Collection, Annotation and Analysis of Gold Standard Corpora for Knowledge-Rich Context Extraction in Russian and German

Anne-Kathrin Schumann
Saarland University/University of Vienna
anne.schumann@mx.uni-saarland.de

Abstract

This paper describes the collection, annotation and linguistic analysis of a gold standard for knowledge-rich context extraction on the basis of Russian and German web corpora as part of ongoing PhD thesis work. In the following sections, the concept of knowledge-rich contexts is refined and gold standard creation is described. Linguistic analyses of the gold standard data and their results are explained.

1 Introduction

Defining statements have long been recognised as a fundamental means of knowledge transfer. Corpus-based research on the description and automated extraction of such statements has produced results for a variety of languages, e.g. English (Pearson, 1998; Meyer, 2001; Muresan and Klavans 2002; Marshman; 2008), French (Malaisé et al., 2005), Spanish (Sierra et al., 2008), German (Storrer and Wellinghoff, 2006; Walter, 2010; Cramer, 2011), Slovenian (Fišer et al., 2010), "Slavic" (Przepiórkowski et al., 2007), Portuguese (Del Gaudio and Branco, 2007) and Dutch (Fahmi and Bouma, 2006; Westerhout, 2009). These studies describe linguistic properties of defining statements, lexico-syntactic patterns or extraction grammars. Not all of them report results of extraction experiments, but many of the papers that do so combine linguistically informed extraction methods with machine learning or heuristic filtering methods.

Only few studies, however, provide a systematic description of the gold standard annotation process (with Walter, 2010, and Cramer, 2011, being notable exceptions), although the identification of defining statements is a non-trivial issue and reliable data is needed for the comparison of experimental results. Moreover, descriptions of the linguistic properties of defining statements, including statistical studies, seem to be largely missing, while results of small-scale studies suggest that the amount of variation in empirical data is not appropriately depicted by the literature (Walter, 2010).

In this paper, we focus on the description of the gold standard annotation process for two languages, namely Russian and German. For Russian, research in the field is still restricted to isolated efforts, whereas for German different kinds of definitions (Walter, 2010, studies legal definitions whereas Cramer, 2011, focuses on lay definitions from web corpora) have been studied. We also provide information concerning the linguistic annotation of the gold standard data and linguistic analyses aimed at revealing typical linguistic properties of knowledge-rich contexts.

2 Knowledge-Rich Contexts and Definitions

Knowledge-rich contexts (KRCs) can be described as pieces of text that may be helpful in a conceptual analysis task. Such tasks are usually performed in the context of terminology work and translation which constitute the main application area of the present work. Examples 1 and 2 present KRCs found in our data.

For a more formal definition of KRCs, it is important to consider that KRC extraction is related to the development of terminological knowledge bases (Meyer et al., 1992) and concept systems. These systems stress the relevance of semantic relations holding between concepts. Consequently, KRC extraction aims at identifying contexts for specialised terms that provide semantic information about the underlying concepts, including information about semantic relations between concepts (see ISO 1087-1: 2000). Moreover, KRCs are related to a set of minimal validity criteria that, however, are less strict than the criteria applied to definitions.

In practice, the boundary between definitions and KRCs is not always clear. Several of the

Proceedings of the Student Research Workshop associated with RANLP 2013, pages 134–141,
Hissar, Bulgaria, 9-11 September 2013.

1) Альтернативный источник энергии — способ, устройство или сооружение, позволяющее получать электрическую энергию (или другой требуемый вид энергии) и заменяющий собой традиционные источники энергии, функционирующие на нефти, добываемом природном газе и угле.
[An alternative source of energy is a method, a machine or a construction that enables the production of electrical (or of another, necessary kind of) energy, thus substituting traditional sources of energy based on oil, natural gas or coal.]

2) Das Verhältnis Energieertrag („Output“) zu Input wird Leistungszahl genannt.
[The relation between energy output and input is called coefficient of performance.]

above-mentioned studies employ the term “definition”, whereas the types of “definitions” subsumed under this term vary considerably. For our own work, we assume that definitions are subtypes of KRCs which echo the categories of “proper definition”, “redundant definition”, “complete definition” and “partial definition” as introduced by Bierwisch and Kiefer (1969) while covering a larger set of semantic relations, e.g. those relations that are relevant to terminological tasks, and satisfying less strict formal criteria.

