Idiomatic Expressions in VerbaLex

Zuzana Nevěřilová

NLP Centre, Faculty of Informatics,
Masaryk University, Botanická 68a,
602 00 Brno, Czech Republic
xpopelk@fi.muni.cz

Abstract. Idiomatic expressions are part of everyday language, therefore NLP applications that can "understand" idioms are desirable. The nature of idioms is somewhat heterogenous — idioms form classes differing in many aspects (e.g. syntactic structure, lexical and syntactic fixedness). Although dictionaries of idioms exist, they usually do not contain information about fixedness or frequency since they are intended to be used by humans, not computer programs.
In this work, we propose how to deal with idioms in the Czech verb valency lexicon VerbaLex using automatically extracted information from the largest dictionary Czech idioms and a web corpus. We propose a three stage process and discuss possible issues.

Key words: idioms, VerbaLex, verb valency, e-lexicography, Czech

1 Introduction

Formulaic sequences are defined as "a sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar." [16] Idioms are one type of formulaic sequence characterized by its non-compositionality, i.e. its meaning cannot be deduced from the meaning of its parts.

Unlike proverbs, idioms can have literal meaning as well as figurative meaning. In practice, only some idioms occur in their literal meaning, other idioms are used solely in the figurative meaning. For example, while *kick the bucket* can mean a physical action of somebody's foot, it usually means *to die*. By contrast, *cry heart out* is used only in the figurative meaning.

The following properties of idioms are often studied:

- their fixedness or degree of prefabrication (continuity, fixedness of the word order),
- syntactic anomalies or lexical constraints (e.g. only words from a given set can be objects of an expression),
- the usage of the literal meaning.

Aleš Horák, Pavel Rychlý, Adam Rambousek (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2017, pp. 59–67, 2017.

The non-compositionality of idioms is the main cause to build and maintain resources of idioms for the purpose of natural language processing (NLP).

The aim of this work is to re-include idioms in the Czech verb valency lexicon VerbaLex. Currently, verb frames in this lexicon with some fixed part are annotated as idioms but no other information is provided. For example, the fixed part is stored in its inflected form, so the idioms with the same headword are not grouped together. Also, the information about the degree of fixedness (fixed or free word order, usage in different tenses or moods, etc.) is missing.

In addition, the idiomatic frames are sometimes grouped with frames similar in the literal meaning, not the figurative one. For example, the idiomatic frame *uplést si na sebe bič* (make a rod for one's own back) is assigned to the synset *uplést/uplétat* (to knit). On the other hand, the idiom *polykat andělíčky* (meaning to be drowning) is (correctly) assigned to the synset *topit se, polykat* (to drown).

In order to solve the problems described above, we propose a methodology on how to describe idiomatic verb frames with respect to their fixedness, syntactic anomalies, lexical constraints and corpus frequency (without considering the literal or figurative meaning). We also suggest how to assign an idiomatic verb frame to the correct verb synset, either an existing or a new one.

1.1 Practical Outputs

Not only automatic processing of idioms is an interesting topic *per se*, the practical applications of this work can be seen.

Sentiment Analysis Affect is one of the properties of most idioms, so it is clear that dealing with idiomatic expressions is part of sentiment analysis, e.g. [11,15].

Machine Translation Presence of idioms in a sentence may have impact on the quality of statistical machine translation and the system need to transfer idioms properly to target language. For evaluation, see e.g. [13]. There are several methods to overcome errors caused by idioms in statistical machine translation, e.g. [3,2].

1.2 Outline

Section 2 describes the initial language resources. In section 3, we show how are idioms described in other language resources. Section 4 proposes a three phase methodology on how to extract information from a dictionary of Czech idioms and convert it to verb frames. In Section 5, we discuss the possible issues concerning the extraction.

2 Current Resources of Idioms in Czech

Currently, Czech idioms are described in two highly overlapping resources. Their extent and usability from the NLP view differs significantly.

idiom	chovat/hřát (si) **hada** na prsou/za ňadry
grammatical constraints	ot, pas, imp
explanation	věřit někomu nekriticky
translation	cherish a serpent in one's bosom

Fig. 1. Example entry in the Dictionary of Czech phraseology and idioms: the headword is in bold, items are order by headwords alphabetically.

