Type-based Search of Idiomatic Expression

Jan Bušta

December 7, 2013

7th Workshop on Recent Advances in Slavonic Natural Language Processing

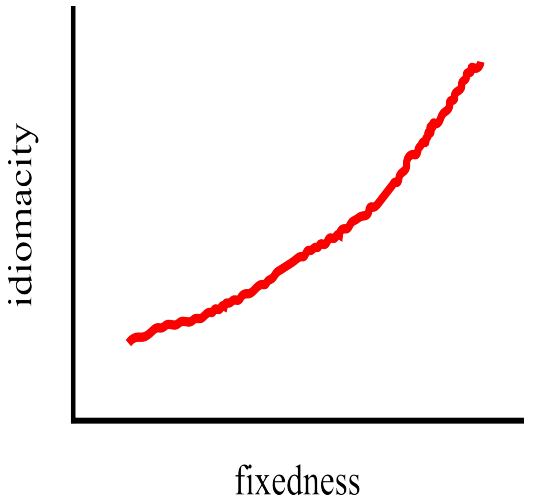
Karlova Studánka

Idiom

- Frege – compositionality
- Čermák – abnormality
- Fixedness – computer processing
  - type-based vs. token-based
- Token based: verb-noun
  - transitive verb (Verbalex: verbs allowing pasivation) + it's direct object (accusative case)

Fixedness

- How much we can change the MWE? Will the chenge preserve the idiomacity?
- More fixed more idiomatic
  - Lexical fixedness
  - Syntactic fixedness



Methods

- PMI

$$PMI(v, n_j) = log \frac{|\mathcal{V} \times \mathcal{N}| f(v, n_j)}{f(v, *) f(*, n_j)}$$

- PMI-based lexical fixedness

$$Fix_{lex}(v, n) = \frac{PMI(v, n) - \overline{PMI}}{s}$$

- Intersegment-based syntactic fixedness

$$Fix_{syn}(v, n) = \frac{max(f(v, n), f(v, d, n), f(v, d, d, n))}{\sum f(v, n), f(v, d, n), f(v, d, d, n)}$$

Results

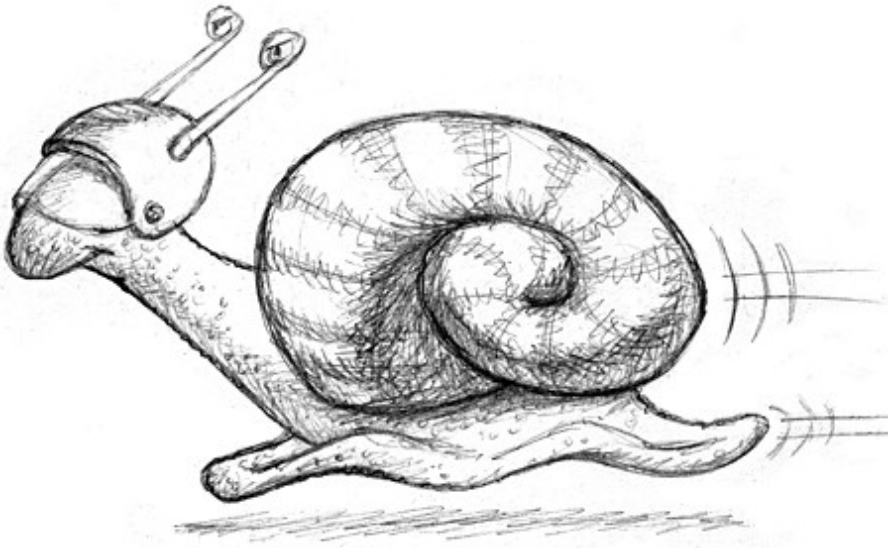
Candidate phrases better than idioms.
(SYN 2000)

Lexikální fixnost – extrahované fráze		
znát slabina	získat kondice	mít koncese
nanést krém	nabírat vzduch	mít rám
∗ najít hrob	∗ zachránit kůže	∗ zamést stopa
vystřelit světlice	zničit úroda	vlastnit karta
být mašina	zrušit pořadník	mít návratnost
mít svědek	nastartovat kariéra	vybrat kombinace
mít jiskra	zatáhnout břicho	mít paluba
být výstup	mít ambice	∗ vidět smrt
mít opletačka	vyhrát poločas	mít důstojnost
vmíchat kůra	získat výpomoc	změnit chod
zlomit odpor	nalévat voda	mít rozchod
dávat inzerát	vytvářet mapa	mít království
mít čelo	vytvářet semeno	mít opice
∗ mít kůže	dělat milost	zachycovat proměna
získat klíč	zastavit voda	být mládež
vypálit rána	natočit šnyt	mít bratranec
∗ najít jmenovatel	vmíchat kečup	mít obojek

Conclusion

- Corpus-based fast generating **idiom candidates.**
- Annotators decide (yes/no decision).
- Faster building the lexicon of idioms.
- Up-to-date language prosessing (no useless idioms anymore).

Questions



Thank you for your attention.