine.eu

Abstract

In this paper we present a novel bilingual (Czech, English) dataset called ShadowSense developed for the purposes of word sense induction (WSI) evaluation. Unlike existing WSI datasets, ShadowSense is annotated by multiple annotators whose inter-annotator agreement represents key reliability score to be used for evaluation of systems automatically inducing word senses. In this paper we clarify the motivation for such an approach, describe the dataset in detail and provide evaluation of three neural WSI systems showing substantial differences compared to traditional evaluation paradigms.

Keywords: WSI, word sense induction, ShadowSense

1. Introduction

Word sense induction (WSI) is an established NLP task focusing on identification of sense usages in natural language. Unlike word sense disambiguation (WSD), word sense induction operates without a predefined sense inventory and its output typically consists of a set of usage clusters (where each usage corresponds to a particular sense occurrence including its context, such as a sentence).

The fact that WSI does not assume a pre-defined discrete sense inventory makes the task more theoretically plausible (see e.g. Kilgarriff, 1997) and as such WSI represents a key role specifically in the context of lexicography where relying on any fixed sense inventory coming from existing lexical resources for WSD application typically creates a vicious circle: making a sense inventory depending on an existing one does not make it possible to divert as for sense granularity or account for new senses resulting from language development or different text types.

While WSI is more relevant than WSD in many usage scenarios, it is also more difficult to evaluate, particularly as grouping sense usages by human annotators has obviously even a lower inter-annotator agreement (see Erk et al., 2009) than classifying in a WSD context which already suffers from a very low inter-annotator agreement either (but benefits from annotators being biased by the predefined sense inventory) as shown e.g. in Kilgarriff and Rosenzweig (2000).

In this paper we focus on the evaluation part of WSI and present a novel dataset that accounts for the unavoidable human disagreements by relying on each instance being annotated by multiple and many (ten) annotators. We argue that such annotation makes it possible to either disregard low-agreement instances in the evaluation or carry out a weighted evaluation where agreement rates represent weights of individual instances. We show that such evaluation differs substantially from prevailing evaluation paradigms relying either on measuring homogeneity and completeness and V-measure as their harmonic mean.

2. Related work

WSI is a long studied task which has been heavily promoted by the SemEval shared task series, particularly in the years 2007, 2010 and 2013 (see Agirre and Soroa, 2007; Manandhar et al., 2010; Jurgens and Klapaftis, 2013, respectively). Each of the turns aimed at various improvements in the evaluation of the task. While 2007 relied on classification-standard F-measure, in 2010 the V-measure was used to better account for the clustering nature of the task and in 2013 a "fuzzy" evaluation was performed relying on annotators having the option to assign multiple senses per instance.

In all three years the reliability of the evaluation data turned out to be one of the biggest evaluation problems. In 2007 and 2010 a small subset of the Wall Street Journal corpus was used hand-annotated with OntoNotes senses (Hovy et al., 2006). In 2013 the data from Erk et al. (2009) was used, however, the test set description[1] contains the following statement:

"The Task 13 test set contains annotations for 4664 instances. Of those, 517 were annotated with two senses (11%) and 25 were annotated with three sense (0.5%). This low percentage of multiple-annotations is in stark contrast with the trial data from the GWS dataset of Erk et al. (2009), which

[1] As found at https://github.com/ai-ku/semeval13-task13/tree/master/test_data since original website is no longer active.

LREC-COLING 2024, pages 14763–14769
20-25 May, 2024.

featured multiple annotations on every instance. A re-analysis of their dataset by trained lexicographers revealed that annotators were often mistaken regarding the specific application of senses and were therefore more likely to rate in applicable senses as applicable. The Task 13 test data adopted a conservative sense annotation approach that involved making sense applicability judgments based on all available information and examples in WordNet 3.1, such as sentence frames, coordinate terms, antonyms, etc., which were not available to the annotators for the trial data."

Without having any doubts on the best intents of the SemEval 2013 Task 13 organizers, this statements raises some concerns. First of all, it disregards important findings made by Erk et al. (2009) on graded sense annotations without substantial investigation. Second, there is no information provided about the inter-annotator agreement of the trained lexicographers, whoever they were – and no information even on how many they were. Finally, the admitted influence of the WordNet 3.1 inventory could have introduced a significant evaluation bias.

