

# On the evaluation of Polish definition extraction grammars

Adam Przepiórkowski, Łukasz Degórski, Beata Wójtowicz

Polish Academy of Sciences, Institute of Computer Science  
ul. Ordona 21, 01-237 Warsaw, Poland  
adam@pipan.waw.pl, {ldegorski, beataw}@bach.ipipan.waw.pl

## Abstract

This paper presents the results of experiments in the automatic extraction of definitions (for semi-automatic glossary construction) from usually unstructured or only weakly structured e-learning texts in Polish. The extraction is performed by regular grammars over XML-encoded morphosyntactically-annotated documents. The results, although perhaps still not fully satisfactory, are carefully evaluated and compared to the inter-annotator agreement; they clearly improve on previous definition extraction attempts for Polish.

## 1. Introduction

The aim of this paper is to report on experiments in the automatic extraction of those fragments of Polish text which could be used as definitions of terms (whether they were written as definitions or not) and to discuss possible ways of evaluating the results of such experiments.

The context of the work reported here is a European (IST) Specific Targeted Research Project aiming at constructing various language technology tools for eLearning, for a number of languages, including Polish.<sup>1</sup> The input to the definition extraction module is XML-encoded<sup>2</sup> morphosyntactically-annotated text, so the definition extraction module must be XML-aware and the effect of the operation of the module should consist in adding XML elements marking the defined term (the *definiendum*) and the defining text (the *definiens*). Such automatically extracted term definitions are to be presented to the author or the maintainer of the Learning Object (LO; i.e., course materials) and, thus, significantly facilitate and accelerate the creation of a glossary for a given LO. From this specification of the task it follows that good recall is much more important than good precision, as it is easier to reject wrong glossary candidates than to browse the LO for term definitions which were not automatically spotted.

The structure of the paper is as follows: §2. mentions previous work on definition extraction, §3. presents a shallow grammar developed for definition extraction from Polish texts, §4. describes the evaluation experiments and their results, §5. compares these results to the state of the art, §6. discusses the inherent difficulty of the task, and §7. concludes the paper.

## 2. Related Work

Definition extraction is an important NLP task, most frequently a subtask of terminology extraction (Pearson, 1996), the automatic creation of glossaries (Klavans and Muresan, 2000; Westerhout and Monachesi, 2007), question answering (Miliaraki and Androutsopoulos, 2004; Fahmi and Bouma, 2006), learning lexical semantic relations (Malaisé et al., 2004; Storrer and Wellinghoff,

2006) and automatic construction of ontologies (Walter and Pinkal, 2006). Tools for definition extraction are invariably language-specific and involve shallow or deep processing, with most work done for English (Pearson, 1996; Klavans and Muresan, 2000; Saggion, 2004) and other Germanic languages (Fahmi and Bouma, 2006; Storrer and Wellinghoff, 2006; Walter and Pinkal, 2006; Westerhout and Monachesi, 2007), as well as French (Malaisé et al., 2004). (Some comparison to this earlier work is made in §5.)

There is very little previous work on definition extraction in Slavic languages, with the exception of some work on Bulgarian reported in Tanev 2004 and Simov and Osenova 2005, and a recent article on Bulgarian, Czech and Polish (Przepiórkowski et al., 2007). In all cases shallow grammars were constructed for the identification of definitions in texts, but only Przepiórkowski et al. 2007 contains the evaluation of these grammars in terms of precision, recall and  $F_2$  measure.<sup>3</sup> Our results improve on the results for the best grammar presented in Przepiórkowski et al. 2007, namely, the grammar for Czech, and far exceed their results for Polish.

