Intelligent Information Systems 9999
ISBN 666-666-666, pages 1–11

Looking for proverbial needles in the proverbial haystack

Filip Graliński

Faculty of Mathematics and Computer Science, Adam Mickiewicz University, Poznań, Poland

Abstract

In this paper, a new method for detection of idiomatic expressions is presented. The proposed method is based on the observation that idiomatic expressions are sometimes marked by metalinguistic expressions like the word *proverbial* or quotation marks. This technique is presented using an example of Polish phraseologisms and the Polish word *przysłowiowy* (*proverbial*). Some approaches to the delimitation of idiomatic expressions are discussed in this paper, including an approach based on pointwise mutual information. Experiments performed on a corpus of Polish web sites are reported here.

Keywords: idiomatic expressions, computer-assisted lexicography, Web corpora

1 Introduction

Idiomatic expressions (idioms)[1], i.e. semantically idiosyncratic phrases, are of interest from a linguistic, sociolinguistic and psycholinguistic as well as natural language engineering perspective. Their non-compositionality provokes theoretical questions and poses practical challenges. In particular, idiomatic expressions must be taken into consideration in machine translation, otherwise grossly misleading and absurd translations might be produced.

As new idiomatic expressions are coined all the time (Bąba and Liberek, 2002) and it is not always straightforward to manually find and identify them in textual material, there is a clear need for computer assistance in phraseological lexicography. With the rapid growth of the Internet, a natural question arises of how to search for new idiomatic expressions in huge Web corpora. The abundance of informal texts on the Internet (e.g. personal blogs, message boards, Web 2.0 sites) makes it even more interesting, as colloquial idioms constitute a substantial part of idiomatic language.

Previous research on the automatic acquisition of English non-compositional expressions focused on their statistical properties (Lin, 1999; Fazly *et al.*, 2009). Furthermore, some methods were limited to a subset of idiomatic expressions, e.g. those of the form *X and/but/or Y* (Widdows, 2005). Such techniques might be applicable for other languages, such as Polish, but not without some challenges.

[1] Usually referred to as *phraseologisms* (*frazeologizm*) in Polish linguistic literature.

TABLE 1: Types of metalinguistic markers. # – number of occurrences of a marker in the Web corpus, I – percentage of idiom- and proverb-related uses estimated on a sample of 100 fragments.

marker	regexp used	#	I
przysłowiowy	`przys[łl]owiow`	29737	50%
tak zwany	`tak zwan`	10248	5%
tzw.	`tzw`	104776	2%
jak to mówią	`jak to m[oó]wi[ąa]`	444	25%
jak to się mówi	`jak to si[eę] m[oó]wi`	532	26%
dosłownie	`dos[łl]ownie`	6554	8%
quotation marks	`"[^\"]{1,50}"`	1741070	1%

For instance, Fazly *et al.* (2009) relies, to some extent, on distribution of articles, which are absent from the Polish language. For a description of Web-mining system dedicated to collocations (a superset of what is considered an idiomatic expression here), see Buczyński (2004).

In this paper, I present a new method for the detection of Polish idiomatic expressions. The proposed method is based on the observation that phraseologisms are quite frequently accompanied by *metalinguistic markers*, the adjective *przysłowiowy* (*proverbial*) being the most distinctive. The idea is to look for such metalinguistic markers, to filter out non-idiom-related occurrences, and to delimit the adjacent idiomatic expressions.

Descriptive and prescriptive linguists have been aware of idiom-related metalinguistic markers (Lewicki, 1974, 2003; Bąba, 1989; Markowski, 2006; Sag *et al.*, 2001). However, such markers have not been exploited for the purposes of automatic detection of idiomatic expressions so far. Their frequency is not very high but the sheer size of the Web makes their number large. They are valuable needles which come in thousands in the huge haystack of the Web.

