



## Recognition of Structured Collocations in An Inflective Language

Bartosz Broda, Magdalena Derwojedowa, and Maciej Piasecki

Institute of Applied Informatics, Wrocław University of Technology, Poland  
Wybrzeże Wyspiańskiego 27, Wrocław, Poland,  
`bartosz.broda,maciej.piasecki@pwr.wroc.pl`  
Institute of Polish, University of Warsaw  
`derwojed@uw.edu.pl`

**Abstract.** We present a method of the structural collocations extraction for an inflective language (Polish) based on the process divided into two phases: extraction and filtering of the pairs of wordforms reduced to baseforms and structural annotation of the extracted collocations with lexico-syntactic patterns. The parameters of the patterns are specified manually but their instances are generated and tested on the corpus automatically. The extracted collocations were evaluated by applying them as rules in morpho-syntactic disambiguation of Polish and by comparing them with a lists of two-word expressions extracted from two Polish dictionaries.

### 1 Introduction

According to the generative power of the natural language, humans are able to produce the infinite number of sentences as well as they can flexibly combine words in compliance with syntactic and semantic rules. However, some sequences of words express more fixed structure than others: their constituents co-occur more often and changes in their structure are very restricted (sometimes even impossible). There is no general name for this broad class of ‘non-atomic language units’ as subsets of the class (varying in the scope of their semantic properties) are called: *collocations*, *fixed expressions*, *terms* or *proper names*. Further on we will call them simply multiword expressions (MEs) or collocations (*sensu largo*).

As collocations introduce a kind of fixed points into the space of possible language expressions, they are very important for the *Natural Language Engineering* (NLE), e.g. identification of collocations can enrich and reduce the description of a document in *Information Retrieval* (IR), improve the accuracy of OCR, or increase the quality of *Machine Translation* (MT). Unfortunately, only a small number of collocations (mostly idioms) is listed in dictionaries, partly because collocation lists are very large and many collocations are domain dependent. That is why the *automatic recognition of collocations* on the basis of large set of text documents—a *corpus*—is very important to the applications in NLE, IR, MT and the similar areas.

There are plenty of methods for recognition of collocations starting with the seminal paper [17]. Most of them are based on the statistical measures of likelihood of the co-occurrence of two *word forms* (WFs) in texts. This general scheme works fine for English but it expresses two significant drawbacks in the case of inflective languages like Polish. Firstly, the fixed order of constituents implicitly assumed in many methods doesn't work for (almost) free word order in Polish; secondly (this is even more important), Polish lexemes are expressed by many WF's. All well known methods treat each WF separately, e.g. two sequences *czerwona kartkę* (*red card*<sub>case=acc, number=sg</sub>) and *czerwonych kartek* (*red card*<sub>case=gen, number=pl</sub>) are analysed as two different collocations regardless of the fact that they are both derived from the expression *czerwona kartka* (*red card*<sub>case=nom, number=sg</sub>) and differ only in the values of the *case* and the *number*. The syntactic structure and meaning of these expressions are the same ('penalty card', as in football).

The aim of this work is to construct a method of collocations recognition that copes with the large number of WF's for one lexeme and identifies the basic syntactic structure of a collocation i.e. the morpho-syntactic dependencies between words in it.

There is no common definition of a collocation. In this paper we adopt the one by Manning & Schütze (it can be regarded as a mainstream definition; [7, pp. 151]):

“*A collocation* is an expression consisting of two or more words that correspond to some conventional way of saying things.” “[...] are characterized by limited *compositionality*”.

Collocations are not compositional in their meaning, i.e. the meaning of a collocation cannot be fully predicted from the meanings of the constituents. It is impossible to exchange one of the collocation constituents to its synonym, e.g. *czerwony arkusz* (*a red sheet*) means something different than *czerwona kartka* (*a red card*), while in many contexts *arkusz* (*a sheet*) is a synonym of *kartka* (*a card*, cf. [3]). Moreover, some types of collocation like *fixed expressions* (cf. [8]) have irregular syntactic structure. Collocations include or at least overlap in large extent with *terminology*, i.e. technical terms and proper names ([7, 5, 10]).