## 3. Regular XML Grammars of Definitions

The complete corpus of instruction texts with manually annotated definitions was split into three parts: a training corpus, a held-out corpus and a testing corpus. The training corpus and the testing corpus consist of 12 different instruction texts each, so they are reasonably well balanced. The held-out corpus, on the other hand, is a specific homogeneous text, namely, the Calimera guidelines (<http://www.calimera.org/>). The quantitative characteristics of these corpora, in particular the number of manually annotated definitions, including the number of definitions split across two or (in one case) three sentences, is given in Table 1. The first version of the grammar, GR, was developed on the basis of the training corpus, in many (well over 100) iterations, where in each iteration the grammar was improved and the results were evaluated both quantitatively (automatically) and qualitatively (manually). This grammar was then applied to held-out data, and subsequently modified, in relatively few iterations (around 5), to increase precision and recall measured

<sup>1</sup>Details of the project are withheld for reasons of anonymity; they will be provided in the final version of the paper.

<sup>2</sup>More precisely, the input adheres to the XML Corpus Encoding Standard (Ide et al., 2000).

<sup>3</sup> $F_\alpha = (1 + \alpha) \cdot (\text{precision} \cdot \text{recall}) / (\alpha \cdot \text{precision} + \text{recall})$ .

on held-out data. This resulted in grammar GR'. Both GR and GR' were then evaluated on the testing corpus, which was not consulted during grammar development and, in fact, was not known to grammar developers at all.<sup>4</sup>

	training	held-out	testing	TOTAL
tokens	139 039	84 288	77 309	300 636
sentences	5 218	2 263	3 349	10 830
definitions	304	82	172	558
(incl. split)	21	9	24	54

Table 1: Corpora used for the development of grammars

The input to the current task of definition extraction is XML-encoded morphosyntactically-annotated<sup>5</sup> text adhering to the XML Corpus Annotation Standard (XCES; Ide et al. 2000). For example, the representation of a Polish sentence starting as *Konstruktywizm kładzie nacisk na* (Eng. ‘Constructivism places (the) emphasis on’) may look as follows:<sup>6</sup>

```
<s id="s9">
<tok base="konstruktywizm" msd="sg:nom:m3"
  ctag="subst" id="t3">Konstruktywizm</tok>
<tok base="kłaść" msd="sg:ter:imperf"
  ctag="fin" id="t4">kładzie</tok>
<tok base="nacisk" msd="sg:acc:m3"
  ctag="subst" id="t5">nacisk</tok>
<tok base="na" msd="acc"
  ctag="prep" id="t6">na</tok>
...
<tok base="." ctag="interp" id="t17">.</tok>
</s>
```

The grammar is a regular grammar implemented with the use of the `lxtransduce` tool (Tobin, 2005), a component of the LTXML2 toolset developed at the University of Edinburgh. An example of a simple rule for prepositional phrases (PPs) is given below:

```
<rule name="PP">
  <seq>
    <query match="tok[@ctag = 'prep']"/>
    <ref name="NP1">
      <with-param name="case" value="'" />
    </ref>
  </seq>
</rule>
```

This rule identifies a sequence whose first element is a token tagged as a preposition and whose subsequent elements are identified by a rule called NP1. This latter rule (not shown here for brevity) is a parameterised rule which finds a nominal phrase (NP) of a given case, but the way it is called above (`value="'"`) ensures that it will find an NP of any case.

The grammar GR contains 44 rules, in a 14K file, most of them more complex than the PP rule above (the average size of a rule is 11.8 lines, compared with 8 lines above).

<sup>4</sup>We intend to provide URLs to the grammars and the respective corpora in the final version of the paper.

<sup>5</sup>A slightly modified version of the tagger presented in Piasecki and Godlewski 2006 was used for this purpose.

<sup>6</sup>Some of the representation has been replaced by ‘...’.

The grammar is split into 4 layers, with rules of each layer possibly calling only rules of the same and previous layers. The first layer contains low-level rules making reference to particular orthographic forms, such as rules correcting the results of the tagger (e.g., a rule gluing back *e* and *-learning* incorrectly split by the tagger), rules finding various textual realisations of the expression corresponding to the English ‘that is’ or ‘namely’ (i.e., *to jest*, *tj.*, etc.), but also rules identifying Polish copulae (*być*, *to*) and a rule identifying verbs characteristic of defining sentences (called *definitor verbs* in Storrer and Wellinghoff 2006), e.g., *oznaczac* ‘signify’, *określać* ‘depict’, *obejmować* ‘comprise’, etc.