2 Idiom-related metalinguistic markers in Polish

Some constructions seem to be used by Polish speakers to metatextually mark idiomatic expressions (Lewicki, 1974; Bąba, 1989; Lewicki, 2003). The most important idiom-related metalinguistic markers in Polish are:

- the adjective *przysłowiowy* (*proverbial*) and the adverb *przysłowiowo* (*proverbially*) (see Table 2a and 2b),
- quotation marks (see Table 2c),
- the phrase *tak zwany* (*so-called*) and its abbreviated form *tzw.* (see Table 2d),
- the phrase *jak to mówią* (*as they say*) or *jak to się mówi* (*as it is said*), see Table 2i,
- the adverb *dosłownie* (*literally*), see Table 2j.

I shall focus on *przysłowiowy*/*przysłowiowo*, as it is the most distinctive idiom-related marker (the frequency of idiom-related uses is lower for quotation marks

TABLE 2: Examples of idiom-related metalinguistic markers

a. *10 lutego to* ***przysłowiowa*** *musztarda po obiedzie*
= *February the 10th - it's proverbial missing the boat*
[lit. mustard after the dinner]

b. *Jak na sezon* ***przysłowiowo*** *ogórkowy, koresponduję niezwykle dużo*
= *For a proverbially silly [lit. cucumber] season, I write a lot*

c. *dla nich już zaczął się "sezon ogórkowy"*
= *the "silly [lit. cucumber] season" has already started for them*

d. *ta pora roku to w polityce* ***tzw.*** *sezon ogórkowy*
= *this time of the year is so-called silly [lit. cucumber] season in politics*

e. *Kontrola typów to jest* ***przysłowiowy*** *pikuś*
= *Type control is a proverbial piece of cake*

f. *Myślisz ze jesteś lepszy od* ***przysłowiowego*** *dresa*
= *You think you are better than a proverbial towny [lit. tracksuit]*

g. *mam nadzieję, że* ***przysłowiowa*** *teściowa jest kochana*
= *I hope that the proverbial mother-in-law is sweet*

h. *Bezrobotni chcieliby* ***przysłowiowego*** *kołacza*
= *The unemployed would like a proverbial fruit [lit. round cake]*

i. *Minęło jak z bicza strzelił,* ***jak to mówią***
= *It was over in no time [lit. as a whip cracks], as they say*

j. *A dziś trzeba,* ***dosłownie****, stawać na głowie, żeby coś wygrać*
= *And today you must, literally, bend your neck*
[lit. stand on your head] to win something

k. *te buciki nie będą* ***przysłowiową*** *"kością niezgody" między nami*
= *these shoes will not be a proverbial "bone of contention" for us*

and the phrase *tak zwany*, see Table 1). However, quotation marks will be taken into account when combined with *przysłowiowy/przysłowiowo*, which sometimes happens (see, for example, Table 2k).

It might seem that *przysłowiowy* should be used in proverb-related contexts (as it originated from the noun *przysłowie* = *proverb*), but this is not the case: in a sample of 1093 sentences with *przysłowiowy/przysłowiowo* (see Section 3) just one proverb-related use was found (see Table 2h, the proverb *bez pracy nie ma kołaczy* = *he that would eat the fruit must climb the tree* is alluded there). In fact, the most common uses of the word *przysłowiowy/-o* mean the following:

- *idiomatic*, *phraseological*, see Table 2a and 2b,
- *colloquial*, usually referring to a single word, see Table 2e and 2f,
- *stereotypical*, see Table 2g.

This list is not exhaustive, there are cases when the word *przysłowiowy* is used for reasons unclear even for a Polish native speaker. In general, it seems that Polish speakers use *przysłowiowy* to distance themselves from a word or a phrase (an idiom, a colloquial or slang word, a stereotype-ridden expression etc.) they are going to invoke (Bąba, 1989).