3 Gold Standard Creation

The gold standard was created in three steps:

- In a *first step*, corpora were collected and KRC candidates were manually selected for annotation. Subcorpora were created to contain annotated KRCs.
- In a *second step*, more KRC candidates were selected from the subcorpora and annotated.
- In a *third step*, the gold standard was consolidated by applying qualitative criteria to the output of the previous two annotation steps.

3.1 Corpus Collection

Russian and German web corpora were crawled using the Babouk corpus crawling engine (de Groc, 2011). The web was chosen as our source of data since for many languages and specialised topics it offers a yet fairly unassessed wealth of data that can hardly be provided by traditional offline resources. Moreover, language workers use online resources extensively while the internet itself, given its known properties such as redundancy and noisiness (Fletcher 2004), has not yet been evaluated with respect to its usefulness for conceptual analysis tasks. Table 1 gives an overview over the Babouk corpora. The german_dev corpus was created within the TTC project[1].

Corpus	Domains	Tokens
russian_dev	cars	~350,000
russian_test	nuclear energy, cars, physics, …	~1,010,000
german_dev	wind energy	~990,000
german_test	IT, alternative energy sources, energy supply	~7,270,000

Table 1: Web corpora crawled with Babouk

From these corpora, KRC candidates (full sentences) were selected by the author, a trained translator, by manually inspecting a part of the texts in each corpus. The selection criteria were:

- the candidate must contain potentially relevant information for a conceptual analysis task,
- it must embody at least one of the following semantic relations: hyperonymy/hyponymy, meronymy, process, position, causality, origin, function, reference,
- at least one target term (a definiendum) can be identified as argument of one of the above-mentioned semantic relations,
- the information provided by the candidate must be currently valid (use of present tense) or temporal restrictions must be clearly marked,
- the candidate must at least be roughly attributable to one domain of interest,
- the information provided by the candidate must be generalisable or shed light on one interesting aspect of the definiendum.

[1] www.ttc-project.eu. The word counts were obtained from the linux wc function on the raw corpora.

Each candidate KRC together with at least one previously annotated definiendum candidate was then presented to two independent annotators, namely Master students of translation. Each annotator was a native speaker of the respective language and had been acquainted with the established validity criteria during an introductory seminar. Annotators were asked to give a simple binary assessment of the KRC status of each KRC candidate given the above validity criteria. For positive judgements, annotators were also asked to give a simple binary assessment of their annotation confidence (1 = "not very confident", 2 = "confident", hence the interval of average confidence for each annotator ranges between 1 and 2). Table 2 summarizes the results of this step by giving acceptance rates and average confidence for each annotator and corpus. Under "agreement", the table also summarises absolute and relative values for agreement on KRC validity judgements and confidence agreement (agreement on "high" and "low" confidence for a given candidate) for those KRCs in the gold standard that were marked "valid" by both annotators. Based on the results of this step, small sub-corpora were extracted from the web corpora to contain the KRC candidates agreed upon by all annotators.

3.2 Annotation Refinement

To achieve maximum coverage of the KRC annotation in the sub-corpora, we manually went through all four sub-corpora again to identify KRC candidates that may have been missed in the first candidate selection step. These new candidates were passed to four new annotators – two native speakers and experienced translators for each language – along with the same annotation criteria. This step resulted in the data summarised in table 3.

3.3 Discussion and Final Gold Standard Creation

Bierwisch and Kiefer (1969) are among the first to point out that linguistic criteria do not fully explain whether a statement can be considered defining or not. Cramer (2011) conducts extensive definition annotation experiments, concluding that the annotators' individual stance towards a candidate statement and the corresponding text, knowledge of the domain and other criteria influence whether a statement is considered defining. For a terminological setting, this is problematic, since these characteristics can be controlled only if the target users are known (e.g. in a small company setting, but not in the case of an online termbase).