The second layer contains linguistically justified rules identifying nouns, NPs, PPs, etc.

The third layer contains various important auxiliary rules, e.g., a rule finding a possible term, i.e., an NP, possibly followed by any number of genitive NPs, PPs or parenthetical expressions, perhaps intermingled.

Finally, the fourth layer contains 12 top-level rules, corresponding to various types of definitions, e.g., two rules for copular definitions (one for the nominative pattern and one for the instrumental pattern), rules for various kinds of parenthetical definitions (one based on parentheses, another for *that is* expressions), a rule for structural definitions with the use of a colon, etc.

After evaluating the grammar on the held-out data, the grammar was extended to GR', containing 13 top level rules (with 48 rules in total, in a 16K file, 12.5 lines for a rule, on the average).

## 4. Experiments and Results

Apart from grammars described above, GR and GR', we constructed three baseline grammars: B1, which marks all sentences as definition sentences; B2, which marks as definitory all sentences containing a possible copula (*jest*, *sq*, *to*), the abbreviation *tj.* ‘i.e.’, or the word *czyli* ‘that is’, ‘namely’; and B3, a very permissive grammar marking as definitions all sentences containing any of the 27 very simple patterns (in most cases, single-token keywords) manually identified on the basis of manually annotated definitions (these patterns include all patterns in B2, as well as various definitor verbs, apparent acronym specifications, the equal sign ‘=’, etc.).

For all five grammars (B1, B2, B3, GR and GR'), experiments were conducted on two corpora: the training corpus, consulted when developing the grammar, and a testing corpus, unseen by grammar developers.<sup>7</sup>

In each experiment, precision and recall were calculated at two levels: at the token level and at the sentence level, as both ways of the evaluation of definition extraction may be found in the literature.<sup>8</sup> At the token level,

<sup>7</sup>We do not provide results for the held-out corpus, as it is a homogeneous and rather unrepresentative corpus of definitions and, hence, the results are systematically worse than either for the training or for the testing corpus.

<sup>8</sup>Carletta 1996, p. 252, advocating the use of the kappa statistic in Computational Linguistics, draws attention to the importance of selecting appropriate units; we believe that it is sentences that are such appropriate units for the task at hand.

	P	R	F <sub>1</sub>	F <sub>2</sub>	F <sub>5</sub>
B1	4.49	100.00	8.59	12.36	22.00
B2	6.83	76.56	12.54	17.38	28.33
B3	6.61	97.29	12.37	17.45	29.59
<b>GR</b>	<b>16.84</b>	66.49	<b>26.87</b>	<b>33.53</b>	<b>44.58</b>
GR'	15.28	68.81	25.01	31.75	43.45

Table 2: Token-level evaluation on the training corpus

	P	R	F <sub>1</sub>	F <sub>2</sub>	F <sub>5</sub>
B1	5.37	100.00	10.19	14.54	25.39
B2	10.01	73.57	17.63	23.61	35.75
B3	9.79	94.64	17.75	24.34	38.72
<b>GR</b>	<b>23.66</b>	72.50	<b>35.68</b>	<b>42.95</b>	<b>53.94</b>
GR'	20.53	75.00	32.23	39.80	52.00

Table 3: Sentence-level evaluation on the training corpus

precision is understood as the number of tokens simultaneously belonging to a manual definition and an automatically found definition, divided by the number of tokens in automatically found definitions. Correspondingly, recall is the ratio of the number of tokens simultaneously in both definition types to the number of tokens in manual definitions. At the sentence level, a sentence is taken as a manual or automatic definition sentence if and only if it contains a (part of a), respectively, manual or automatic definition. Given that, precision and recall are calculated in a way analogous to token level precision and recall.