3. Annotation methodology

In this paper we decided to take an approach which can be seen as somewhat complementary to the one used by Erk et al. (2009). Instead of relying on graded multi-sense annotations, i.e. allowing annotators to assign multiple senses to one context, we asked multiple annotators to perform the actual word sense clustering with arbitrary clusters and labels on a large web corpus enTenTen2008 (Jakubíček et al., 2013).

To make the task manually tractable, we exported word sketch annotations from Sketch Engine (Kilgarriff et al., 2014), a corpus platform hosting the enTenTen2008 corpus. Sketch Engine uses morphosyntantic collocation descriptions in the form of a so called sketch grammar (formalized as regular expressions in the corpus query language facilitating morphological annotation of the corpus) that is applied to corpora and together with the logDice association score (Rychlý, 2008) used to find collocations candidates categorized by syntactic relations.

We have manually selected 24 polysemous words in English and 25 in Czech, and exported top 150 collocations from Sketch Engine, across all syntactic relations, sorted by the logDice score, as well as all concordance lines pertaining to each collocation. We dropped weakly-bound collocations with logDice score smaller than 5.5, so there are fewer than 150 collocations for some of the rarer headwords in the dataset. Altogether, we extracted 8,178,835 concordance lines representing individual sense contexts. For each collocation, we also exported its common full text usage called longest-commonest match in Sketch Engine (Kilgarriff et al., 2015).[2]

The annotators were presented with the extracted collocation in the form of *(headword, relation, collocate)* triples together with the longest-commonest match and a link to the corresponding Sketch Engine collocation concordance. Instead of clustering individual usages (concordance lines), they were asked to cluster the collocations, possibly inspecting concordance evidence if in doubt. Order of the lines for each headword was randomized for each of the annotators.

The Czech part was annotated by 5 native speakers, while the English part was annotated by 6 native speakers and by 4 non-native speakers with the CEFR B2 level of English ability.

We provided no predetermined sense inventory to the annotators. Instead, the annotators were instructed to choose the senses as they see fit; the criterion whether one collocation is used in the same sense as another collocation is whether the particular context is related or unrelated. Therefore, sense inventories of each annotator is different. The annotators have been instructed to skip sense-unspecific or mixed collocations and also those for which their certainty was low.

Overall, the annotators have assessed 3,352 collocations backed by 1,573,671 concordance contexts for English and 3,476 collocations backed by 6,605,164 concordance contexts for Czech. For evaluating of their inter-annotator agreement, we calculated the Rand score (Hubert and Arabie, 1985) across all annotations.

For the English set, the annotators selected 8.11 senses for each word on average ($\sigma = 2.74$), for the Czech set, the annotators selected 3.6 senses on average ($\sigma = 0.53$).

Table 1 shows the average proportion of unassigned collocations to senses by each annotator. For the ESL speakers, the large imbalance can be explained by the speakers' uncertainty over the command of the English language.

On the other hand, the native speakers were significantly more averse to leave a collocation unassigned. A possible explanation might be that all of the native annotators were hired only for this particular task and paid by the hour, which may motivate them to make the annotation seem more complete at a first glance. On the other hand, the annotators for the Czech test set and the ESL annotators are people we have a longer relationship with, which makes them not feel the need to prove themselves on this task.

[2]The longest-commonest match is, roughly speaking, the most common extended collocation context that covers at least 25% of all usages (concordance lines), i.e. it is a common super-phrase that the collocation occurs in.

Language	Annotator	Unassigned
Cs	1	24.3 %
Cs	2	25.2 %
Cs	3	21.1 %
Cs	4	9.4 %
Cs	5	6.3 %
En Native	1	0.0 %
En Native	2	1.0 %
En Native	3	0.2 %
En Native	4	0.1 %
En Native	5	3.8 %
En Native	6	6.3 %
En ESL	7	25.0 %
En ESL	8	10.1 %
En ESL	9	48.9 %
En ESL	10	5.0 %

Table 1: Unassigned Collocations By Annotator

3.1. Inter-Annotator agreement

Traditional methods, such as Cohen's κ are of little use here due to the structure of the data, as there are multiple annotators, each using different, unknown labelling. We compared the annotations using the Rand Index (Hubert and Arabie, 1985).