The results of our own (small) annotation experiment seem to support Cramer's (2011) claim that individual criteria of the annotators influence the annotation process, resulting in different rates of acceptance/rejection and varying levels of confidence as summarised in tables 2 and 3: Although all annotators marked the vast majority of the KRC candidates presented to them as "valid", average confidence varies considerably between annotators, but also between corpora and annotation cycles. The different confidence levels and acceptance rates of the individual annotators indeed suggest that annotators develop individual annotation strategies while sudden confidence jumps (or drops) with, however, stable acceptance rates may be the result of changes in these strategies that, however, cannot be linked directly to linguistic criteria. Agreement seems to be generally higher in the first annotation cycle for both Russian and German which may be an effect of a more admissive pre-selection of candidates for the second cycle resulting in a potentially lower quality of candidates. The slightly, but consistently higher values achieved for russian_test in comparison to russian_dev may be an effect of the less 'technical' material in this corpus, since russian_dev contains a considerable amount of instructional texts which may not suit the annotators' expectations.

κ scores, if computed on the data, are low, however, it seems questionable whether they are applicable to this voting task in which no clearly negative examples were presented to the annotators. Moreover, it is unclear which κ level would be acceptable for a task as complex and fuzzy as this one. Finally, the small number of annotators (1 for the complete sub-corpora, 2 more for each pre-selected KRC candidate) does not allow for statistical generalisations concerning the KRC status of the annotated candidates. Given these reasons, we decided to apply qualitative criteria in order to improve the consistency of the data, e.g. by spotting false negatives (KRC candidates wrongly marked as "invalid" by at least one annotator) and false positives (KRC candidates wrongly marked as "valid" by the annotators). For example, we removed KRC candidates from the gold standard that had been annotated more than once, that turned out to be not compliant with the validity criteria, were longer than one sentence or that

Corpora	Annotators				Agreement	
	De1		*De2*			
	proportion of KRC candidates marked as "valid"	average confidence	proportion of KRC candidates marked as "valid"	average confidence	agreement on positive and negative judgements	agreement on high and low confidence
german_dev	347 (93%)	1.66	341 (92%)	1.84	326 (88%)	185 (68%)
german_test	290 (97%)	1.70	263 (88%)	1.83	262 (88%)	162 (70%)
	Ru1		*Ru2*			
russian_dev	289 (97%)	1.98	294 (98%)	1.83	290 (97%)	198 (83%)
russian_test	229 (100%)	1.99	225 (98%)	1.85	225 (98%)	159 (90%)

Table 2: Results of the first annotation cycle

Corpora	Annotators				Agreement	
	De3		*De4*			
	proportion of KRC candidates marked as "valid"	average confidence	proportion of KRC candidates marked as "valid"	average confidence	agreement on positive and negative judgements	agreement on high and low confidence
german_dev	63 (79%)	1.71	66 (83%)	1.50	51 (64%)	21 (46%)
german_test	45 (82%)	1.53	45 (82%)	1.51	41 (75%)	18 (50%)
	Ru3		*Ru4*			
russian_dev	64 (88%)	1.80	64 (88%)	1.59	65 (89%)	27 (63%)
russian_test	99 (94%)	1.86	102 (97%)	1.75	98 (93%)	67 (80%)

Table 3: Results of the second annotation cycle.

exhibited strongly erroneous language. With respect to boundary cases or linguistic defects of the KRCs, the resulting gold standard seems to be rather inclusive. Table 4 summarises the finalised gold standard.

Corpus	Tokens	KRCs
sub_german_dev	~ 160,000	337
sub_german_test	~ 170,000	295
sub_russian_dev	~ 99,000	292
sub_russian_test	~ 75,000	268

Table 4: Overview over finalised gold standard[2].

