Building a Multilingual Valency Lexicon

Jana Šindlerová

sindlerova@ufal.mff.cuni.cz

PIRE meeting 11.06.2009

ÚSTAV FORMÁLNÍ A APLIKOVANÉ LINGVISTIKY

Outline

- Motivation
  - Linguistic
  - Experimental
- The Idea
- Work Done So Far
- Work to Be Done

Outline

- Motivation
  - Linguistic
  - Experimental
- The Idea
- Work Done So Far
- Work to Be Done

A Few Quotations as a Starter

- “It is quite possible that there is a universal inventory of possible argument structure constructions relating form and meaning, and that particular languages make use of a particular subset of this inventory.” [Goldberg, 1995]
- “Only by looking at which distinctions are made crosslinguistically can we determine what the semantically relevant aspects of verb meaning are that determine the basis of the clustering into subclasses.” [Goldberg, 1995]

Multilingual ValLex as a Means for Crosslinguistic Research

- an interlinked database of argument structure constructions available for each verb of the language(s)
- enables crosslinguistic comparison of valency structures
- based on parallel corpora – a large source of linguistic data

Linguistic Theories Used

- Functional Generative Description
  - framework used for valency description in PDT-VALLEX and EngValLex
- Construction Grammar
  - an inspirative approach to argument structure which we use in order to improve the structure of MultiValLex

Multilingual ValLex as a Means for MT Experiments

- New utilizable feature
- Helping with diverse behavior
- Valency as the core of syntactic dependency relations
- Mapping of arguments – mapping of structures

Applications Used

- The „Prague“ corpora – PCEDT, the „vallexes“
- Tred, PDT-VALLEX frame-editor
- Tecto-MT, t-trees aligner of David Mareček et al.

Outline

- Motivation
  - Linguistic
  - Experimental
- The Idea
- Work Done So Far
- Work to Be Done

The Idea

- Let’s have an interlinked translational electronic valency lexicon
- Take “the P(CE)DT-vallexes”
- Make links between the corresponding frames
- Make links between the corresponding arguments
- Add the semantic verb classification
- First for Czech-English

Outline

- Motivation
  - Linguistic
  - Experimental
- The Idea
- Work Done So Far
- Work to Be Done