2.1 Dictionary of Czech Phraseology and Idioms

As a source of Czech idioms, we took the *Dictionary of Czech Phraseology and Idioms* (DCPI, [17]). It distinguishes four types of idioms: similes, expressions without verb, verb expressions, and sentences. For our work, we only picked the third part. It contains 19,121 idioms, however the number does not include variants (see below).

An example entry can be seen in Figure 1. The digital version of DCPI contains mostly visual markup and therefore it is not straightforward to extract all variants of the idiom. In the example, all correct variants are:

- chovat hada na prsou
- chovat si hada na prsou
- hřát hada na prsou
- hřát si hada na prsou
- chovat hada za ňadry
- …

2.2 VerbaLex

Verb valency lexicons usually consist of the following units:

- verb synset – a set of synonymic verbs describing an action, event or state
- verb frame – syntactic and semantic description of sentence constituents dependent on the verbs from the synset
- slot – description of each dependent constituent

VerbaLex [7] is the largest verb valency lexicon for Czech. It contains 6,244 verb synsets, 19,158 verb frames, 10,449 unique verbs. An example frame can be seen in Figure 2.

Each verb synset contains information on whether it is used in the passive form. Each slot contains description of some of its syntactic properties: the case of the noun phrase and the preposition if applicable.

Semantic information is available on two levels:

- *semantic role* (also known as thematic role or thematic relation) that a sentence constituent plays with respect to the action or state. VerbaLex contains 33 semantic roles such as agent, patient, location or substance.

Fig. 2. Example frame from VerbaLex. The verbs in the synset sometimes form pairs of perfective/imperfective verbs. This particular frame means *to cradle* somebody in one's arms, lap etc.

– *semantic constraint* on a hypernym (e.g. person). This second level is related to WordNet hypernym [4] (e.g. `person:1`, where `person` is a literal and `1` is the sense number).

Currently, other constraints (e.g. a set of words that can fill a slot, information about word order fixedness) are not implemented.

[6, 5.3.5] describes the annotation of idiomatic frames as follows:

– only idioms from DCPI with frequency higher than 10 in the corpus ALL[1] are included in VerbaLex
– some unspecified idioms not present in DCPI are also included in VerbaLex
– the whole fixed part is described as one slot (its "semantic role" is DPHR, meaning *dependent part of phrase*)
– information about the meaning of the idiom is described in the frame definition
– information about syntactic anomalies of lexical constraints is not included

Currently, VerbaLex contains 1,109 frames with DPHR. [10] distinguishes univalent, bivalent, trivalent, and quadrivalent valencies in English and states that trivalent idioms are very rare and quadrivalent ones seem not to exist. Apparently, the situation in the Czech lexicon is very similar: the vast majority of idioms in VerbaLex are univalent.

3 Related Work

In this section, we describe in short other works that deal with idioms in the context of NLP. We do not describe lexicons of idioms in other languages but focus on idioms in language resources usable in NLP. Idioms occur in all the resources mentioned below since they describe stereotypical patterns.

[1] A 600 million corpus created in Natural Language Processing Centre.

VALLEX In [9], the authors mention that VALLEX contains some very frequent idioms but the focus of the work is on verb in their primary meanings.

VerbNet [14] only mentions that the coverage of VerbNet was extended by WordNet [4] and these verb were also part of idiomatic expressions. Moreover, the idiomatic expressions were grouped together according to their meaning, not the surface structure, e.g. *kick the bucket* is grouped with *to die*.

FrameNet [12, 3.2.7] distinguish idioms and support predicates[2] verb+noun. In FrameNet, idioms are treated as multi-word targets. In both cases, the expression is defined in an appropriate frame according to its meaning.

Pattern Dictionary of English Verbs (PDEV) [5] uses the word *norm* for both literal meaning and conventionalized metaphors and idioms. By contrast, the dynamic metaphors are called *exploitations*. Since our focus is in a dictionary, we deal only with norms. PDEV contains idioms such as *to sing praises of something* but they are not distinguished from other verb patterns. Also, there is no grouping by meaning (i.e. no relationship between *sing praises* and *praise*).

4 Proposed Methods

In order to include idioms into VerbaLex, we propose several steps: extracting the idioms from DCPI, searching the idioms in corpus, and creating verb valency frames for frequently used idioms.