3.4 Coverage of the Annotation

Since one of the aims of the annotation was to achieve maximum coverage of identified KRCs in the gold corpora, we estimated the percentage of inadvertently missed KRCs in each sub-corpus, that is, we estimated an error rate based on KRC candidate misses. To this end, we randomly selected 500 sentences from each sub-corpus and assessed them with respect to their KRC status (given the validity criteria): Identified KRCs were counted as wanted hits, non-KRCs as wanted misses. Potential KRCs that had not been included in any of the annotation cycles were counted as unwanted misses. Based on these analyses, we calculated the proportion of unwanted misses along with 95% confidence intervals on each sub-corpus (see Sachs and Hedderich, 2009). The maximum proportion resulted to be of 0.02 (10 sentences on sub_german_test), resulting in a confidence interval of [0.0096, 0.0365]. We conclude that the proportion of unidentified (and thus unannotated) KRC candidates in our data is unlikely to be above 4% and therefore lies within still acceptable limits.

4 Corpus Annotation

The corpora crawled by Babouk come as plain text files along with separate XML headers containing metadata such as the online source of the text, seed terms used for crawling and the date when the text was extracted from the web. We performed preprocessing and linguistic annotation of the gold standard corpora and then formatted the data in XML. In a first step, we

[2] Word counts were obtained again with the linux wc function after sentence splitting.

used the Perl Lingua::Sentence module[3] for splitting the Russian and German corpora into single sentences. Exact duplicate sentences were removed with a simple Perl script. On all subcorpora, we performed POS tagging, lemmatisation and dependency parsing. Tagging and lemmatisation was performed for Russian using TreeTagger (Schmid, 1994) along with the tagset developed by Sharoff et al. (2008)[4]. For parsing Russian we used the model and pipeline for MaltParser (Nivre et al., 2007) provided by Sharoff and Nivre (2011). For the linguistic annotation of the German corpora we used the Mate toolsuite (Bohnet, 2010).

A simple XML format was developed for all Russian and German corpora. In this format, each token is annotated with the linguistic information outputted by the analysis tools. Moreover, a boolean attribute "isterm" is used to indicate whether a token matches one of the definienda identified as target terms during the gold standard annotation process for each corpus. KRCs identified during the annotation process are kept in tab-separated files together with their respective definienda and the annotators' confidence votes.

5 Linguistic Analyses

5.1 Method

Linguistic analyses of the gold standard KRCs were performed in order to arrive at a description of the specific linguistic properties of the KRCs. More specifically, we studied frequencies of different phenomena comparing the KRC data with an equal amount of randomly selected non-KRCs from the gold standard corpora as well as with frequencies from two non-specialised web corpus samples, a 2011 news crawl from the Leipzig corpus portal for German (NCL, Quasthoff et al., 2006) and an older version of the Russian internet corpus (RIC, Sharoff, 2006). We believe that with this double comparison we can distinguish between differences that occur between texts with a different level of specialisation (gold vs. RIC and gold vs. NCL) and differences that mark a stable feature of our gold data as compared to non-KRCs (KRCs vs. non-KRCs from the gold corpora). The Chi-Square and Fisher Tests were used to test for differences between the datasets. We used 95% confidence intervals for estimating the size of the differences between observed proportions, as suggested by Baroni and Evert (2008).

5.2 Results

Since results can be presented here only summarily due to space restrictions, we focus on observations on the levels of lexis and morphology. On the lexical level, we studied *POS* and *lemma frequencies.* Table 5 summarises the POS tags for which distributional differences were found between the Russian KRCs and both the RIC sample and the random non-KRCs from the Russian gold standard corpora while the numbers given are those for the comparison between gold standard and RIC. The tagset used is "Russian small"[5].

Tag	**Prop. KRCs**	**Prop. RIC**	χ^2	**p**	**CI**
S	0.439	0.365	112.20	< 0.01	[0.06, 0.09]
A	0.196	0.109	283.21	< 0.01	[0.08, 0.10]
ADV	0.013	0.032	76.75	< 0.01	[-0.02, -0,01]
PART	0.006	0.029	156.02	< 0.01	[-0.03, -0,02]
ADV-PRO	0.003	0.013	61.47	< 0.01	[-0.01, -0.01]
PRAE-DIC	0.001	0.006	38.74	< 0.01	[-0.01, 0]

Table 5: Results for comparison of POS frequencies Russian gold standard vs. RIC.