The focus here is on idiomatic expressions, but it should be noted that most of the non-idiom-related uses of *przysłowiowy/-o* are of interest to a lexicographer as well. In particular, sentences with *przysłowiowy/-o* seem to be a rich source of new colloquial or slang words, often not attested in existing dictionaries.

The idiom-related use of *przysłowiowy* (and other similar metalinguistic markers) has been criticised from a prescriptive point of view (Chwałowski, 2002; Markowski, 2006). Nevertheless, it is very popular among speakers of Polish and, paradoxically, we exploit what is considered a mistake or a language abuse.

3 Corpus pre-processing

A Web corpus of 732M words was used. The corpus contained 29737 paragraphs with *przysłowiowy/-o* and its inflected forms, including the variants with letter *l* instead of *ł*[2]. About half of the paragraphs with *przysłowiowy/-o* were collected from web sites obtained by querying Web search engines for *przysłowiowy/-o*.

This material was processed in the following way:

- tokenise a paragraph,
- find *przysłowiowy/-o* tokens,
- take at most 8 tokens to the left and to the right,
- stop at periods and other end-of-sentence characters.

This procedure yielded 27160 unique snippets with the word *przysłowiowy/-o*. A random sample of 1093 snippets was selected. The sample was divided into the development set (100 snippets) and the test set (993 snippets). The idiomatic expressions referred to by *przysłowiowy/-o* in the sample snippets were manually tagged and delimited. (Bąba and Liberek (2002) and Kłosińska *et al.* (2005) were

[2]The regular expression *przys[tł]owiow* was simply used as the first argument to `grep`.

manually consulted, some new idioms were accepted as well, if clearly attested in the Web corpus). The number of idioms found was 408 (37.3%)[3].

4 Identification of idiomatic expressions

Two issues must be addressed:

1. how to identify from among *przysłowiowy/-o* snippets those which contain idiomatic expressions?
2. how to delimit such idiomatic expressions?

Both issues are intertwined in the proposed solution and, contrary to what may be expected, given a snippet with *przysłowiowy/-o*, we first try to delimit a phrase and if the delimitation fails or if a single word was delimited, the snippet is classified as non-idiom-related. The reason why the single-word condition was introduced is that an idiomatic expression, by definition, must be composed of at least *two* words (Lewicki, 2003, pp. 195–203).

4.1 Delimitation procedure

The idiom delimitation algorithm is presented in Table 3. First, the *przysłowiowy/-o* token is located (line 2). Then we take tokens to the right of the *przysłowiowy/-o* token until some condition Γ_R (more on this later) is not true (lines 3–6). This way, the idiom right boundary (b_r) is determined (line 7). Then, we switch to the left side and, analogically, take tokens until condition Γ_L is not met (lines 8–11) and determine the idiom left boundary (b_l, see line 12). Finally, we strip punctuation marks and the *przysłowiowy/-o* token from the beginning (lines 13–15) and the end (lines 16–18).

If at least two tokens were delimited (line 19), the phrase is classified as an idiomatic expression and returned (the *przysłowiowy/-o* token is discarded, see line 21).

In this paper, the following subconditions are considered as contributing to Γ_R:

- $\mu_R(m)$: $i - k \leq m$ (at most m tokens to the right are considered),
- π_R: t_i is a punctuation mark (excluding quotation marks),
- ξ: t_i is the second quotation mark on the right,
- $\phi_R(f)$: $\Phi(t_{k+1} \ldots t_i) \geq f$ (the corpus frequency of $t_{k+1} \ldots t_i$ is greater than or equal to f),
- $\tau_R(t)$:
$$\frac{\Phi(t_{k+1} \ldots t_i)}{\Phi(t_{k+1} \ldots t_{i-1})\Phi(t_i)} > t.$$

In a comparable manner, the following subconditions are considered for Γ_L:

- $\mu_L(m)$: $k - i \leq m$ (at most m tokens to the left are considered),

[3]Only idiomatic expressions marked with *przysłowiowy/-o* were counted.