4.1 Extraction from DCPI

During the extraction phase, the most difficult part is dealing with the visual markup and dealing with errors in the markup. We propose the following steps:

1. extract idiom and its explanation from the LaTeXmarkup (i.e. delete other content: historical context of the idiom, translations to other languages than English etc.),
2. expand content in parentheses into several variants (e.g. *hřát (si)* means *hřát* or *hřát si*)
3. expand variants delimited by slashes (e.g. *na prsou/za ňadry*)
4. remove LaTeXmarkup

The most difficult part is expanding variants delimited by slashes since it is not clear how much content is in each variant. We therefore suggest to over-generate the expansion and reduce the unused variants in the next phase. For example from the idiom presented in Figure 1 chovat/hřát (si) hada na prsou/za ňadry, the algorithm can generate the following variants:

[2] A support predicate is a governing verb in cases the syntactic and semantic heads differ, i.e. it "does not reliably have the same meaning independently of the frame-evoking element" [12]

1. *chovat/hřát (si) hada na prsou ňadry
2. chovat/hřát (si) hada na prsou
3. *chovat/hřát (si) hada na za ňadry
4. chovat/hřát (si) hada za ňadry

4.2 Corpus Search

In order to check usage of the idiom in current language, we propose to search it in a large web corpus. At first, we decided *not* to process idioms containing auxiliary or modal verbs. The main reason is that auxiliary verbs are not included in VerbaLex at all, modal verbs are included in several frames not in a consistent way.

The proposed steps are as follows:

1. parse idiom in order to identify verb phrase and the rest (objects, adverbials)
2. if the idiom contains auxiliary of modal verb and no other verb, do not process it
3. optionally exclude non-standard Czech
4. recognize variables (e.g. somebody, something, somewhere) in the idiom and in its explanation
5. construct CQL query for the idiom
6. if the idiom is not found in the corpus, do not process it

We do not have information about word order fixedness, so we propose to construct several CQL queries with different word orders. Again, the over-generation is not a problem. We propose to search the verb phrase as a lemma in order to find different word forms. On the other hand, we propose to search exact word forms of the other parts of the idiom since it can contain non-standard words that can easily be lemmatized incorrectly.

After the corpus search, the sum of the frequencies of different word orders should be above a certain threshold. In addition, the distribution of these frequencies can provide information about the word order fixedness.

4.3 Integration to VerbaLex

The last step is the creation of the appropriate verb frames. We are not sure this process can be fully automated. At least verb frames can automatically be proposed and then checked manually. We propose the following steps:

1. create individual slots for different parts of the idiom
2. parse the explanation and identify verb phrase and the rest (objects, adverbials)
3. find an existing verb frame with a similar meaning

We are aware that many idioms do not have a one word equivalent since their meaning is rather complex (see the discussion below). The final decision whether place the idiom among other frames of a verb synset or whether create a new verb synset will be left on lexicographers.

5 Discussion

We expect that the idioms selected by the procedures described above will strongly overlap with those already present in VerbaLex. However, by using our methods the idiom description should be more detailed.

5.1 Standard vs. Non-standard Language

DCPI contains idioms in both standard and non-standard Czech. However, NLP tools are mainly focused on standard language and also only VerbaLex only contains standard Czech. We propose to exclude non-standard Czech at first and to compare frequencies of standard and non-standard idioms in corpora.

5.2 Different Degree of Fixedness

Currently, no information about the degree of fixedness is present in VerbaLex. However, DCPI contains information on syntactic constraints (e.g. the tenses the idiom can be used in). We propose to add this information as a metainformation to each frame and optionally check the constraint in corpus. We also propose to add information about word order fixedness based on the corpus search (described in 4.2).

The most difficult part will be a description of specific syntactic constraints on the argument. For example, in the idiom *aby* `<PAT>` *husa kopla* (meaning strong disagreement), we did not find any other arguments in the patient slot than pronouns.

5.3 Assignment of an Idiomatic Frame to Verb Synset

Apparently, not many idioms can be explained by one verb. One such example is *to kick the bucket* which can be "translated" as *to die*. In some cases, the idioms intensify a one verb expression, for example, *to laugh your head off* means to laugh (a lot). In these two cases, connection with the close meaning verb (in our examples, *to die* and *to laugh* respectively) is the desired state. Nevertheless, some idioms convey a complex action and no single verb can replace the idiom. For example, *ocitnout se v křížové palbě* (*find oneself in the cross fire*) means to feel pressure from several sides. The border between intensified meaning and a complex action is probably fuzzy.