The table summarises proportions on the two corpora, chi-square and p-values as well as the 95%-confidence interval for the difference between proportions as outputted by the R[6] function prop.test().

On the level of *lemmata*, the same analysis showed that certain general nouns such as *вид* ("type", "kind") and *совокупность* ("the whole") for Russian or *Begriff* ("concept"), for German, were found significantly more often in the gold standard, whereas qualifying adjectives (*новый*, "new", *gut*, "good") and sentential adverbs (*даже*, "even", *nur*, "only") appear with a significantly lower frequency in the gold data.

[3] http://search.cpan.org/~achimru/Lingua-Sentence-1.00/lib/Lingua/Sentence.pm.

[4] The tagging model is available from: http://corpus.leeds.ac.uk/mocky/russian.par.gz.

[5] http://corpus.leeds.ac.uk/mocky/.

[6] http://www.r-project.org/.

Category	Prop. KRCs	Prop. RIC	χ^2	p	CI
perfective aspect	0.2168	0.6298	408.0662	< 0.0001	[-0.4498, -0.3762]
imperfective aspect	0.7814	0.3679	408.6745	< 0.0001	[0.3767, 0.4503]
imperative	0.0091	0.0195	3.7124	0.0540	[-0.0206, -0.0001]
passive	0.2168	0.0990	62.3199	< 0.0001	[0.0876, 0.1480]
infinitive	0.0747	0.1902	65.8157	< 0.0001	[-0.1429, -0.0881]
participle	0.0719	0.1504	35.2414	< 0.0001	[-0.1041, -0.0528]
first person	0.0009	0.0694	74.3668	< 0.0001	[-0.0833, -0.0536]
second person	0.0109	0.0366	15.1299	0.0001	[-0.0385, -0.0129]
third person	0.5383	0.3157	119.5548	< 0.0001	[0.1828, 0.2624]
present tense	0.7058	0.4170	198.2847	< 0.0001	[0.2499, 0.3278]
past tense	0.1949	0.3110	41.1045	< 0.0001	[-0.1514, -0.0807]
future tense	0.0118	0.0624	38.8885	< 0.0001	[-0.0661, -0.0350]
singular	0.5501	0.5090	3.8523	0.0497	[0.0001, 0.0821]

Table 6: Distributional differences of morphological markers between verbs in Russian KRCs and RIC.

Russian also shows fewer occurrences of modals (e.g.*должен*, "he must" and *мочь*, "may, can").

In another step, we studied *morphological properties* of verbs in the KRC samples in comparison, again, to similarly-sized samples from the reference web corpora (NCL for German, RIC for Russian) and samples of non-KRCs from the gold corpora. To this end, we analysed the morphological tags outputted by TreeTagger (for Russian) and mate (for German). The categories for which both comparisons gave significant results on Russian are summarised in table 6. The analysis shows that verbs in Russian KRCs are more often in imperfective aspect, passive voice and third person present tense. Less frequently in the gold standard we find imperative forms, verbal infinitives (maybe due to a lack of modals that need to be followed by an infinitive, see above) and participles. As previously, the German data echoes these results. A manual analysis of the *syntactic realisation* of the predicates in the KRCs gave evidence that Russian "unpersonal-definite" constructions (subjectless sentences with a verb in third person plural serving as predicate) and German presentatives may be light indicators for KRCs.

5.3 Discussion

Our results on the *lexical* level amount to a tendency towards an unpersonal style exhibited by KRCs in both languages. On the other hand, typical elements of defining statements (e.g. generalising adverbs or mentions of specific disciplines) that are described in the literature could not be found in high quantity. Obviously, larger datasets are necessary for an in-depth study of the lexical properties of KRCs. The *morphological properties* of verbs in the KRCs seem to support our hypothesis of an unpersonal, fact-oriented style, while imperfective aspect, present tense, presentatives and subjectless sentences can be understood as generalisation signals.