Table 3: Idiom delimitation procedure

$(t_1, \ldots, t_n) \leftarrow$ the list of the tokens of the given *przysłowiowy/-o* snippet
$k \leftarrow$ the index of the *przysłowiowy/-o* token

$i \leftarrow k + 1$
while $i \leq n$ and Γ_R **do**
  $i \leftarrow i + 1$
end while
$b_r \leftarrow i - 1$

$i \leftarrow k - 1$
while $i > 0$ and Γ_L **do**
  $i \leftarrow i - 1$
end while
$b_l \leftarrow i + 1$

while $b_l \leq b_r$ and (t_{b_l} is a punctuation mark or $b_l = k$) **do**
  $b_l \leftarrow b_l + 1$
end while

while $b_r \geq b_l$ and (t_{b_r} is a punctuation mark or $b_r = k$) **do**
  $b_r \leftarrow b_r - 1$
end while

if $b_r - b_l \geq 1$ **then**
  if $b_l < k$ and $k < b_r$ **then**
    $I \leftarrow (t_{b_l}, \ldots, t_{k-1}, t_{k+1}, \ldots, t_{b_r})$
  else
    $I \leftarrow (t_{b_l}, \ldots, t_{b_r})$
  end if
  remove from I quotation-mark tokens if any
  return I
else
  return false
end if

- π_L: t_i is a punctuation mark (including quotation marks),[4]
- $\phi_L(f)$: $\Phi(t_i \dots t_{k-1} t_{k+1} \dots t_{b_r}) \geq f$ (the corpus frequency of $t_i \dots t_{k-1} t_{k+1} \dots t_{b_r}$ is greater than or equal to f),
- $\tau_L(t)$:

$$\frac{\Phi(t_i \dots t_{k-1} t_{k+1} \dots t_{b_r})}{\Phi(t_i)\Phi(t_{i+1} \dots t_{k-1} t_{k+1} \dots t_{b_r})} > t.$$

4.2 Baseline

For the baseline, the following conditions were used:

- $\Gamma_L = \mu_L(0)$,
- $\Gamma_R = \mu_R(4) \wedge \neg\pi_R \wedge \neg\xi$.

The subconditions used ($\mu_R(4)$, in particular) were derived from the development set. $\mu_L(0)$ means that no tokens to the left of the *przysłowiowy/-o* token are attached (this simplification is fairly good for idiomatic expressions having the syntactic status of noun phrases). Note that no external resources (neither lexicons nor corpora) are involved in the baseline.

For example, for snippet (a) in Table 2 the baseline procedure would yield the idiom *musztarda po obiedzie* (true positive), for (b) — no idiom would be returned as only the single word *ogórkowy* would be delimited (false negative), for (g) — false-positive phrase *teściowa jest kochana* would be returned, for (k) — the correct idiom *kością niezgody* would be delimited.

4.3 Frequency-based approach

This time, the corpus was queried (the same corpus from which the *przysłowiowy/-o* snippets were collected, see Section 3). An idiom is accepted if it occurred at least twice in the corpus, i.e. the following conditions were used:

- $\Gamma_L = \phi_L(2) \wedge \neg\pi_L$,
- $\Gamma_R = \phi_R(2) \wedge \neg\pi_R \wedge \neg\xi$,

As the corpus and the query phrases are lemmatised, some language-specific knowledge is involved, namely a lexicon of inflected forms. However, the lemmatisation process is very crude, no POS tagging was done and in case of ambiguity one lemma is simply chosen at random.

4.4 PMI-based approach

The pure frequency-based approach tends to delimit phrases which are too long (an unwanted frequent word is often attached to the idiom). The solution might be to attach a word only if the association strength between it and the phrase collected so far exceeds some threshold. The pointwise mutual information (PMI), a standard information-theoretic measure (Gale *et al.*, 1991) was chosen as the association strength measure. PMI for two strings w_1, w_2 is defined as:

[4] Note that π_L and π_R are not symmetrical.

$$\mathrm{PMI}(w_1, w_2) = \log \frac{P(w_1 w_2)}{P(w_1)P(w_2)},$$

where $P(w)$ is the probability of string w. If PMI is just compared against a threshold value, frequencies could be used instead of probabilities. Such a simplification was assumed in the definition of τ_L/τ_R subconditions (see Section 4.1).