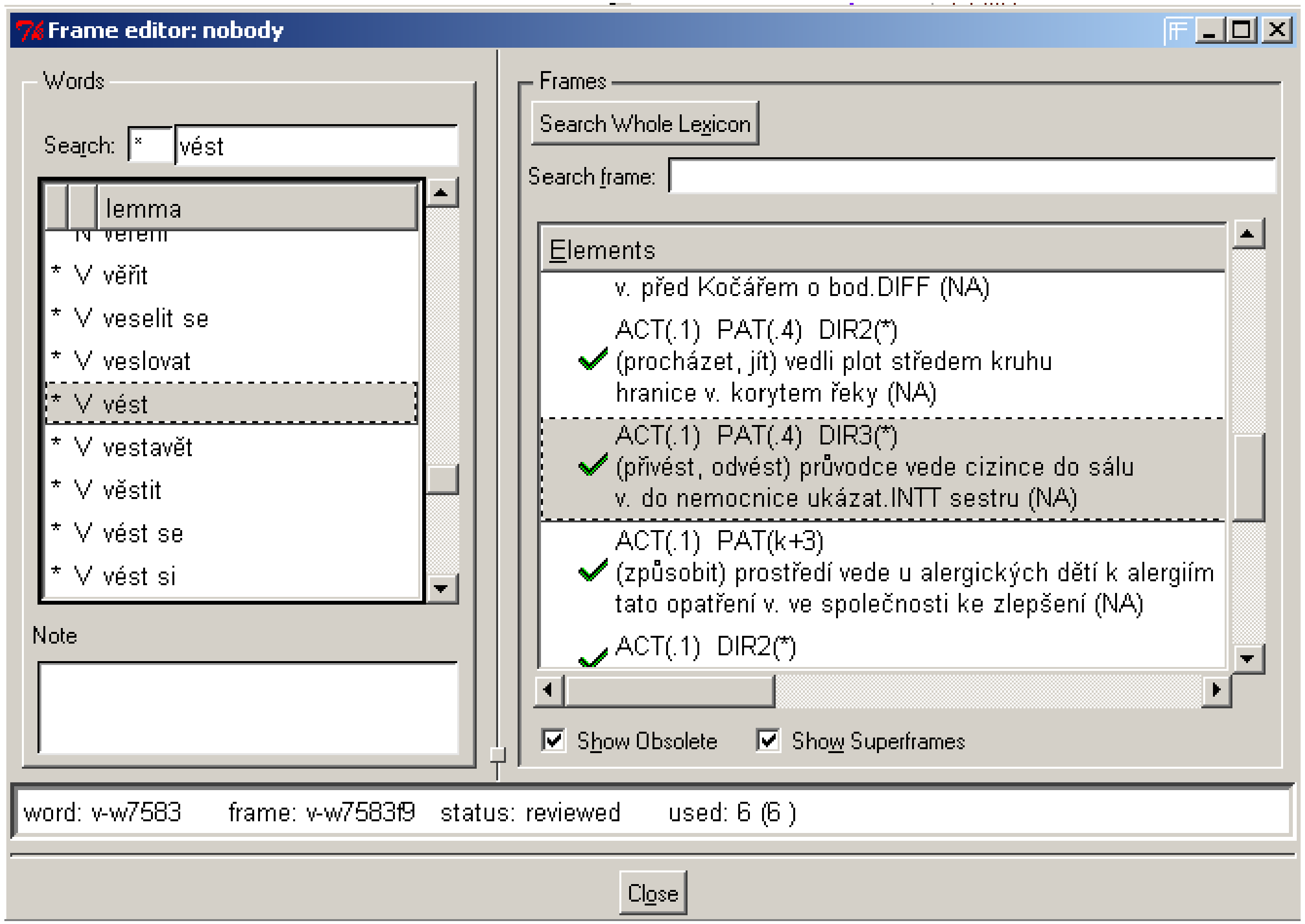
PDT-VALLEX

- 9191 valency frames
- 5510 verbs

EngValLex

- 6006 valency frames
  - Less data
- 3576 verbs
  - Less prefixation, no aspect counterparts etc.

PDT-VALLEX



Search: * vést

- N věření
- * V věřit
- * V veselit se
- * V veslovat
- * V vést
- * V vestavět
- * V věstit
- * V vést se
- * V vést si

Search frame:

v. před Kočářem o bod.DIFF (NA)

✓ ACT(.1) PAT(.4) DIR2(*)
(procházet, jít) vedli plot středem kruhu
hranice v. korytem řeky (NA)

✓ ACT(.1) PAT(.4) DIR3(*)
(přivést, odvést) průvodce vede cizince do sálu
v. do nemocnice ukázat.INTT sestru (NA)