6 Conclusions and Future Work

In this paper, we proposed a methodology for the task of annotating a gold standard for KRC extraction. Our analysis suggests that decisions concerning the KRC-status of candidate statements are influenced by a range of factors that are not related to the linguistic surface of the KRC candidates themselves. Clearly, more empirical research on text-based knowledge acquisition is needed to arrive at more adequate models. The annotations carried out in the course of this study are transparent in that annotators' judgements can be used as hints for a more detailed study of boundary cases or external influencing factors. Nevertheless, further annotation work should use linguistic features of defining statements as optional signal. Our analysis of linguistic properties of KRCs supports hypotheses found in the literature, but also indicates that other, frequently described properties occur only rarely. Future work will deal with the question whether more linguistic information can improve KRC extraction.

Acknowledgement

The work described in this paper was partly funded by the CLARA project (EU/FP7), grant agreement n° 238405. I am also grateful to my anonymous reviewers for their helpful remarks. Last but not least, I am indebted to my colleagues José Martinez Martinez and Ekaterina Lapshinova-Koltunski for many interesting discussions and suggestions.

References

Baroni, M. and Evert, S. 2008. "Statistical methods for corpus exploitation". In Lüdeling, A. and Kytö, M. (eds), *Corpus Linguistics. An International Handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft 29). Mouton de Gruyter, Berlin.

Bierwisch, M. and Kiefer, F. 1969. "Remarks on Definitions in Natural Language". In Kiefer, F. (ed), *Studies in Syntax and Semantics.* (Foundations of Language 10). Reidel, Dordrecht.

Bohnet, B. 2010. "Top Accuracy and Fast Dependency Parsing is not a Contradiction". COLING 2010, August 23-27, Beijing, China: 89-97. Available online at http://www.aclweb.org/anthology/C/C10/C10-1011.pdf.

Cramer, I. M. 2011. *Definitionen in Wörterbuch und Text: Zur manuellen Annotation, korpusgestützten Analyse und automatischen Extraktion definitorischer Textsegmente im Kontext der computergestützten Lexikographie.* Published PhD thesis. Technical University Dortmund. Available online at https://eldorado.tu-dortmund.de/bitstream/2003/27628/1/Dissertation.pdf.

de Groc, C. 2011. "Babouk: Focused web crawling for corpus compilation and automatic terminology extraction". IEEE/WIC/ACM: International Conference on Intelligent Agent Technology, August 22-27. Lyon, France: 497-498.

Del Gaudio, R. and Branco, A. 2007. "Automatic Extraction of Definitions in Portuguese: A Rule-Based Approach". In Neves, J., Santos, M.F. and Machado, J.M. (eds.) *Progress in Artificial Intelligence.* (Lecture Notes in Artificial Intelligence 4874). Springer, Berlin.

Fahmi, I. and Bouma, G. 2006. "Learning to Identify Definitions using Syntactic Features". Workshop on Learning Structured Information in Natural Language Applications at EACL 2006, April 3, Trento, Italy: 64-71. Available online at http://ai-nlp.info.uniroma2.it/eacl2006-ws10/WS10-eacl2006-proceedings.pdf.

Fišer, D., Pollak, S. and Vintar, Š. 2010. "Learning to Mine Definitions from Slovene Structured and Unstructured Knowledge-Rich Resources". LREC 2010, May 19-21, Valletta, Malta: 2932-2936. Available online at http://www.lrec-conf.org/proceedings/lrec2010/pdf/141_Paper.pdf.

Fletcher, W. H. 2004. "Making the Web More Useful as a Source for Linguistic Corpora". In Connor, U. and Upton, T. A. (eds), *Applied Corpus Linguistics. A Multidimensional Perspective*. Rodopi, Amsterdam/New York.

International Organization for Standardization. 2000. International Standard ISO 1087-1: 2000 – Terminology Work – Vocabulary – Part 1: Theory and application. ISO, Geneva.

Malaisé, V., Zweigenbaum, P. and Bachimont, B. 2005. "Mining defining contexts to help structuring differential ontologies". *Terminology 11 (1)*: 21-53.

Marshman, E. 2008. "Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in English and French specialized texts". *Terminology 14 (1)*: 124-151.

