Automatic Identification of Valency Frames in Free Text

Martin Wörgötter

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

3rd December 2016

Outline

- valencies and VerbaLex [1]
- motivation for valency frame indentification
- implementation
- examples
- evaluation using a test set
- conclusions

VerbaLex

- lexical database of $\sim$ 10,500 Czech verbs
- valency frames
- obligatory $\times$ facultative constituents
- morpho-syntactic and semantic information
- synsets, WordNet [2]
- grammar

The task and motivation

Identify verbs and all the corresponding valency frames in a sentence.

- word sense disambiguation
- corpora evidence for valency frames
- enhancing VerbaLex
- e.g. resolving parsing ambiguities [3]
- machine translation

Algorithm and source properties

Input: syntactically and morphologically analysed text
Resources: VerbaLex, Czech WordNet [4] – preprocessed
Output: *vertical text*, with lists of frame identifiers
Speed: $\sim$ 0.03 s per sentence
VerbaLex: patched, added *ct/ci* infinitive variants, enumerated frames
WordNet: created a (`lemma` $\rightarrow$ `list of hypernyms`) dictionary

Algorithm description

- clause scope, order discarded
- a set of tests:
  - reflexivity
  - surface grammar constraints
  - category of personality/impersonality
  - complementation: prepositions, adverbs, subordinate clauses, ...
  - phrasemes
- boolean results
- labeling the constituents with the matched WordNet literals

“*Dokument pozbyl svého významu v roce 1806.*”

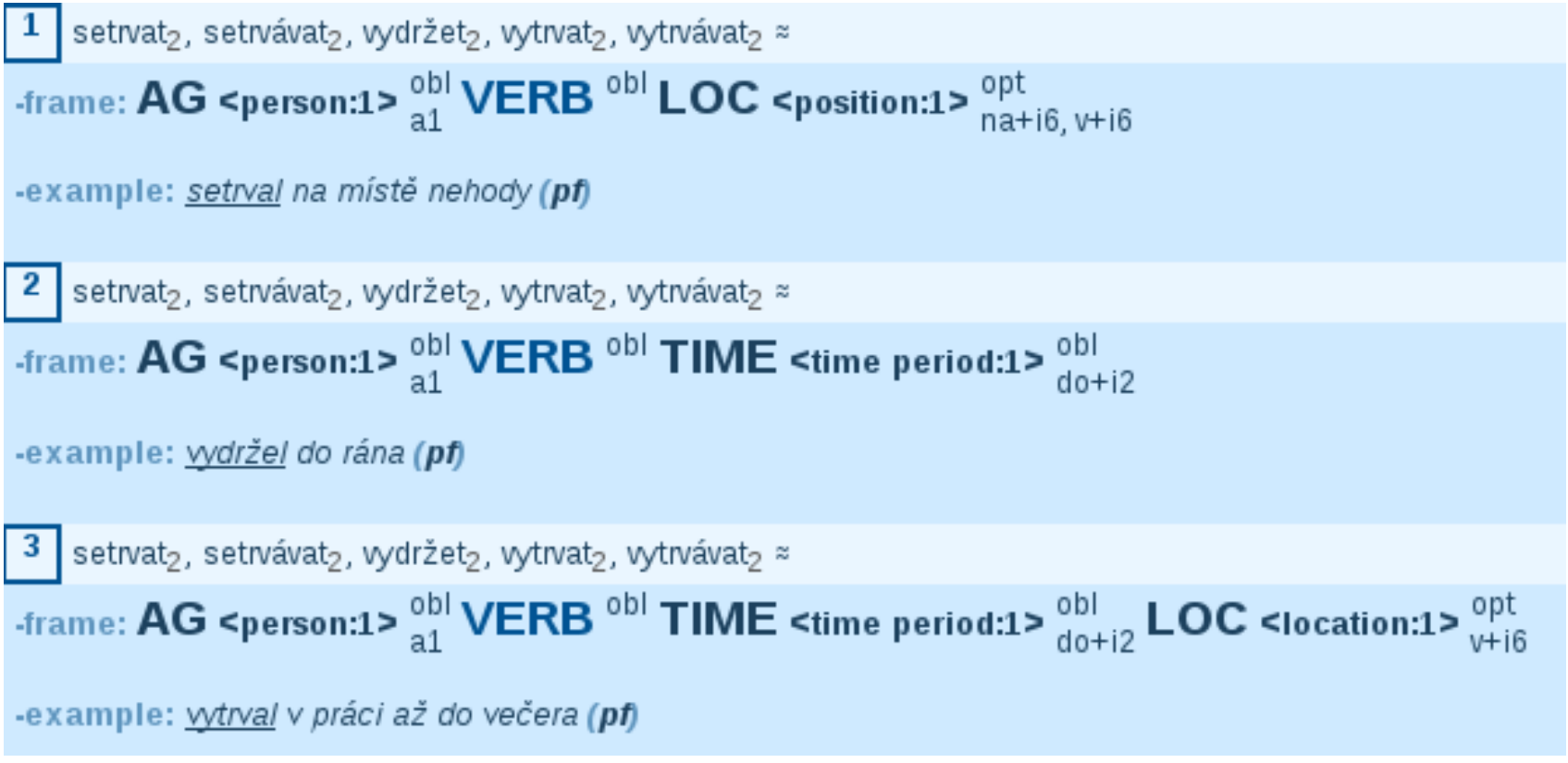
“*The document finally lost its meaning in 1806.*”