Strictly speaking, the following conditions were used in the PMI-based approach:

- $\Gamma_L = \tau_L(1.3 \times 10^{-08}) \wedge \phi_L(2) \wedge \neg\pi_L$,
- $\Gamma_R = \tau_R(1.3 \times 10^{-08}) \wedge \phi_R(2) \wedge \neg\pi_R \wedge \neg\xi$,

The threshold value for τ_L and τ_R (1.3×10^{-08}) was tuned to the development set.

4.5 Additional filter

The approaches described in Subsections 4.2–4.4 can be combined with an additional procedure which strips words which are unlikely to occur at the beginning or at the end of an idiomatic expression (Merkel and Andersson, 2000). Such an additional filter could consist in supplementing lines 13 and 16 (see Table 3) with appropriate conditions.

The following assumptions were made:

- idioms may not start with a conjunction,
- idioms may not end with a conjunction nor a preposition.

A list of 148 Polish conjunctions and a list of 134 Polish prepositions were manually constructed for the purposes of the additional filter.

5 Evaluation

The final results are presented in Table 4. Note that precision, recall and F-measure were calculated by taking as true positives only those idioms that were correctly identified *and* delimited.

Interestingly, frequency-based approach performed worse than the baseline (unless combined with the additional filter described in Section 4.5). It might seem that the positive effects of taking PMI into account and using the additional filter would overlap, but it was not the case — the best results were obtained for the PMI-based approach combined with the additional filter.

A sample of 10 snippets processed with the PMI-based method combined with the additional filter is presented in Table 5. The total number of idiomatic expressions reported in the entire set of 27160 snippets was 13648.

TABLE 4: Results obtained for the test set. C – number of correctly recognised and delimited idiomatic expressions, T – number of all reported idiomatic expression, P – precision, R – recall, F – F-measure. Precision, recall and F-measure are calculated by taking as true positives only those idioms that were correctly identified *and* delimited.

	C	T	P	R	F
baseline	117	583	0.201	0.315	0.245
baseline + filter	116	566	0.205	0.313	0.248
frequency-based approach	115	888	0.130	0.310	0.183
frequency-based approach + filter	162	810	0.200	0.437	0.274
PMI-based	120	529	0.227	0.323	0.267
PMI-based + filter	140	479	0.292	0.377	**0.329**

6 Conclusions and future work

We have shown that it is possible, to some extent, to identify idiomatic expressions using metalinguistic markers like the word *przysłowiowy* (*proverbial*) or quotation marks provided that the corpus is large enough. For better results, one should probably turn to syntactic analysis (at least a chunker) and/or machine learning.

If the procedure described in this paper is applied to a large Web corpus, an extensive list of idiomatic expressions currently in use can be obtained. As far as machine translation is concerned, such expressions should be entered into the bilingual lexicon and manually translated.

The approach proposed here should also be feasible for other languages, English in particular[5].

Acknowledgements

The paper is based on research funded by the Polish Ministry of Science and Higher Education (Grant No 003/R/T00/2008/05).

References

Aleksander BUCZYŃSKI (2004), *Pozyskiwanie z Internetu tekstów do badań lingwistycznych*, Master's thesis, Uniwersytet Warszawski, Warszawa.

Stanisław BĄBA (1989), *Innowacje frazologiczne współczesnej polszczyzny*, Wydawnictwo Naukowe UAM, Poznań.

Stanisław BĄBA and Jarosław LIBEREK (2002), *Słownik frazologiczny współczesnej polszczyzny*, Wydawnictwo Naukowe PWN, Warszawa.