Meyer, I. 2001. "Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework". In Bourigault, D., Jacquemin, C. and L'Homme, M.-C. (eds), *Recent Advances in Computational Terminology.* (Natural Language Processing 2). John Benjamins. Amsterdam/Philadelphia.

Meyer, I., Skuce, D., Bowker, L. and Eck, K. 1992. "Towards a New Generation of Terminological Resources: An Experiment in Building a Terminological Knowledge Base". COLING 1992, August 23-28, Nantes, France: 956-960. Available online at http://acl.ldc.upenn.edu/C/C92/C92-3146.pdf.

Muresan, S. and Klavans, J. 2002. "A Method for Automatically Building and Evaluating Dictionary Resources". LREC 2002, May 29-31, Las Palmas, Spain: 231-234. Available online at http://www.lrec-conf.org/proceedings/lrec2002/.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S. and Marsi, E. 2007. "MaltParser: A language-independent system for data-driven dependency parsing". *Natural Language Engineering* 13(2): 95-135.

Pearson, J. 1998. *Terms in Context*. (Studies in Corpus Linguistics 1). John Benjamins, Amsterdam/Philadelphia.

Przepiórkowski, A., Degórski, Ł., Spousta, M., Simov, K., Osenova, P., Lemnitzer, L., Kuboň, V. and Wójtowicz, B. 2007. "Towards the automatic extraction of definitions in Slavic". BSNLP workshop at ACL 2007, June 29, Prague, Czech Republic: 43-50. Available online at

http://langtech.jrc.it/BSNLP2007/m/BSNLP-2007-proceedings.pdf.

Quasthoff, U., Richter, M. and Biemann, C. 2006. "Corpus Portal for Search in Monolingual Corpora". LREC 2006, May 24-26, Genoa, Italy: 1799–1802. Available online at: http://www.lrec-conf.org/proceedings/lrec2006/.

Sachs, L. and Hedderich, J. 2009. *Angewandte Statistik. Methodensammlung mit R.* Springer, Berlin/Heidelberg.

Schmid, H. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". International Conference on New Methods in Language Processing, Manchester, England: 44–49. Available online at: ftp://ftp.ims.uni-stuttgart.de/pub/corpora/tree-tagger1.pdf.

Sharoff, S. 2006. "Creating general-purpose corpora using automated search engine queries". In M. Baroni and S. Bernardini (eds) *WaCky! Working papers on the Web as Corpus*. Gedit. Bologna. Available online at: http://www.comp.leeds.ac.uk/ssharoff/publications/wacky-paper.pdf.

Sharoff, S., Kopotev, M., Erjavec, T., Feldmann, A. and Divjak, D. 2008. "Designing and evaluating a Russian tagset". LREC 2008, May 28-30, Marrakech, Morocco: 279-285. Available online at http://www.lrec-conf.org/proceedings/lrec2008/.

Sharoff, S. and Nivre, J. 2011. "The proper place of men and machines in language technology. Processing Russian without any linguistic knowledge." Dialogue 2011, May 25-29, Bekasovo, Russia: 591-604. Available online at http://www.dialog-21.ru/digests/dialog2011/materials/ru/pdf/58.pdf.

Sierra, G., Alarcón, R., Aguilar, C. and C. Bach. 2008. "Definitional verbal patterns for semantic relation extraction". *Terminology 14 (1)*: 74-98.

Storrer, A. and Wellinghoff, S. 2006. "Automated detection and annotation of term definitions in German text corpora". LREC 2006, May 24-26, Genoa, Italy: 2373-2376. Available online at http://www.lrec-conf.org/proceedings/lrec2006/.

Walter, S. 2010. *Definitionsextraktion aus Urteilstexten*. Published PhD thesis. Saarland University. Available online at: http://www.coli.uni-saarland.de/~stwa/publications/DissertationStephanWalter.pdf.

Westerhout, E. 2009. "Definition Extraction Using Linguistic and Structural Features". First Workshop on Definition Extraction, Borovets, Bulgaria: 61-67. Available online at http://www.aclweb.org/anthology-new/W/W09/W09-4410.pdf.