40 pozbýt$_2$, pozbývat$_2$ ≈

-frame: **OBJ** <object:1> obl i1 **VERB** obl **ATTR** <attribute:2> obl i2

1/49 frames

“*Konstantinovi vojáci zde setrvali až do poloviny léta 312.*”
“*Constantin’s soldiers rested there until mid-summer 312.*”



3/40 frames, facultative participants not present

“*Během jeho nepřítomnosti vládl nad Soluní jeho bratr Manuel Agelos.*”
“*During his captivity, his brother Manuel Angelos ruled over Thessalonica.*”

13 panovat$_1$, vládnout$_1$ ≈

-frame: **AG** <person:1> obl a1 **VERB** obl **LOC** <location:1> obl nad+i7, v+i6

18 panovat$_1$, vládnout$_1$ ≈

-frame: **AG** <person:1> obl a1 **VERB** obl **PAT** <group:1> obl nad+i7 **LOC** <location:1> opt v+i6

2/18 frames

"Řeky tekoucí ze severu na jih vymílají více východní břeh."
"Rivers flowing from north to south eat away more of the eastern bank."

1 erodovat$_1$, vymílat$_1$ ≈

-frame: **PHEN** <phenomenon:1> | **STATE** <state:4> obl i1 **VERB** obl **SUBS** <soil:2> | **OBJ** <natural object:1> obl i4

-example: *déšť eroduje kámen* (***impf***)
-example: *vlhkost eroduje půdu* (***impf***)

2 erodovat$_1$ ≈

-frame: **SUBS** <soil:2> | **OBJ** <natural object:1> obl i1 **VERB** obl **ATTR** <attribute:2> opt i7

-example: *půda eroduje vlhkem* (***impf***)

No frame accepted, the frames above are available in synset *erode:2, eat away:1, fret:11*.

5 odervat$_1$, odtrhnout$_1$, odtrhovat$_1$, urvat$_3$, utrhnout$_1$ ≈

-frame: **ENT** <stream:1> obl i1 **VERB** obl **OBJ** <natural object:1> obl i4

entity:1>thing:12>body of water:1, water:2>stream:1, watercourse:2

Establishing a test set

Na jejím místě se ve středověku nacházel trh s rybami.

submit

nacházet

Context: Na jejím místě se ve středověku **nacházel** trh s rybami.

Type of annotation: ○ No allowed frame ◉ Matched ○ No match ○ Not a verb ○ Auxiliary ○ Infinitive

najít se nacházet se

Czech Synset: ENG20-02624183-v
definition: *nečekaně se objevit*

1 ☐ nacházet se$_3$, najít se$_1$

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

-example: *našli se i zrádci* ***(pf)***

2 ☐ nacházet se$_3$

-frame: **OBJ** **<object:1>** | **SUBS** **<substance:1>** obl i1 **LOC** **<location:1>** | **ATTR** **<shape:2>** obl vi6

-example: *sůl se nachází ve formě krystalů* ***(impf)***

nacházet se být

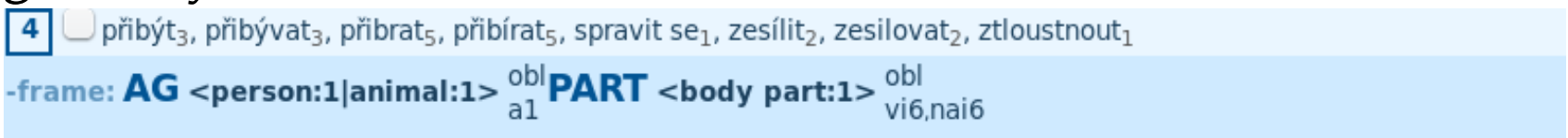
English equivalent: ENG20-02669122-v
definition: *prodlévat v nějakém stavu*