Table 2 shows the pairwise Rand Index between all pairs of the five annotators of the Czech dataset. The mean Rand Index is 0.914.

Table 3 shows the pairwise Rand Index between all pairs of the ten annotators of the English dataset. The mean Rand Index is 0.817. The mean Rand Index slightly increases to 0.822 if only the senses annotated by the native speakers are considered.

	2	**3**	**4**	**5**
1	.88	.94	.87	.92
2		.92	.90	.91
3			.92	.94
4				.91

Table 2: Pairwise Rand Index for Czech annotators

	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	.79	.84	.82	.84	.86	.84	.82	.84	.85
2		.79	.81	.81	.80	.81	.81	.78	.81
3			.81	.81	.82	.82	.81	.81	.83
4				.82	.81	.82	.82	.83	.83
5					.83	.81	.80	.82	.82
6						.83	.86	.80	.84
7							.84	.83	.86
8								.82	.83
9									.84

Table 3: Pairwise Rand Index for English annotators

3.2. Structure of the Dataset

ShadowSense is available at https://github.com/lexicalcomputing/ShadowSense.

The dataset is provided as in a TAB separated columnar format, encoded using the UTF-8 character encoding. The data is separated is one file for each language.

The English dataset contains instances from the enTenTen2008 corpus (Jakubíček et al., 2011). One line in the data file represents a single instance. For each instance, we provide the target headword in a lemmatized form, with its PoS tag appended, the Word Sketch triple which was used for the annotation. The context is a sentence, within which the target instance is marked by the < and > characters. Should your application require different context processing, the token sequence number within the corpus is also provided and arbitrary context can be extracted from the full corpus text, which is available for download. The sense annotations have the form aX.sY – X-th annotator assigned Y-th sense to the instance. Note that there is no relationship between the sense numbers of different annotators or different headwords. Each of the annotators uses their own sense inventory.

The Czech dataset is derived in the same way from the csTenTen17 corpus (Suchomel, 2018).

For convenience, and as the naive evaluation of the test statistics can be expensive, a high-performance scorer program is distributed along the data. Its usage is described in the repository.

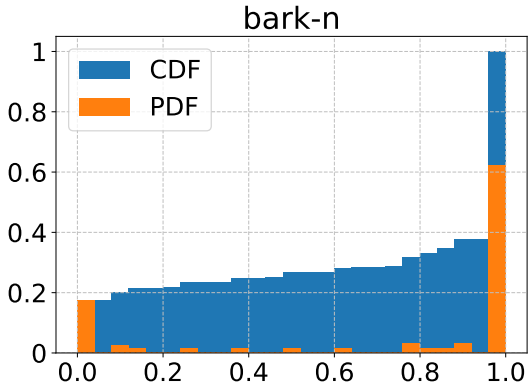

Figure 1: Pairwise annotation agreement distribution for the noun *bark*

4. Evaluation methodology

We will show the intuition behind the evaluation on two extreme examples from the dataset and then provide a rigorous description.

Each entry from the dataset consists of a headword, a corpus context and a sequence of sense

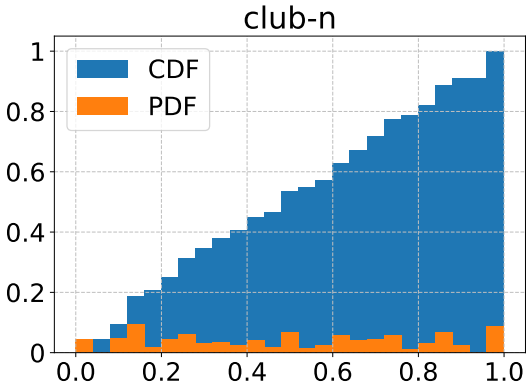

Figure 2: Pairwise annotation agreement distribution for the noun *club*

annotations, one annotation per annotator. Some of the annotations can be empty and they are not directly comparable between the annotators in any way, their labelling is arbitrary.