Most methods of collocation recognition are based on the identification of such sequences of WF's that are more frequent than it would be expected from the probabilistic distributions of their constituents. Several statistical and heuristic measures based on statistics have been proposed. An extensive list of 84 measures is surveyed in [10]. In [1], the work on the recognition of collocations in Polish corpus, 16 different measures are tested. Statistical identification of significantly frequent sequences is often accompanied by additional pre- and post-processing, especially in the case of languages of rich inflection. During preprocessing the text is first filtered against *stop lists* of meaningless, too general or unknown WF's and then analysed morphosyntactically in order to annotate them with a PoS and values of the morphosyntactic categories, e.g. case, gender, number, tense. Moreover, *morphological base forms* (or lemmas, BF's) can be also assigned to WF's in text. In the case of Serbian (cf. [9]), the preprocessing was extended

with syntactic filters (implemented as regular expressions) identifying potential terminology.

There is a limited number of works on Slavic languages ([9, 16, 18]) and only one for Polish: Buczyński's *Kolokacje* system (cf. [1]) is based exclusively on statistical recognition of significantly frequent two-word sequences of WFs.

## 2 Basic Statistical Recognition

Polish is a language of rich inflection, which means, that a lexeme is (typically) a set of many wordforms, e.g. up to 14 WFs for a noun and even up to 119 for a verb (including participles, gerunds etc.). The application *Kolokacje* (cf. [1, 2]) works on texts in Polish and implements 16 different statistical measures for binary collocations on the level of WFs. The properties of statistical measures were subjects of many studies, so we decided to use one of the measures implemented in *Kolokacje*, and concentrate on the problems of the Polish inflection and free word order.

Contrary to [9], we wanted to keep the first phase of processing, i.e. statistical recognition, as simple as possible. To do that, we apply linguistic filtering in the post-processing, when the possible collocations are already identified. The cost of syntactic analysis of occurrences of selected potential collocations is much lower than the syntactic analysis of the entire corpus. Moreover, we wanted to make the syntactic filtering more automatic. We also wanted to avoid manual construction of the detailed syntactic rules. This is especially difficult because of the free word order in Polish.

The MEs recognition process has been divided into three phases:

1. *reduction of WFs* — all WFs are reduced to BFs,
2. *statistical recognition* — frequent sequences of BFs are identified and are marked *potential collocation*,
3. *statistical syntactic filtering* — frequency of potential collocation matching syntactic constraints is tested and a list of structurally annotated collocations is generated.

Polish wordforms are often ambiguous among several possible BFs, e.g. *mam* can be a WF of the following lexemes (BFs are listed): *mama* (*mummy*<sub>case=gen,num=pl</sub> infml. 'mother'), *mieć* (*to have*<sub>person=1st,num=sg,tense=present</sub>) and *mamić* (*to delude*<sub>imperative</sub>). This type of ambiguity can be solved only by analysing the context. For the disambiguation we applied *TaKIPI* — a morpho-syntactic tagger of Polish (cf. [12]). Its accuracy is 93.44%, when measured for all tokens and the complete morpho-syntactic description (86.3% for ambiguous words only). The accuracy of the base form disambiguation has not been measured yet. We can expect it lower, but close to the PoS disambiguation i.e. 98.8% (91.64% for ambiguous words only).

For all experiments, we used the largest corpus of Polish, namely IPI PAN Corpus (henceforth IPIC; [14]). During the reduction phase, all documents of

IPIC, (254 524 624 tokens in total), have been disambiguated by TaKIPI and saved as sequences of BFs. Next, the *Kolokacje* application slightly modified in order to make the processing of so large corpus possible was used for the statistical recognition. A list of potential collocations was produced according to the selected measure. Because of the technical properties of *Kolokacje* we limited ourselves to binary collocations. We tested several measures implemented in *Kolokacje*, achieving the best results (according to the selective manual evaluation) for the *Frequently Biased Symmetric Conditional Probability* (FSCP):

$$R_{FSCP} = \frac{c(w, w')^3}{c(w)c(w')} \quad (1)$$

where  $w, w'$  are words, and  $c(w)$ ,  $c(w, w')$  are frequencies of a word and a pair, respectively.