The results of all experiments are given in the four Tables 2–5, separately for each corpus and for each evaluation level. Apart from precision (P) and recall (R), also the usual F-measure is given (F<sub>1</sub>), as well as F<sub>2</sub> used in Przepiórkowski et al. 2007 and F<sub>5</sub> (apparently) used in Saggion 2004.<sup>9</sup> Note that for the task at hand, where recall is more important than precision, the latter two measures seem appropriate, although whether recall is twice as important as precision (F<sub>2</sub>) or five times as important (F<sub>5</sub>) is ultimately an empirical issue that should be settled by user case evaluation experiments.

## 5. Comparisons

While the recall of the definition extraction grammar described in §3. is much lower than that of the very permissive baseline grammars, all other measures, i.e., precision and various F-measures, show the clear improvement of the grammar over all three baselines, including the relatively sophisticated baseline grammar B3: according to the least favourable (for GR') F<sub>5</sub> measure, GR' is a 25% token-level improvement over B3 when calculated on the testing corpus, and 10% sentence-level improvement. The gain is much clearer when, e.g., F<sub>2</sub> is taken into account (62% and 35%, respectively).

The results presented above also constitute a clear improvement over earlier definition extraction attempts for West Slavic, reported in Przepiórkowski et al. 2007. The best results given in that paper concern Czech, for which

<sup>9</sup>It should, however, be noted that Saggion 2004 uses F<sub>5</sub> to evaluate definition answers to particular questions.

	P	R	F <sub>1</sub>	F <sub>2</sub>	F <sub>5</sub>
B1	4.96	100.00	9.45	13.53	23.83
B2	7.82	64.20	13.95	18.87	29.17
B3	8.21	91.49	15.07	20.88	34.00
GR	<b>19.27</b>	48.96	27.65	32.34	38.95
<b>GR'</b>	18.77	56.55	<b>28.18</b>	<b>33.84</b>	<b>42.34</b>

Table 4: Token-level evaluation on the testing corpus

	P	R	F <sub>1</sub>	F <sub>2</sub>	F <sub>5</sub>
B1	5.43	100.00	10.31	14.71	25.64
B2	9.69	61.54	16.74	22.11	32.53
B3	10.54	88.46	18.84	25.54	39.64
GR	<b>19.34</b>	51.65	28.14	33.18	40.40
<b>GR'</b>	18.69	59.34	<b>28.42</b>	<b>34.39</b>	<b>43.55</b>

Table 5: Sentence-level evaluation on the testing corpus

— at token-level evaluation<sup>10</sup> — precision, recall and F<sub>2</sub> of, respectively, 18.3%, 40.7% and 28.9 are reported, to be compared to our 18.77%, 56.55% and 33.84, as measured on the testing corpus. Perhaps more directly comparable are the results concerning Polish, where Przepiórkowski et al. 2007 report the precision of 14.8%, recall of 22.2% and F<sub>2</sub> of 19.0: the grammar presented here constitutes an improvement of 78%, as counted on the basis of F<sub>2</sub> for the testing corpus.<sup>11</sup>

Comparison to other work is more difficult for a number of reasons. First, some papers, especially, those reporting on definition extraction for Question Answering, use evaluation measures different than the ones assumed here. For example, Miliaraki and Androutsopoulos 2004 evaluate their system in terms of the percentage of questions that the system handles successfully (i.e., produces 5 snippets at least one of which contains an answer); a similar evaluation method, but taking into account the ranking of the answers and the position of the correct answer in that ranking, is assumed in Tanev 2004.