Evaluation

- five test sets
- full agreement required
- IAA of all annotators is $\sim 17\%$
- pairwise agreement reaches 70%
- the algorithm achieves $\sim 30\%$ precision and $\sim 20\%$ recall

Errors of the algorithm

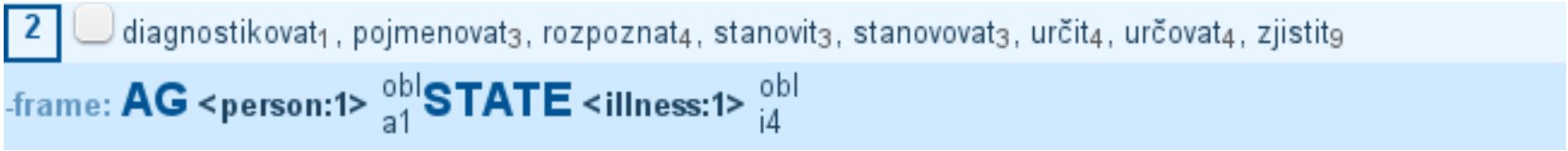
"*Pokud se to ignoruje a pokračuje se v sestupu, bolest v uších stále zesiluje.*"
"*If it is ignored and one continues the descend, the pain in ears gradually increases.*"



- error in syntactic analysis, „*bolest*" is not an *object*
  WordNet definition: „*stát se tlustším, silnějším, mohutnějším, zvýšit svou váhu*" $\sim$ put on weight

Errors of the algorithm

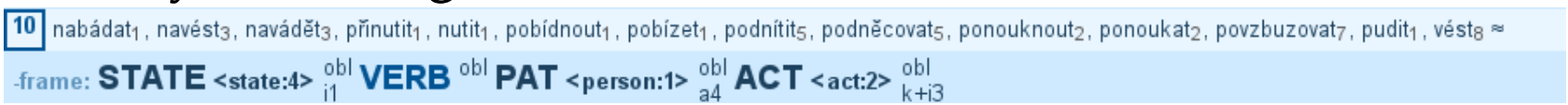
"*Zapomněl jsem heslo pro přístup, kde ho zjistím?*"
"*I have forgotten my password, where can I find it?*"



- no semantic disambiguation using the main clause (the algorithm returns another 23 frames)

Errors of the annotators

"*Jednoho vede vidina odměny a druhého pomsta.*"
"*Someone is encouraged by imagining the reward and someone other by the revenge.*"



- missing obligatory participant $ACT[act:2]^{obl}_{k+i3}$
  (no accepted frame by the algorithm)

Errors of annotators

"Optimální spalovací proces umožňuje získat z paliva maximum energie a zejména snižuje emise oxidu uhličitého."
"The optimal combustion process allows to gain maximum energy from the fuel and in particular reduces carbon dioxide emissions."

- subcategorization features conflict
  (no accepted frame by the algorithm)

Assesing agreement

- unlimited number of assigned frames
- frequent, polysemous verbs
- the type of texts in the test set
- dubitable annotations?

Conclusion and further work

- fast algorithm
- word sense labeling
- analyse the non-acceptance of frames
- validate existing VerbaLex example sentences

Thank you for your attention


HLAVÁČKOVÁ, Dana. *Databáze slovesných valenčních rámců VerbaLex* [online]. 2008 [visited on 2016-05-19]. Available from: `http://is.muni.cz/th/17907/ff_d/`. Disertační práce. Masarykova univerzita, Filozofická fakulta, Brno. Supervised by Karel PALA.


FELLBAUM, Christiane. *WordNet: An Electronic Lexical Database* [online]. Cambridge: MIT Press, 1998 [visited on 2015-11-15]. Language, speech, and communication. Available from: `https://books.google.cz/books?id=Rehu8OOzMIMC`.


JAKUBÍČEK, Miloš. *Enhancing Czech Parsing with Complex Valency Frames* [online]. 2010 [visited on 2016-05-19]. Available from: `http://is.muni.cz/th/172962/fi_m/`. Diplomová práce. Masarykova univerzita, Fakulta informatiky, Brno. Supervised by Aleš HORÁK.


BLAHUŠ, Marek; PALA, Karel. Extending Czech WordNet Using a Bilingual Dictionary. In: CHRISTIANE FELLBAUM, Piek Vossen (ed.). *6th International Global Wordnet Conference Proceedings* [tištěná verze "print"]. Matsue, Japan: Toyohashi University of Technology, 2012, pp. 50–55. ISBN 978-80-263-0244-5.