FSCP, proposed in [1], produces similar results to *Log Frequency Biased Mutual Dependency*, but is more efficient. As a result, 304,139 binary potential collocations, for which the value of (rounded to third decimal place) FSCP was greater than 0, were identified. The reduction to BFs decreases the complexity of texts by making all different forms of lexemes equal. On the other hand, it can result in accidental association of words that are not syntactically linked, because morpho-syntactic properties of WFs are not expressed on the level of BFs. E.g. after transforming the sentence below to BFs:

- WFs: *Dalem **długopis czerwony** koledze.*  
(*I gave a red ballpoint to a colleague.*)
- BFs: *Dać **długopis czerwony** kolega.*

there is no information left that *czerwony* (*red*) modifies *długopis* (*a ballpoint*) and not *kolega* (*a colleague*). But the most unwanted side-effect of this method is that some MEs, which are fixed not only lexically, but also grammatically (mostly verbal and prepositional collocations), can be reduced to unrecognizable BFs. For example three possible noun-noun constructions — both nouns in the same case. i.e. an apposition, e.g. *królowa matka*<sub>case=nom</sub> (*queen mother*); noun and subordinate noun in genitive, e.g. *pies*<sub>case=nom</sub> *sąsiada*<sub>case=gen</sub> (*neighbour's dog*) and noun that has its own requirement, usually inherited from verb in the lexical derivation, e.g. *pomoc*<sub>case=nom</sub> *ofiarom*<sub>case=dat</sub> *wypadku* (*help for the victims of the accident*) — are presented as the pairs of the same base forms, although the linguistic mechanism of each of those collocations is totally different. This loss of morphosyntactic information can be a problem if such data are used for (theoretical) linguistic purposes.

### 3 Statistical Syntactic Filtering

The main goal of the filtering phase is to separate accidentally associated pairs of BFs from the ones representing real syntactic units in the corpus. After the manual inspection of potential collocation we identified several classes among them

corresponding to the interesting collocation types, namely: **Adj-Noun**, **Noun-Adj**, **Noun-Noun**, **Prep-Noun**. The last class was introduced experimentally to identify fixed associations of nouns with prepositions (any regularities could be very useful in automatic identification and classification of some types of adjuncts).

Each class is characterised by different syntactic relations between the elements in the pair. It is possible to express these relations by a formal *constraint*, called *constructional constraint* (CC) which must be satisfied for any pair corresponding to the given potential collocation. Let's assume that in the case of the sentence in Sec. 2 two collocations are recognised: *czerwony (red)-dlugopis (a ballpoint)* of the class **Noun-Adj** and *czerwony (red)-kolega (a colleague)* of the class **Adj-Noun**. In order to find which of the two is really supported by the original sentence, we need to identify the corresponding WFs in text and to check if they agree in number, case and gender (then a CC requires that both have the same value of those three categories). This agreement takes place in the case of the first pair—*dlugopis czerwony*, but it is absent in the second pair. As each token in IPIC was previously annotated with the morpho-syntactic information by TaKIPI, it is enough to move a *text window* of the size two across IPIC to identify the corresponding pairs of WFs.

The formal tool for expressing the constraints and checking them for pairs of wordforms in IPIC we used is the JOSKIPI language of the syntactic constraints and its implementation in the TaKIPI engine (cf. [11]). The constraints are applied to each position of a text window. Let's take the CC of the class **Noun-Adj** as an example: `agrpp(0,1,nmb,gnd,cas,3)`, where `agrpp` is an operator testing the agreement on number, gender and case between the first and the second WF in the text window.