Second, various authors concentrate on the precision of their grammars, and do not provide recall or F-measure. For example, Fahmi and Bouma 2006 give the baseline precision of 59% for their Dutch system (so high “probably due to the fact that the current corpus consists of encyclopedic material only”; Fahmi and Bouma 2006, fn. 4), showing that this number may be raised to 75.9% after taking sentence position into account, and to 92.21% by applying maximum entropy classifiers. Similarly, Malaisé et al. 2004 mention the precision of 61–66%, while Walter and Pinkal 2006 cite 44.6–48.6%, depending on the eval-

<sup>10</sup>Przepiórkowski et al. 2007 do not present sentence-level evaluation results; instead they consider another way to calculate precision and recall, based on overlaps of manual and automatic definitions. For reasons of space we do not present here detailed results of this evaluation method, but measured this way our grammar achieves F<sub>2</sub> = 33.70 (measured on the testing corpus), compared to their 28.4 for Polish and 33.9 for Czech.

<sup>11</sup>Note also that their input text contained manually added annotations of keywords, many of which were natural candidates for defined terms. No such clues are present in the current experiments.

uator (with the numbers raised to 71.8–75.2% for a subset of 17–18 best rules from the 33-rule grammar). While we can only envy such precision results, it is not clear how the overall definition extraction results, as measured by the F-measures, compare to our system.

Third, in some of the remaining work, the evaluation method is not completely clear or a little different than in the current paper, or the evaluation is limited to a few types of definitions only. For example, Saggion 2004, participating in the TREC QA 2003 competition, measures (apparently)  $F_5$  separately for answers to different definition questions, giving the “combined F-score of 0.236”, but it is not clear if this F-score is directly comparable to our  $F_5$  of 0.4355 (for GR’, testing corpus). Similarly, Storrer and Wellinghoff 2006 give 34% precision and 70% recall for their German definition extraction system, although it is not fully clear whether these measures are token-based, sentence-based, or calculated in some other way, neither is it clear whether the evaluation was performed on previously unseen testing data. In any case, their system seems to be one of the best systems reported so far, probably exceeding our results.<sup>12</sup> On the other hand, Westerhout and Monachesi 2007 provide results directly comparable to our token-level results, albeit only for two most frequent and typical types of definitions: copular definitions (precision = 26%, recall = 72%) and definitions involving other connector verbs (precision = 44%, recall = 56%). We are planning to perform similar evaluation of particular definition types in the future.

## 6. Inter-annotator Agreement

The task of finding definitory contexts in instruction texts is relatively ill-defined: such definitory contexts often provide definitions of terms in implicit or indirect way, without any structural definitional clues. It is sometimes controversial whether a given sentence in fact provides a definitory context, as in the following sentence actually marked as a definition: *Użytkownicy Internetu spotykają i korzystają z portali, aby dotrzeć do poszukiwanych przez siebie informacji*<sup>13</sup> ‘Internet users meet and use portals in order to find the information they seek’. One way to measure the inherent difficulty of the task is to calculate the inter-annotator agreement (IAA).

Przepiórkowski et al. 2007 provide the IAA in terms of the usual Cohen’s kappa statistic,  $\kappa$ , calculated at token-level, i.e., the classification task is understood here as classifying each token as either belonging to a definitory context or not. Calculated this way, they report the IAA for Polish equal to 0.31, while our result is 0.259, i.e., in either case very low for any classification task, and especially low for a binary classification task; this confirms that definition extraction is intrinsically very difficult.

<sup>12</sup>If these are sentence-level results, as seems probable, the values of  $F_1$ ,  $F_2$  and  $F_5$  are, respectively, 45.77, 51.74 and 59.50, compared to our 35.68, 42.95 and 53.94 for GR and for the training corpus.

<sup>13</sup>This sentence actually contains a grammatical error, as the intransitive *spotykać się* is an inherently reflexive verb, while the sentence lacks the reflexive marker *się*.

