Idiomatic Expressions in VerbaLex

Zuzana Nevěřilová, Adam Rambousek

Natural Language Processing Centre, Faculty of Informatics
Masaryk University, Brno, Czech Republic

RASLAN 2017, Karlova Studánka, 01 Dec 2017

Idioms

. . . multi-word expression with some degree of fixedness (prefabrication) that can be used in both literal and figurative meanings.

a carrot and stick, spill the beans . . .

Idioms

. . . multi-word expression with some degree of fixedness (prefabrication) that can be used in both literal and figurative meanings.

a carrot and stick, spill the beans . . .

The figurative meaning dominates.

Meaning of Idioms

Knowing the meaning of idioms is part of the language knowledge.
Native speakers use idioms spontaneously.

Meaning of Idioms

Knowing the meaning of idioms is part of the language knowledge. Native speakers use idioms spontaneously.

Often, idioms can easily be translated (i.e. idiom with the same meaning exists in the target language) but very often, they have to be translated as a whole (i.e. it is not possible to translate an idiom word by word):

a carrot and stick = cukr a bič (sugar and whip)

Meaning of Idioms

Knowing the meaning of idioms is part of the language knowledge. Native speakers use idioms spontaneously.

Often, idioms can easily be translated (i.e. idiom with the same meaning exists in the target language) but very often, they have to be translated as a whole (i.e. it is not possible to translate an idiom word by word):

a carrot and stick = cukr a bič (sugar and whip)

Dictionary of Czech Phraseology and Idioms

Idioms are described in specialized dictionaries. For Czech, it is the Dictionary of Czech Phraseology and Idioms by F. Čermák et al.

vyzvonit° to/všechno

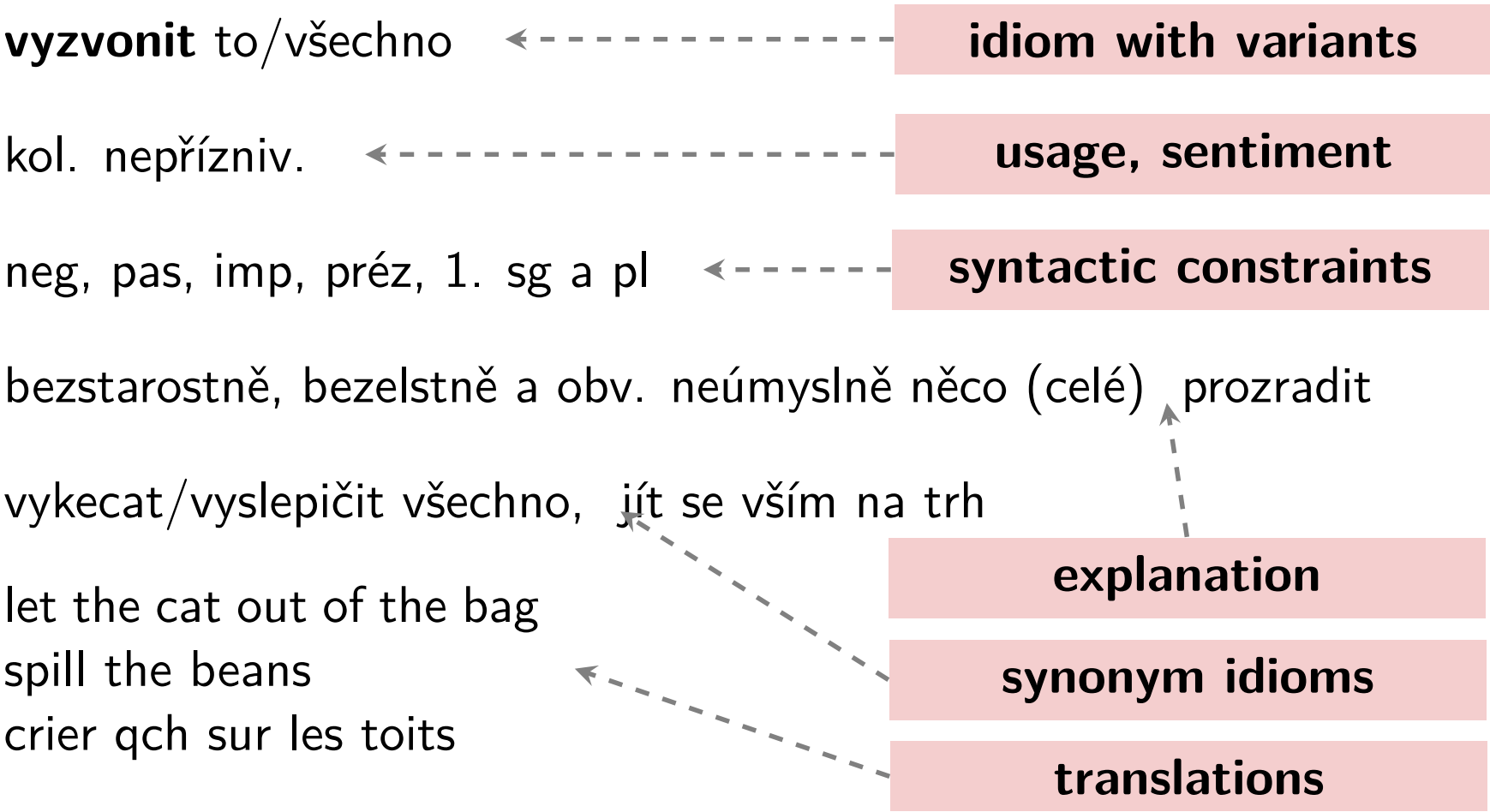
(kol; nepřízn) **0** neg, pas, imp, préz, 1. sg a pl *(Zvl. čl. naivní, nezkušený ap. v postoji ke svěřené důvěrné informaci, tajemství ap.:) bezstarostně, bezelstně a obv. neúmyslně něco (celé) prozradit, někomu oznámit.* Děti chtěly rodiče překvapit, ale malá Markétka to všechno vyzvonila. **Cf** prozradit něco, vykecat/vyslepičit všechno, jít se vším na trh ●**A** let the cat out of the bag, spill the beans **N** (etw ausplaudern) **F** crier qch sur les toits