5.4 Literal vs. Figurative Meaning

We are aware that some idioms are used in both literal and figurative meaning, e.g. *break the ice*. Many studies measure what meanings are activated in these cases, e.g. [1]. There are also works that detect figurative meaning by computational means, e.g. [8]. However, in processing Czech idioms, we are not that far. So, we propose to postpone the research on literal vs. figurative meanings so far.

Acknowledgments. This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071.

References

1. Cacciari, C., Tabossi, P.: Idioms: Processing, Structure, and Interpretation. Taylor & Francis (2014), https://books.google.cz/books?id=RgrsAgAAQBAJ
2. Durrani, N., Schmid, H., Fraser, A., Koehn, P., Schütze, H.: The operation sequence model—combining n-gram-based and phrase-based statistical machine translation. Computational Linguistics (2015)
3. Elloumi, Z., Besacier, L., Kraif, O.: Integrating multi word expressions in statistical machine translation. MULTI-WORD UNITS IN MACHINE TRANSLATION AND TRANSLATION TECHNOLOGIES MUMTTT2015 p. 83 (2015)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)
5. Hanks, P., Pustejovsky, J.: Common Sense About Word Meaning: Sense in Context. In: Sojka, P., Kopecek, I., Pala, K. (eds.) Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 3206, pp. 15–17. Springer Berlin / Heidelberg (2004), http://dx.doi.org/10.1007/978-3-540-30120-2_2, 10.1007/978-3-540-30120-2_2
6. Hlaváčková, D.: Databáze slovesných valenčních rámců VerbaLex. Ph.D. thesis, Masarykova univerzita, Filozofická fakulta, Ústav českého jazyka (2007)
7. Hlaváčková, D., Horák, A.: VerbaLex – New Comprehensive Lexicon of Verb Valencies for Czech. In: Proceedings of the Slovko Conference (2005)
8. Li, L., Sporleder, C.: Using Gaussian Mixture Models to Detect Figurative Language in Context. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 297–300. HLT '10, Association for Computational Linguistics, Stroudsburg, PA, USA (2010), http://dl.acm.org.ezproxy.muni.cz/citation.cfm?id=1857999.1858038
9. Lopatková, M., Bojar, O., Semecký, J., Benešová, V., Žabokrtský, Z.: Valency Lexicon of Czech Verbs VALLEX: Recent Experiments with Frame Disambiguation. In: Matoušek, V., Pavelka, T., Mautner, P. (eds.) LNCS/Lecture Notes in Artificial Intelligence/Proceedings of Text, Speech and Dialogue. vol. 3658, pp. 99–106. Springer Verlag Heidelberg, Karlovy Vary, Czech Rep., Sept. 12-16 (2005)
10. Müller, E.A.: Valence and Phraseology in Stratificational Linguistics. In: Lockwood, D., Fries, P., Copeland, J. (eds.) Functional Approaches to Language, Culture, and Cognition: Papers in Honor of Sydney M. Lamb. J. Benjamins (2000), https://books.google.cz/books?id=quxDFHwth4oC
11. Rill, S., Scheidt, J., Drescher, J., Schütz, O., Reinel, D., Wogenstein, F.: A Generic Approach to Generate Opinion Lists of Phrases for Opinion Mining Applications. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining. pp. 7:1–7:8. WISDOM '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2346676.2346683
12. Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Scheffczyk, J.: FrameNet II: Extended Theory and Practice. Tech. rep., ICSI (Aug 2006), http://framenet.icsi.berkeley.edu/book/book.pdf
13. Salton, G., Ross, R., Kelleher, J.: An empirical study of the impact of idioms on phrase based statistical machine translation of english to brazilian-portuguese (2014)

14. Schuler, K.K.: VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, Faculties of the University of Pennsylvania (2005), http://verbs.colorado.edu/kipper/Papers/dissertation.pdf
15. Williams, L., Bannister, C., Arribas-Ayllon, M., Preece, A., Spasić, I.: The role of idioms in sentiment analysis. Expert Systems with Applications 42(21), 7375 – 7385 (2015), http://www.sciencedirect.com/science/article/pii/S0957417415003759
16. Wray, A.: Formulaic Language and the Lexicon. Cambridge University Press, New York (2002), http://linguistlist.org/pubs/books/get-book.cfm?BookID=2109
17. Čermák, F., et al.: Slovník české frazeologie a idiomatiky I-IV (Dictionary of Czech Phraseology and Idioms, SČFI). Academia, Praha (1983)