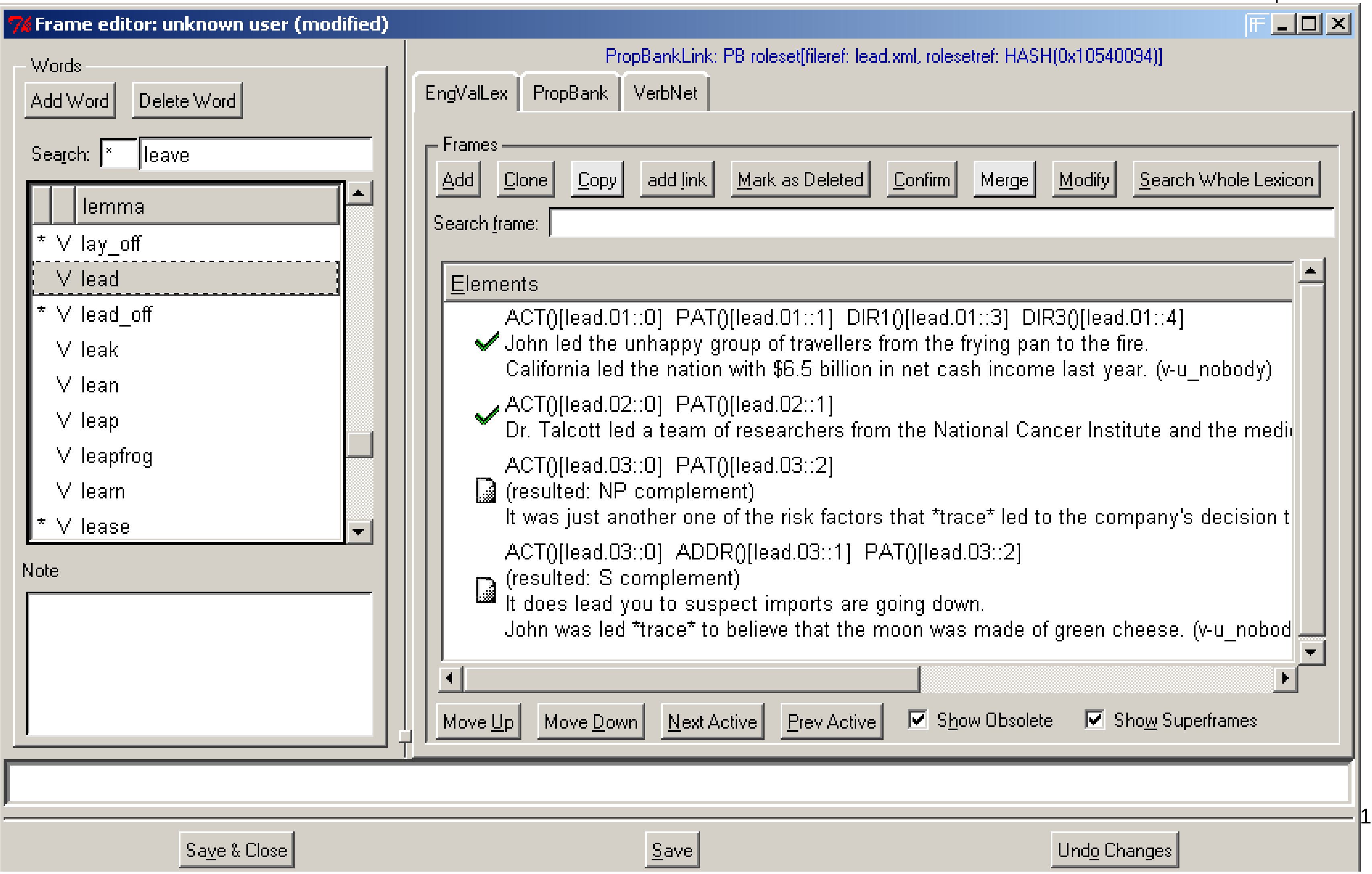
✓ ACT(.1) PAT(k+3)
(způsobit) prostředí vede u alergických dětí k alergiím
tato opatření v. ve společnosti ke zlepšení (NA)

✓ ACT(.1) DIR2(*)

☑ Show Obsolete ☑ Show Superframes

word: v-w7583 frame: v-w7583f9 status: reviewed used: 6 (6)

EngValLex



PropBankLink: PB roleset[fileref: lead.xml, rolesetref: HASH(0x10540094)]

✓ ACT()[lead.01::0] PAT()[lead.01::1] DIR1()[lead.01::3] DIR3()[lead.01::4]
John led the unhappy group of travellers from the frying pan to the fire.
California led the nation with $6.5 billion in net cash income last year. (v-u_nobody)

✓ ACT()[lead.02::0] PAT()[lead.02::1]
Dr. Talcott led a team of researchers from the National Cancer Institute and the medi

ACT()[lead.03::0] PAT()[lead.03::2]
(resulted: NP complement)
It was just another one of the risk factors that *trace* led to the company's decision t

ACT()[lead.03::0] ADDR()[lead.03::1] PAT()[lead.03::2]
(resulted: S complement)
It does lead you to suspect imports are going down.
John was led *trace* to believe that the moon was made of green cheese. (v-u_nobod

“Vallexes” in comparison

- unbalanced
  - PDT-VALLEX as a standard – bigger, older, refined
  - EngValLex – based on PropBank, based on different amount and type of data
- now proceeding refinement of EngValLex
  - adding new frames, adding new verbs, phrasal verbs adjustment, control and adjustment of existing frames
  - supposed deadline for refinement – September 2009