For each potential collocation  $\langle b_i, b_j \rangle$ , we need to check if the number of the WF pairs satisfying the appropriate CC, written  $CC(\langle b_i, b_j \rangle)$ , is significantly large in comparison to some accidental value. We used the standard *t-score test*:

$$\frac{|CC(\langle b_i, b_j \rangle) - \frac{n}{V}|}{\sqrt{\frac{n}{V}}} \quad (2)$$

where  $n$  is the number of WFs corresponding to  $\langle b_i, b_j \rangle$ , and  $V$  is the number of possible combinations of values tested by the given CC, e.g. in the case of the CC presented above we have 2 possible numbers, 5 genders and 7 cases, that gives  $V = 4900$  combinations— $1/4900$  probability of the CC equals `true` in the null hypothesis;  $V$  is specified for each CC separately.

There can be several CCs defined for a class of potential collocations, because pairs of BFs can result from different types of syntactic relations, e.g. in the case of the **Noun-Noun** we distinguished three possible types of syntactic constructions:

1. the second noun is in genitive and modifies the first one, e.g. *szuźba zdrowia* ( $V = 7$ ):  
`and(equal(cas[1],gen),not(equal(cas[0],cas[1])))`;
2. a symmetric construction—the first noun is in genitive ( $V = 7$ ):  
`and( equal(cas[0],gen), not( equal(cas[0],cas[1]) ) )`;

3. both nouns are in the same case—(typically these are) proper names, e.g. *Jan Paweł (John Paul)* ( $V = 49$ ): `equal(cas[0],cas[1])`.

The constructional properties of a collocation are necessary, but not the only ones features, e.g. *Gwiezdne Wojny (Star Wars)* occurs in text only in plural—this is not necessary in the syntax, but results from the semantics. In order to identify such properties we introduced additional set of constraints for each class of potential collocations, called *specifying constraints* (SCs). Each SCs is defined as a template of all possible significant syntactic regularities of WF pairs corresponding to potential collocation. The template is written in the form of a sequence of JOSKIPI operators— $o_1, \dots, o_k$ , and the regularities are statistically significant patterns of values of the operators, i.e.  $o_1 = v_{1,i}, \dots, o_k = v_{k,j}$ , e.g. for the Noun-Noun class and the second CC above, the following SC is defined: `and(nmb[0],nmb[1])`—we check whether there is a statistically significant pattern of occurrences constraining values of numbers of the two word forms. For example for *Dynamo Kijów (Dynamo Kyiv)* such pattern was found automatically—when those word forms co-occur forming ME then both are in singular.

In order to distinguish significant patterns we use the *t-score test* again. The test is limited to WF pairs of a given potential collocation. Such pairs that match the given CC—we look for syntactic regularities only across instances of the given collocation. The null hypothesis is that all patterns of operator values are equally possible. Besides the sequence of operators, each SC is specified with the list of the numbers of possible values of operators— $val(o_1) * \dots * val(o_k)$ , that the corresponding JOSKIPI operator can produce. This list is the parameter of the null hypothesis:

$$\frac{|SC(o_1, \dots, o_k)| - \frac{|CC((b_i, b_j))|}{val(o_1) * \dots * val(o_k)}}{\sqrt{\frac{|CC((b_i, b_j))|}{val(o_1) * \dots * val(o_k)}}} \quad (3)$$

where  $SC(o_1, \dots, o_k)$  is an instance of SC, i.e.  $o_1 = v_{1,i}, \dots, o_k = v_{k,j}$ , and  $|SC(o_1, \dots, o_k)|$  is the size of the set of WF pairs satisfying this SC instance.

All possible instances of SC (of any subsequence of operators) for the given CC and potential collocation are generated and tested. The instances satisfying the test (with 99.5% confidence) are saved to a file as significant regularities, e.g. for the collocation *mistrzostwa<sub>num=pl</sub> świata<sub>num=sg</sub> (championship)* one significant instance of the SC template: `and(nmb[0],nmb[1])`, was found: `nmb[0]=pl,nmb[1]=sg`.