Robert CHWAŁOWSKI (2002), *Typografia typowej książki*, Helion, Gliwice.

Afsaneh FAZLY, Paul COOK, and Suzanne STEVENSON (2009), Unsupervised Type and Token Identification of Idiomatic Expressions, *Computational Linguistics*, 35(1):61–103.

[5]For example, the Google search engine returned 176 results for `"proverbial red rag"`, 64 results – for `"proverbial pain in the neck"` and 33 – `"pain in the proverbial neck"` (March 2010).

TABLE 5: A sample of 10 snippets processed with the PMI-based method combined with the additional filter. The position of the *przysłowiowy/-o* token is marked with #. The phrase identified as an idiomatic expression is underlined.

Forma oparta na wspomnieniach jest # "strzałem w dziesiątkę"
= *The form based on memories is a # "bullseye"*

stracona bramka, podziałała na Polonię, jak # płachta na byka i gospodarze ruszyli do śmiałych
= *lost goal, it was for Polonia like a # red rag to a bull and the hosts rushed to bold*

przestępczość z bronią w ręku nie jest wart # funta kłaków
= *crime under arms is not worth a straw*

mojej zabawy z Paintem, który przyjmijcie z # przymrużeniem oka
= *with my playing with Paint, which you will take with # a pinch of salt*

a może inaczej... koncentracji w meczach z # "ogórami", jak te z Polonią
= *and maybe another way... concentration in matches with # "bunglers", as those with Polonia*

Jeśli tak to gratuluję # postawy psa ogrodnika
= *If so I congratulate on the # attitude of dog in the manger*

chłopców, jak i dziewcząt, klną jak # szewcy
= *boys and girls, they swear like # troopers*

Rok temu była # kicha, a teraz dbam o śliweczki,
= *A year ago it was a # bummer, but now I care for the plums*

w którym wszystko jest poukładane i dopięte na # ostatni guzik
= *in which everything is sorted out and # buttoned up*

się Święta z pewnością tu i ówdzie rozpalą # już spór na temat wyższości Świąt Bożego Narodzenia
= *Holidays for sure here and there will kindle the already # dispute on the primacy of Christmas*

W.A. Gale, K.W. Church, P. Hanks, and D. Hindle (1991), Using Statistics in Lexical Analysis, in U. Zernik, editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*, pp. 115–164, Lawrence Erlbaum Associates, Hillsdale, N.J.

Anna Kłosińska, Elżbieta Soboal, and Anna Stankiewicz, editors (2005), *Wielki słownik frazeologiczny z przykładami*, Wydawnictwo Naukowe PWN, Warszawa.

Andrzej Maria Lewicki (1974), Zwroty frazeologiczne, czyli predykaty w formie składników nieciągłych, *Studia gramatyczne I*, pp. 135–143.

Andrzej Maria Lewicki (2003), *Studia z teorii frazologii*, Oficyna Wydawnicza Leksem, Łask.

Dekang Lin (1999), Automatic Identification of Non-compositional Phrases, in *In Proceedings of ACL-99*, pp. 317–324.

Andrzej Markowski, editor (2006), *Wielki słownik poprawnej polszczyzny*, Wydawnictwo Naukowe PWN, Warszawa, entry *przysłowiowy*.

Magnus Merkel and Mikael Andersson (2000), Knowledge-Lite Extraction of Multi-Word Units with Language Filters and Entropy Thresholds, in *In Proceedings of RIAO'2000, Collége de*, pp. 737–746.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger (2001), Multiword Expressions: A Pain in the Neck for NLP, in *In Proc. of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pp. 1–15.

Dominic Widdows (2005), Automatic extraction of idioms using graph analysis and asymmetric lexicosyntactic patterns, in *In Proceedings of ACL'05 Workshop on Deep Lexical Acquisition*, pp. 48–56.