There are at least two possible objections to this way of measuring IAA. First, token-based IAA seems to be flawed as it assumes a probabilistic model in which an annotator throws a weighed coin for each token, which results in many short (often one-token long) “definitions”, rather than a few long ones. In order to address this problem, we also calculated IAA at the sentence level, approximating definition extraction as a classification problem of marking a sentence as definitory or not. For that we used a subcorpus of the training corpus, i.e., over 83K-token parts of a popular book about computers, in which the original annotator marked 158 definitions (including 13 definitions split across 2 sentences). It turned out that the second annotator found as many as 595 definitory contexts (none split across sentences) in the same text! The comparison resulted in the very low *sentence-level*  $\kappa$  of 0.307, which confirms cross-linguistic *token-level* results of Przepiórkowski et al. 2007 and Westerhout and Monachesi 2007 ranging from 0.26 (Westerhout and Monachesi 2007, for one pair of annotators, for Dutch; cf. also our result for Polish given above), through 0.31 (Przepiórkowski et al. 2007, for Polish) and 0.42 (Westerhout and Monachesi 2007, for another pair of annotators, for Dutch), to 0.44 (Przepiórkowski et al. 2007, for Czech). Our contingency tables for both ways of calculating IAA are given in Tables 6–7.<sup>14</sup>

	def.	not def.	TOTAL
def.	1593	5702	7295
not def.	1740	74197	75937
TOTAL	3333	79899	83232

Table 6: Token-level contingency table

	def.	not def.	TOTAL
def.	127	419	546
not def.	39	2968	3007
TOTAL	166	3387	3553

Table 7: Sentence-level contingency table

Second, as discussed by Di Eugenio and Glass 2004, and also in the current context noted by Westerhout and Monachesi 2007, Cohen’s  $\kappa$  does not take into account prevalence and bias effects, both conspicuous in Tables 6–7: non-definitory tokens/sentence are strongly prevalent<sup>15</sup> and there is a strong bias between annotators, with one annotator classifying over twice as many tokens (over three times as many sentences) as definitory than the other. Various ways of dealing with this problem are controversial and much discussed in the medical literature; one relatively well-known (and also used in Westerhout and Monachesi 2007) measure is PABAK (prevalence-adjusted bias-adjusted kappa), which completely removes prevalence

<sup>14</sup>Note that some sentences contain 2 or more definitions, hence, the numbers of definitory *sentences* in Table 7 do not exactly correspond to the numbers of definitory *contexts* reported above.

<sup>15</sup>Depending on the annotator, non-definitory tokens constitute 91.2% or 96.0% of all tokens; for sentences the proportions are 84.6% or 95.3%.

and bias effects by replacing, in the contingency table, the values of the two (diagonal) agreement cells with their average, and the values of the two (off-diagonal) disagreement cells with their average, and then proceeding as in case of Cohen's  $\kappa$ . The token-level and sentence-level values of PABAK for our experiments are, respectively, 0.821 and 0.742.

However, we side with those authors who note that some effects of bias and prevalence on the magnitude of kappa are actually meaningful and consider PABAK on its own as "uninformative" (Sim and Wright, 2005, p.264). Instead, it is more instructive to compare  $\kappa$  with the maximum value  $\kappa$  could attain given the actual proportions of decisions by annotators (Dunn, 1989; Sim and Wright, 2005); in order to calculate such  $\kappa_{\max}$ , as much as possible weight should be moved from the disagreement cells in a contingency table to the agreement cells, but without changing the marginal totals. Such maximal  $\kappa$  values are 0.606 (token-level) and 0.425 (sentence-level), to be compared with 0.259 and 0.307, respectively. This comparison clearly demonstrates that the IAA agreement is much higher at the sentence level (0.307 out of 0.425) than at the token level (0.259 out of 0.606), which provides one more argument for evaluating definition extraction at sentence level, as common in Question Answering, rather than at token level, as in Westerhout and Monachesi 2007 or Przepiórkowski et al. 2007.

## 7. Conclusion

This paper presents the results of a series of experiments on definition extraction for Polish, an inherently very difficult task, which seem to be comparable to the state of the art. Unlike most other reports on such experiments, we carefully distinguish between training, held-out and testing data, with the testing corpus never, even indirectly, consulted by grammar developers. We also explicitly distinguish between token-level and sentence-level results, although the latter seem to be more relevant for the task at hand: a developer or maintainer of a Learning Object, when creating a glossary, should probably be presented with full sentences possibly containing definitory contexts, rather than shorter snippets of texts. Measured at this level, our system achieves 75% recall and over 20% precision on previously seen text, or almost 60% recall and almost 19% precision on new texts, and it compares favourably to previous Slavic definition extraction experiments, far exceeding previous results for Polish.