The only way the annotations can be compared is to inspect a **pair** of annotations for two different contexts. Based on the sense labelling, we can tell whether each annotator believes that these two contexts belong into a single sense, or whether they should be kept separate and represent two different senses. As every instance is annotated by multiple annotators, we can quantify the agreement between the annotators.

Figure 1 shows the cumulative and density distribution of all annotation pairs for the lemma *bark* (mainly *tree bark*, *dog's bark*) according to the degree of agreement. The x-axis represents the decision of the annotators for a pair of annotations. The leftmost bin contains those pairs for which every annotator agreed that the two instances should belong into distinct senses, while the rightmost bin represents the pairs for which the annotators claim that the instances represent into the same sense. The bins in-between the two extremes contain those pairs of annotations, where only some of the annotators claim that the sense is the same. In the case of the word *bark*, the senses are well-separated and the grey area, where the annotators do not fully agree, is small. Notice that it is not possible to say anything specific about the number of senses from the pairwise agreement distribution.

Figure 2 shows the pairwise sense annotation agreement for the word *club*. Here the situation is significantly less clear and the sense separation is much worse. Annotators do not agree on whether, for example, *golf club* and *cricket club* belong into the same sense, but also whether *golf club* and *football club* should be kept together or separate, so most of the pairs lie in the gray area.

4.1. Evaluation statistics

We provide two evaluation methodologies for ShadowSense that differ in how they treat disagreements of clustering carried out by the annotators. In the first setting, which we call the shadow Rand index (sRI), we only consider the pairs of annotations which agree at least 75 % of time, while in the second setting, the weighted shadow Rand index (wsRI), we employ the agreement between the annotators to assign lesser weights to the annotations which do not agree strongly, so that the dataset can be exploited more efficiently.

Both of the statistics are based on the Rand index (Vinh et al., 2009) and operate over pairs of annotations.

For i-th headword instance out of m instances in total, we have the sense annotations from n annotators $A_i = (s_{i,1}, \cdots, s_{i,n})$, and from the evaluated WSI system we obtain the assignment to a cluster c_i.

For each pair of instances, we calculate the annotation agreement on the annotations that were not left empty by either of the annotators as

$$r_{ij} = \frac{\sum_{k, s_{ik} \neq \perp, s_{jk} \neq \perp} \begin{cases} 1 & s_{ik} = s_{jk} \\ 0 & otherwise \end{cases}}{\sum_{k, s_{ik} \neq \perp, s_{jk} \neq \perp} 1}$$

For the calculation, we only consider pairs for which at least half of the annotators annotated both of their instances in the instance pair.

4.2. sRI

We calculate confusion matrix elements, but only for the pairs of annotations which agree 75 % of time or more. The pairs are then influence the result with equal weight, no matter how strong the agreement is. The rest of the pairs are not considered for the calculation.

$$tp = \sum_{i,j \in 1..m} \begin{cases} 1 & c_i = c_j, r_{ij} >= 0.75 \\ 0 & otherwise \end{cases}$$

$$tn = \sum_{i,j \in 1..m} \begin{cases} 1 & c_i \neq c_j, r_{ij} <= 0.25 \\ 0 & otherwise \end{cases}$$

$$fp = \sum_{i,j \in 1..m} \begin{cases} 1 & c_i = c_j, r_{ij} <= 0.25 \\ 0 & otherwise \end{cases}$$

$$fn = \sum_{i,j \in 1..m} \begin{cases} 1 & c_i \neq c_j, r_{ij} >= 0.75 \\ 0 & otherwise \end{cases}$$

The sRI statistic is then calculated as

$$\mathrm{sRI} = \frac{2(tp \cdot tn - fp \cdot fn)}{(tn + fn)(tp + fp) + (tn + fp)(tp + fn)}$$

The range of the statistic is bounded by 1 from the top, but lower bound can be slightly below zero due to the adjustment for chance (Erk et al., 2009).

4.3. wsRI

The wsRI statistic exploits the annotator agreement to provide weighting of the pairs entering the evaluation. This improves the explanatory power for words for which the agreement is low and enables better usage of the test set.