## 4 Evaluation

A proper evaluation of the collocations extraction is a permanent problem (cf. [7]), because dictionaries of collocations are created rarely and their coverage is selective and limited. There is no available electronic dictionary of collocations

for Polish. Thus, a sound evaluation process based on the precision and recall calculated in relation to some manually created pattern set is not possible for Polish. Additionally to a limited manual assessment we decided to perform two tests:

1. applying the extracted collocations as a knowledge source in the morphosyntactic disambiguation of Polish — the improved accuracy was expected,
2. comparing the extracted collocations with two lists of two-word lexical units extracted from the electronic source, i.e. [13], and from [15] (by queries on the WWW interface formulated on the basis of WFs from IPIC).

For the needs of the first test, we used a statistical morphosyntactic tagger of Polish (cf. [6]) based on the basic *bi-gram Markov Model* (cf. [7]). The accuracy of this tagger is 91.3% on all words. During the test 102,286 instances of collocations were found. Tagger has accuracy of 94.8% on them. We transformed the extracted collocations and their CCs into rules of morpho-syntactic tag elimination, which removed all tags not fulfilling the CC for found collocations. The rules removed the proper description only in 0.5% of WFs and the number of tags was reduced by 44.4%, the ambiguity was not resolved completely. The accuracy of the tagger was increased to 91.8% on the whole while measured only for the words in collocations to 95.7%. It means that the application of the rules had a positive impact on the work of the whole statistical tagger.

In the second test we used the joint list of two-word lexical units extracted from both dictionaries, i.e. on 8,601 pairs. Next we compared the joint list with the list of extracted collocations. In the case of all classes of collocations, the recall is 46.7%, see Tab. 1. In the case of the classes *Adj-Noun* and *Noun-Adj*, it is difficult to calculate the exact value of the recall because for WFs in dictionaries no Part of Speech is assigned. However, on the basis of manual inspection, it is much higher than in the first case. In both cases precision is low — quite expectable result for the small general dictionary as a source of collocations.

**Table 1.** Comparison of the extracted collocations with the two dictionaries.

|                                     | collocations | common | missed |
|-------------------------------------|--------------|--------|--------|
| all classes                         | 682,454      | 4,015  | 4,586  |
| <i>Adj-Noun</i> and <i>Noun-Adj</i> | 338,467      | 3,360  | —      |

We decided to take a look into extracted ME, but because their number is very large we have selected a random sample for each class of ME for evaluation by a qualified linguist (one of the co-authors). Size of the samples has been determined using tables from [4]. Population size was rounded up to values chosen by Israel. Using assumption of 95% confidence level he used the following formula:  $n = \frac{N}{1+N(e)^2}$  ( $n$  is sample size,  $N$  is population size and  $e$  is desired level of precision). Together 3,149 out of total 94,558 ME were rated (see table 2).

All selected collocations were analysed manually and assigned to six types:

**Table 2.** Size of samples in relation to multiword expression types. Precision level is 5% and confidence level is 95%.  $Sum_N$  is number of merged test cases within broader class.

| Class       | Adj-Verb | Noun-Adj | Noun-Noun | Noun-Verb | Verb-Adj | Verb-Noun |
|-------------|----------|----------|-----------|-----------|----------|-----------|
| $Sum_N$     | 512      | 23914    | 55278     | 5310      | 628      | 8916      |
| Sample size | 308      | 394      | 1122      | 552       | 323      | 450       |
| N           | 21.75%   | 11.93%   | 30.48%    | 41.12%    | 6.81%    | 23.33%    |

**Table 3.** Results [%] of evaluation.

| Class | Adj-Verb | Noun-Adj | Noun-Noun | Noun-Verb | Verb-Adj | Verb-Noun |
|-------|----------|----------|-----------|-----------|----------|-----------|
| N     | 21.75    | 11.93    | 30.48     | 41.12     | 6.81     | 23.33     |
| B     | 52.27    | 12.44    | 24.15     | 23.55     | 78.33    | 6.67      |
| NW    | 0.00     | 5.33     | 11.05     | 0.18      | 0.00     | 0.00      |
| Nwb   | 0.00     | 0.25     | 0.89      | 0.00      | 0.00     | 0.00      |
| K     | 7.47     | 55.84    | 0.53      | 21.20     | 3.72     | 58.00     |
| Kb    | 16.88    | 1.78     | 30.57     | 13.04     | 9.60     | 9.11      |
| F     | 0.65     | 12.44    | 1.52      | 0.18      | 0.31     | 1.78      |
| Fb    | 0.97     | 0.00     | 0.80      | 0.72      | 0.93     | 0.89      |