## References

Carletta, Jean, 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22:249–254.

Di Eugenio, Barbara and Michael Glass, 2004. The kappa statistic: A second look. *Computational Linguistics*, 30:95–101.

Dunn, G., 1989. *Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors*. London: Edward Arnold.

Fahmi, Ismail and Gosse Bouma, 2006. Learning to identify definitions using syntactic features. In *Proceedings of the EACL 2006 workshop on Learning Structured Information in Natural Language Applications*.

Ide, Nancy, Patrice Bonhomme, and Laurent Romary, 2000. XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Linguistic Resources and Evaluation Conference*. Athens, Greece.

Klavans, Judith L. and Smaranda Muresan, 2000. DEFINDER: Rule-based methods for the extraction of medical terminology and their associated definitions from on-line text. In *Proceedings of the Annual Fall Symposium of the American Medical Informatics Association*.

Malaisé, Véronique, Pierre Zweigenbaum, and Bruno Bachimont, 2004. Detecting semantic relations between terms in definitions. In Sophia Ananadiou and Pierre Zweigenbaum (eds.), *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*. Geneva, Switzerland.

Miliaraki, Spyridoula and Ion Androutsopoulos, 2004. Learning to identify single-snippet answers to definition questions. In *Proceedings of COLING 2004*. Geneva, Switzerland.

Pearson, Jennifer, 1996. The expression of definitions in specialised texts: a corpus-based analysis. In M. Gellerstam, J. Järborg, S. G. Malmgren, K. Norén, L.Rogström, and C. Pappmehl (eds.), *Proceedings of the Seventh Euralex International Congress*. Göteborg.

Piasecki, Maciej and Grzegorz Godlewski, 2006. Effective architecture of the Polish tagger. In *Proceedings of Text, Dialogue and Speech (TSD) 2006*.

Przepiórkowski, Adam, Łukasz Degórski, Miroslav Spousta, Kiril Simov, Petya Osenova, Lothar Lemnitzer, Vladislav Kuboň, and Beata Wójtowicz, 2007. Towards the automatic extraction of definitions in Slavic. In Jakub Piskorski, Bruno Pouliquen, Ralf Steinberger, and Hristo Tanev (eds.), *Proceedings of the BSNLP workshop at ACL 2007*. Prague.

Saggion, Horacio, 2004. Identifying definitions in text collections for question answering. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004*. Lisbon: ELRA.

Sim, Julius and Chris C Wright, 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85:257–268.

Simov, Kiril Ivanov and Petya Osenova, 2005. BulQA: Bulgarian-Bulgarian Question Answering at CLEF 2005. In *CLEF*.

Storrer, Angelika and Sandra Wellinghoff, 2006. Automated detection and annotation of term definitions in German text corpora. In *Proceedings of LREC 2006*.

Tanev, Hristo, 2004. Socrates: A question answering prototype for Bulgarian. In *Recent Advances in Natural Language Processing III, Selected Papers from RANLP 2003*. John Benjamins.

Tobin, Richard, 2005. *Lxtransduce, a replacement for fsmatch*. University of Edinburgh. <http://www.cogsci.ed.ac.uk/~richard/ltxml2/lxtransduce-manual.html>.

Walter, Stephan and Manfred Pinkal, 2006. Automatic extraction of definitions from German court decisions. In *Proceedings of the Workshop on Information Extraction Beyond The Document*. Sydney, Australia: Association for Computational Linguistics.

Westerhout, Eline and Paola Monachesi, 2007. Extraction of Dutch definitory contexts for eLearning purposes. In *Proceedings of CLIN 2007*.