We introduce a weighting factor w_{ij} for a pair of annotations i and j, assigning linear weights to the pairs based on their distance from the midpoint, so the weight is 0 when half of the annotators agree that two instances belong to the same sense, while the second half disagrees. The weight is 1 if all annotators agree or disagree.

$$w_{ij} = 2|0.5 - r_{ij}|$$

The confusion matrix elements are then

$$tp = \sum_{i,j \in 1..m} \begin{cases} w_{ij} & c_i = c_j, r_{ij} > 0.5 \\ 0 & otherwise \end{cases}$$

$$tn = \sum_{i,j \in 1..m} \begin{cases} w_{ij} & c_i \neq c_j, r_{ij} < 0.5 \\ 0 & otherwise \end{cases}$$

$$fp = \sum_{i,j \in 1..m} \begin{cases} w_{ij} & c_i = c_j, r_{ij} < 0.5 \\ 0 & otherwise \end{cases}$$

$$fn = \sum_{i,j \in 1..m} \begin{cases} w_{ij} & c_i \neq c_j, r_{ij} > 0.5 \\ 0 & otherwise \end{cases}$$

The resulting wsRI statistic is then calculated as

$$\mathrm{wsRI} = \frac{2(tp \cdot tn - fp \cdot fn)}{(tn + fn)(tp + fp) + (tn + fp)(tp + fn)}$$

Compared to sRI, wSRI does not use a hard cutoff within the calculation, so the statistic is more stable when facing highly ambiguous test sets or test sets annotated by a large number of annotators.

5. Evaluation of sample WSI systems

We evaluated ShadowSense against three WSI systems. In addition, we evaluated the same systems on the SemEval 2013 Task 13 WSI dataset for comparison using the Fuzzy B-cubed and Fuzzy NMI statistics for reference.

The systems we evaluated were SymPatternWSI[3] (Amrami and Goldberg, 2018), BertWSI[4] (Amrami and Goldberg, 2019), which are based on language modelling approaches, and an in-house reimplementation of Adaptive Skip-Gram[5] (Bartunov et al., 2016), which uses multisense word embeddings.

BertWSI is the clear leader in every test. SymPatternWSI provided respectable scores for the SemEval 2013 Task 13 dataset, but consistently failed to induce more than a single sense for most headwords from our dataset, possibly due to hyperparameter tuning. However, the scores for each system comfortably beat every system competing in the original SemEval 2013 Task 13 shootout (Jurgens and Klapaftis, 2013).

The SemEval 2013 Task 13 test set assigns multiple, weighted, senses to some of the gold instances, as do the evaluated systems. Where possible, we kept the multiple assignments when using the original metrics, but for the purposes of ShadowSense evaluation, we chose the most probable sense from the possibilities offered. As the organizers of the shared task note, this is not a common occurrence in the gold set, in contrast to the training set.

The sRI and wsRI correlate with the fuzzy NMI and fuzzy B-cubed metrics, but make fewer assumptions on the annotated data. The number of clusters does need to be known, no sense inventory needs to be present. The statistics have other desirable properties, such as yielding a result close to zero when a WSI method assigns every instance into the same cluster.

Difference between sRI and wsRI is small, but wsRI should be able to provide more stratified result over different WSI methods when the test set has many annotators or is the interannotator agreement is low.

6. Conclusions

We described ShadowSense, a test set for evaluating word sense induction systems for Czech and English. ShadowSense aims to provide an unbiased evidence about the senses of words, showing

[3] https://github.com/asafamr/SymPatternWSI

[4] https://github.com/asafamr/bertwsi.

[5] The original repository is available at https://github.com/sbos/AdaGram.jl.

	SemEval2013		ShadowSense	
	fNMI	**fBC**	**sRI**	**wsRI**
SymPatternWSI	.115	.572	-.004	-.004
BertWSI	.209	.641	.757	.761
AdaGram ($\alpha = 0.1, d = 256, w = 4$)	.065	.455	.285	.282

Table 4: Performance Comparison of Select WSI Systems

the decisions of multiple annotators. The methodology does not require a common sense inventory or any coordination between the annotators.

We benchmarked multiple WSI systems against ShadowSense and shown that the shadow Rand index and weighted shadow Rand index have desirable properties.