- B — error, (e.g. “Al-Kaida” separated in two words),
- K — real collocations,
- Kb — real collocations but with some grammatical properties not described,
- NW — proper names (collocations, too),
- Nwb — proper names but with some grammatical properties not described,
- N — insignificant or accidental association (originating in the unbalance of the corpus),
- F — phraseology,
- Fb — phraseology with some grammatical properties not described.

ME of the type **Prep-Noun** were excluded from the manual evaluation, as we had not expected to find any significant real collocations. Our aim here was to identify some more significant associations of prepositions and nouns that can be used during tagging in order to disambiguate the case in both constituents. ME of this type were applied as rules during tests with the tagger.

In these ME in which both constituents are associated by morpho-syntactic agreement (involving adjectives and nouns), the number of wrongly associated word forms is very low. However, if an extracted ME is only the result of co-occurrence in sequence in the text, the number of errors and insignificant associations is high. It is important that on average, a significant percentage of the extracted ‘collocations’ are just stronger syntactic-semantic associations of the type N, e.g. *złamane ramię* (a broken arm), *wymóg religii* (a religion requirement), or *gwaltowna fala* (an instantaneous rapid wave). Such pairs are not real collocations according to the definitions, but are very useful in text processing,

**Table 4.** Results [%] summed in groups.

| Class            | Adj-Verb | Noun-Adj | Noun-Noun | Noun-Verb | Verb-Adj | Verb-Noun |
|------------------|----------|----------|-----------|-----------|----------|-----------|
| K+Kb             | 24.35    | 57.61    | 31.11     | 34.24     | 13.31    | 67.11     |
| F+Fb             | 1.62     | 12.44    | 2.32      | 0.91      | 1.24     | 2.67      |
| NW+Nwb           | 0.00     | 5.58     | 11.94     | 0.18      | 0.00     | 0.00      |
| N+B              | 74.03    | 24.37    | 54.63     | 64.67     | 85.14    | 30.00     |
| K+Kb+NW+Nwb      | 24.35    | 63.20    | 43.05     | 34.42     | 13.31    | 67.11     |
| K+Kb+F+Fb        | 25.97    | 70.05    | 33.42     | 35.14     | 14.55    | 69.78     |
| K+Kb+NW+Nwb+F+Fb | 25.97    | 75.63    | 45.37     | 35.33     | 14.55    | 69.78     |

e.g. in the morpho-syntactic disambiguation. Moreover, many pairs of the type N are accidental associations of an adjective describing a colour, geographic origin, time or those, which have very general meaning like *mały* (*small*), *nowy* (*new*), etc. Such pairs can be eliminated by simple additional post-processing.

## 5 Conclusions

If we take into account the raw numbers of precision and recall we can say that the approach failed. However, we have to consider that the used dictionaries are general and quite small. Discovering general collocations in a large general corpus is very difficult, especially in the case of an inflective language like Polish. Application of this method to some domain corpus could result in better figures. Errors caused by the *TaKIPi* are very seldom, one of them is *Al-Kaida*, that was mistakenly separated and interpreted as the association of two nouns.

The application of all extracted collocations in morpho-syntactic disambiguation was quite successful. Moreover, the manual inspection of the extracted collocations showed that in spite of the substantial number of false collocations observed it is still relatively easy to notice the real ones and separate them by editing. Thus, the created tool can be used for semi-automatic collocation extraction.

The syntactic was extracted so the manual work was reduced significantly in comparison to other approaches, e.g. [9]. We did not have to create detailed syntactic rules. Especially, the automatic check of SCs brought interesting results with the minimal human effort. Moreover, the processing is quite efficient — results for very large corpus are processed on an average contemporary PC in less than a day.