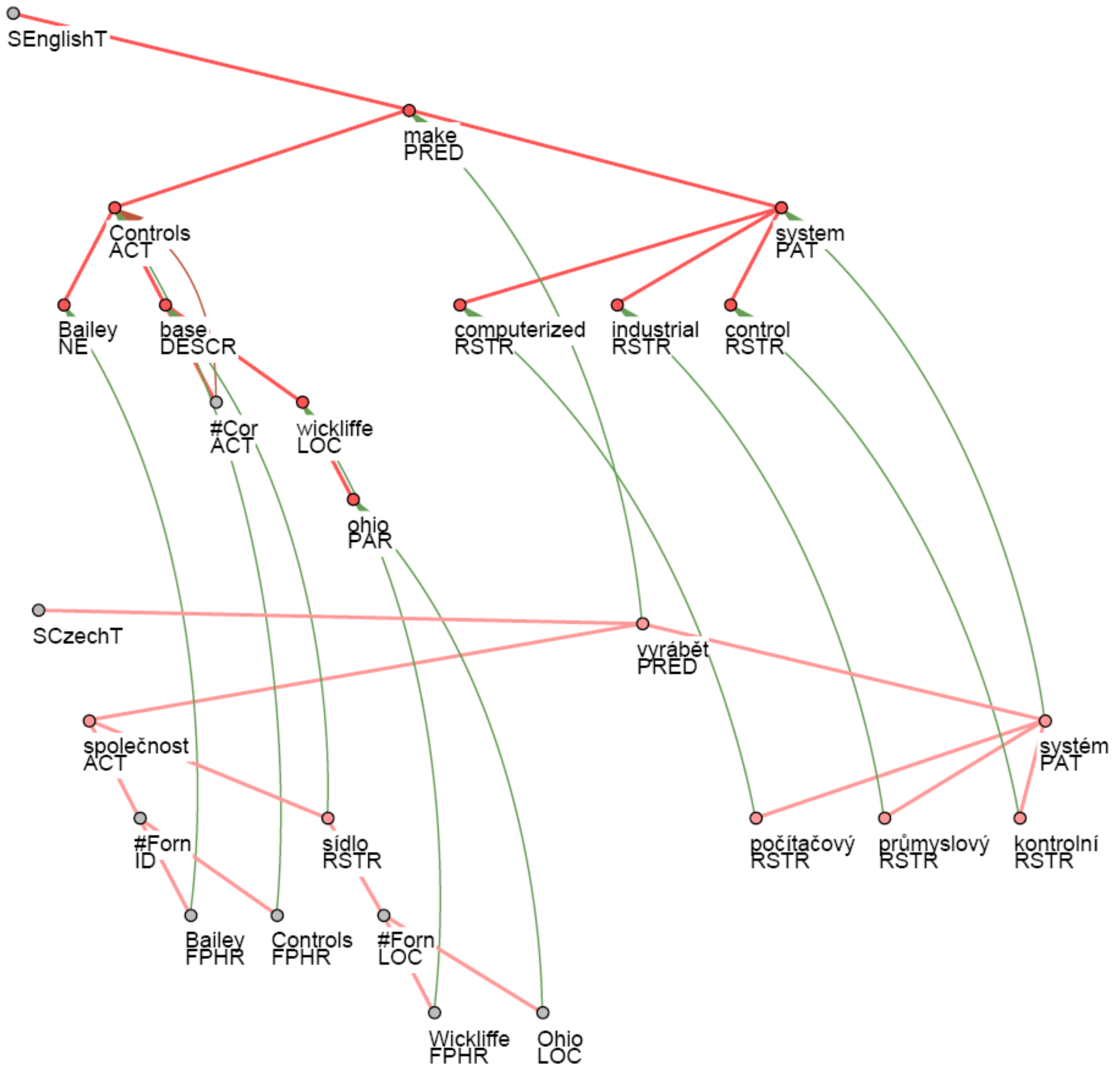
Linking

- Both directions: Czech-English, English-Czech
  - bring: přinést, zavést, přilákat...
  - přinést: bring, serve, take...
- Automatic procedure (2009)
  - Based on manually annotated t-data from PCEDT
  - Using automatic alignment of nodes in t-trees
- Manual completion (2010)

Automatic alignment

- t-layers of both parts of PCEDT
- files annotated manually and controlled in both corpora
- 11547 sentences, about 200000 nodes
- about 1/5 of the PCEDT
- alignment based on GIZA++, using set of manually designed features (both lexical and structural)
- filelists of translations of individual verbs

Bailey Controls, based in Wickliffe, Ohio, makes computerized industrial controls systems....
Bailey Controls, sídlem v Wickliffe, Ohio, vyrábí počítačové průmyslové kontrolní systémy.



Verbs in the aligned data

- 26881 Czech t-nodes marked v
- 2858 unique Czech verbal t-lemmas
- 30889 English t-nodes marked as v
  - participles and gerunds
- 2492 unique English verbal t-lemmas
- The alignment will be used as the basis for automatic extraction of translational equivalents and for the linking of their arguments

Outline

- Motivation
  - Linguistic
  - Experimental
- The Idea
- Work Done So Far
- Work to Be Done

Work to Be Done

- Incorporation of the automatic alignment information into the vallexes
- Manual annotation of additional links
- Linguistic research leading to
- Verb Class Assignment
- MT Experiments made on PCEDT