7. Limitations

The obvious limitation of the annotation procedure is that it heavily relies on the assumption that a collocation belongs typically to one sense only, and if it does not, that the annotators can successfully conclude so (and omit the collocation from annotating, manifested by the ".x" suffix class). Obviously, there will be cases where annotators would intentionally assign a collocation to a cluster while concordance evidence would suggest that the collocation features multiple senses. The only way that we are aware of as for mitigating this situation is relying on as many annotators as possible in order to minimize such scenarios.

Another limitation is that the evaluation method does not consider fuzzy assignment of a context to multiple senses, but current WSI systems commonly assign multiple senses to a target instance. However, we would prefer the WSI algorithm to choose the clusters and their number so that mixed instances do not appear in their output, and therefore we believe that this is not an important limitation for most applications, but we intend to explore the possibility of extending the metric to the result of a fuzzy WSI method.

8. Future work

The downstream applications, for which we employ the results of the WSI algorithms, operate on many languages. The results which can be obtained on two language test set are of limited use for other languages. Tuning a WSI method on the current ShadowSense version is unlikely to provide parameters generalizable to languages from other language familes or employing different scripts and tokenizations. To increase the coverage, we intend to create similar datasets for other languages.

An useful improvement of the test statistics would separate the result into two parts analogous to precision and recall, or provide a way of weighting the statistics, as in many applications it is desirable that the clusters obtained from a WSI method are not too coarse or narrow. For example clustering result containing a few mixed clusters is very difficult to split, while lumping fine clusters is much easier. Both sRI and wSRI prefer to balance the granularity to lie in the midpoint.

Another disjoint avenue of research we intend to pursue is the automated extraction of a sense hierarchy from the data, but the requirements for the amount of simultaneous annotations is likely much higher than what the current dataset provides.

9. Acknowledgments

The work described herein has been supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062 LINDAT/CLARIAH-CZ and by Lexical Computing through its industrial partnership funding scheme at Faculty of Informatics, Masaryk University.

10. Bibliographical References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 task 02: Evaluating word sense induction and discrimination systems. In *Proceedings of the fourth international workshop on semantic evaluations (semeval-2007)*, pages 7–12.

Asaf Amrami and Yoav Goldberg. 2018. Word sense induction with neural bilm and symmetric patterns. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4860–4867.

Asaf Amrami and Yoav Goldberg. 2019. Towards better substitution-based word sense induction. *arXiv preprint arXiv:1905.12598*.

Sergey Bartunov, Dmitry Kondrashkin, Anton Osokin, and Dmitry Vetrov. 2016. Breaking sticks and ambiguities with adaptive skip-gram. In *artificial intelligence and statistics*, pages 130–138. PMLR.

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2:193–218.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *7th international corpus linguistics conference CL*, pages 125–127.

Otto Jespersen. 1922. *Language: Its Nature, Development, and Origin*. Allen and Unwin.

David Jurgens and Ioannis Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 290–299.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31:91–113.

Adam Kilgarriff, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlỳ, and Vít Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, 1(1):7–36.

Adam Kilgarriff, Vít Baisa, Pavel Rychlỳ, and Miloš Jakubíček. 2015. Longest–commonest match. In *Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 conference*, pages 11–13. Trojina, Institute for Applied Slovene Studies/Lexical Computing Ltd ....

Adam Kilgarriff and Joseph Rosenzweig. 2000. Framework and results for english senseval. *Computers and the Humanities*, 34:15–48.

Suresh Manandhar, Ioannis Klapaftis, Dmitriy Dligach, and Sameer Pradhan. 2010. Semeval-2010 task 14: Word sense induction &disambiguation. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 63–68.

Pavel Rychlỳ. 2008. A lexicographer-friendly association score. In *RASLAN*, pages 6–9.

Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2009. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th annual international conference on machine learning*, pages 1073–1080.

11. Language Resource References

Jakubíček, Miloš and Kilgarriff, Adam and Kovář, Vojtěch and Rychlỳ, Pavel and Suchomel, Vít. 2011. *enTenTen*. PID http://hdl.handle.net/11858/00-097C-0000-0001-CCDF-8. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Suchomel, Vít. 2018. *Czech Web Corpus 2017 (csTenTen17)*. PID http://hdl.handle.net/11234/1-4835. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.