---

Slovník české frazeologie a idiomatiky: výrazy slovesné

. . . suitable for humans, not understandable by computer programs

Dictionary of Czech Phraseology and Idioms



Describing the Meaning of Idioms

We focus on idioms containing a verb.

Syntactic description:

- valence (most idioms are uni- or bivalent)
- arguments (animate/inanimate, cases, prepositions)
- constraints on word order
- other constraints (tense, negative, number)

Semantic description:

- meaning (explanation or one-word synonym)
- sentiment

VerbaLex

verb valency lexicon of Czech verbs:

6,244 verb synsets, 19,158 verb frames, 10,449 unique verbs

about 30 semantic roles (AG, PAT, LOC, . . .)

chovat$^{\text{impf}}_{1}$ **pohoupat**$^{\text{pf}}_{1}$ **pohupovat**$^{\text{impf}}_{1}$ **pochovat**$^{\text{pf}}_{2}$

1 chovat$_1$, pohoupat$_1$, pohupovat$_1$, pochovat$_2$ ≈

-frame: **AG** <person:1> obl a1 **VERB** obl **PAT** <child:1> obl a4 **PART** <body part:1> opt na+i6

-example: *matka pohoupala dítě na klíně* (***pf***)

to cradle a child in one's bodypart

Idioms in VerbaLex

about 1,000 verb frames with the semantic role DPHR

topit se$^{\text{impf}}_{1}$ **polykat**$^{\text{impf}}_{3}$

1 topit se$_1$ ≈

-frame: **AG** <person:1> obl a1 **VERB** obl

-example: *dítě se topí* (***impf***)

2 polykat$_3$ ≈

-frame: **AG** <person:1> obl a1 **VERB** obl **DPHR** <andělíčky> obl

-example: *dítě polyká andělíčky* (***impf***)

to be drowning

Idioms in VerbaLex

sometimes the meaning of the idiom is assigned to a bad verb synset (here *to knit*)

2 splést$_2$, splétat$_2$, uplést$_2$, uplétat$_2$, zaplést$_2$, zaplétat$_2$ ≈

-frame: **AG** <person:1> obl a1 **VERB** obl **PAT** <person:1> obl a3 **PART** <hair:1> obl i4 **ATTR** <shape:2> obl do+i2

-example: *maminka upletla dceři vlasy do copánků* (***pf***)

3 uplést$_2$, uplétat$_2$ ≈

-frame: **AG** <person:1> obl a1 **VERB** obl **PAT** <person:1> obl a3 **PAT** <person:1> obl na+a4 **DPHR** <bič> obl

-example: *upletla si na sebe bič* (***pf***)

make a rod for one's own back

The plan is . . .

. . . to re-annotate idioms in VerbaLex

- add information about syntactic properties (word order fixedness, constraints on tense, negative, number)
- add information about sentiment
- assign the meaning with the correct verb synset (where applicable)
- find idiom in corpora and provide corpus evidence

The plan is . . .

. . . to re-annotate idioms in VerbaLex

- add information about syntactic properties (word order fixedness, constraints on tense, negative, number)
- add information about sentiment
- assign the meaning with the correct verb synset (where applicable)
- find idiom in corpora and provide corpus evidence

Phase One: Conversion of the Dictionary

- convert Dictionary of Czech Phraseology and Idioms into machine understandable format
- expand entries such as:
  chovat/hřát (si) **hada** na prsou/za ňadry (cherish a serpent in one's bosom)
  to
  - chovat hada na prsou
  - chovat si hada na prsou
  - hřát hada na prsou
  - hřát si hada na prsou
  - chovat hada za ňadry
  - ...

Phase Two: Corpus Search

convert dictionary entries into CQL queries:

- consider changes in the word order
- consider gaps
- consider non-standard uses (e.g. být vs. bejt)

Discussion

- so far, VerbaLex contains standard Czech, however, some idioms are rather non-standard language
- different degrees of fixedness: fixed: aby tě husa kopla (meaning refusal)
  less fixed: chovat si hada na prsou (cherish a serpent in one's bosom)
- literal vs. figurative meanings: we propose not to consider different meanings

And that's it!

Thanks Adam for presenting