In further research we will concentrate on the reduction of the nominal pairs. This goal can be achieved by elimination of the associations with too general adjectives based on information theory and by the application of a semantic stop list including adjectives expressing time or geographical origin. We want to introduce the additional measures based on the contexts in which the given pair is used.

The extracted ME and the method can be also applied in OCR correction of handwriting or in Speech Recognition in a similar manner to morpho-syntactic

tagging task. In postprocessing phase one can directly use collocations and their syntactic descriptions to correct errors in recognition of multiword expressions. Our long-term goal is the extraction of lexical units for the needs of semi-automatic extension of a lexicon.

**Acknowledgement.** Work financed by the Polish Ministry of Education and Science, project No. 3 T11C 018 29.

## References

1. Buczyński A.: *Pozyskiwanie z internetu tekstów do badań lingwistycznych*, Msc thesis, Wydz. Mat., Inform. i Mech., Uniwersytet Warszawski (2004).
2. Buczyński A., Okniński T.: *Program Kolokacje*, <http://www.mimuw.edu.pl/polszczyzna/kolokacje/> (2006).
3. Derwojedowa M., Piasecki M., Szpakowicz S., Zawislawska M.: *plWordNet—the Polish Wordnet*, WWW: <http://plwordnet.pwr.wroc.pl> (2007).
4. Israel G.: *Determining Sample Size*, University of Florida Tech. Rep., 1992.
5. Jacquemin C.: *Spotting and Discovering Terms through Natural Language Processing*, The MIT Press (2001).
6. Kukła P.: *Tager dla języka polskiego oparty na kombinacji metod statystycznych*, Msc thesis, Wydz. Inf. i Zarządz., Politechnika Wroclawska (2007) In preparation.
7. Manning C. D., Schütze, H.: *Foundations of Statistical Natural Language Processing*, The MIT Press (2001).
8. Moirón V. M. B.: *Data-driven identification of fixed expressions and their modifiability*, PhD thesis, Rijksuniversiteit Groningen (2005).
9. Nenadić G., Spasić I., Ananiadou S.: *Morpho-syntactic clues for terminological processing in Serbian*, In: Proceedings of Workshop on Morphological Processing of Slavic Languages, EACL 2003, Budapest, Hungary (2003).
10. Pecina P.: *An extensive empirical study of collocation extraction methods*, In: Proceedings of the ACL Student Research Workshop, Ann Arbor, Michigan, Association for Computational Linguistics (2005) 13–18.
11. Piasecki M.: *Hand-written and Automatically Extracted Rules for Polish Tagger*, In Sojka, P. et. al. (ed.) Proc. of the Text, Speech and Dialog 2006 LNAI, Springer (2006).
12. Piasecki M., Godlewski G.: *Effective architecture of the Polish tagger*, In Sojka, P. et. al. (ed.) Proc. of the Text, Speech and Dialog 2006 LNAI, Springer (2006).
13. Piotrowski T., Saloni Z.: *Kieszonkowy słownik angielsko-polski i polsko-angielski*, Wyd. Wilga, Warszawa (1999).
14. Przepiórkowski A.: *The IPI PAN Corpus Preliminary Version*, Institute of Computer Science PAS (2004).
15. PWN: *Słownik języka polskiego*, Published on WWW: <http://sjp.pwn.pl/> (2007).
16. Sharoff S.: *What is at stake: a case study of Russian expressions starting with a preposition*, In Tanaka T., Villavicencio A., Bond F., Korhonen A., eds.: Second ACL Workshop on Multiword Expressions: Integrating Processing, Barcelona, Spain, Association for Computational Linguistics (2004) 17–23.
17. Smadja F.: *Retrieving collocations from text: Xtract*, Computational Linguistics **19**(1) (1993) 143–177.
18. Spasic I.: *A Machine Learning Approach to Term Classification*, PhD thesis, Information Systems Research Centre School of Computing, Science and Engineering University of Salford, Salford, UK (2004).