****

CURRENT STATUS OF THE PDEV PROJECT

Outline of the presentation

Prague, Dec 7, 2010

by Martin Holub and Lenka Smejkalova

Institute of Formal and Applied Linguistics
Charles University in Prague

*** 1) Prerequisites

We aim at building PDEV as an NLP-applicable source. To check if PDEV can be useful for NLP we need a resonable sample of PDEV data that is

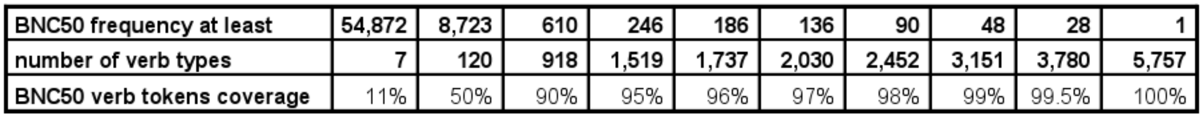
```
   * consistent in all main components, i.e.
       - pattern database
       - manually tagged reference samples of corpus data
       - system of semantic types

   * representative in the sense of corpus coverage

   * clear enough so that trained humans are able to achieve a
     reasonable degree of inter-annotator agreement on corpus data
```

```
!!! This is what we need to show that "PDEV can work well"!
    Such a test should be "statistically significant"!
```

*** 2) Verbs in BNC50 and the current PDEV

* Basic BNC50 statistics

- The total number of lexical verb tokens is 4,673,003.



BNC50 frequency at least	54,872	8,723	610	246	186	136	90	48	28	1
number of verb types	7	120	918	1,519	1,737	2,030	2,452	3,151	3,780	5,757
BNC50 verb tokens coverage	11%	50%	90%	95%	96%	97%	98%	99%	99.5%	100%

Table 1. The coverage of BNC50 verb tokens. For example, 918 most frequent verbs, each of which occurs at least 610 times in BNC50, cover more than 90% of all BNC50 lexical verb tokens.

Table 1 shows, among other things, the fact that verbs with f < 250 cover only about 5% of all lexical verb tokens in BNC50 corpus.

* Existing complete PDEV entries and the corpus coverage

Table 2 shows the number of existing PDEV entries with status “complete” (checked by Hanks) and the corresponding number of patterns created.

	verb entries	patterns
all	695	2,662
f ≥ 50	419	2,136
f ≥ 100	266	1,679
f ≥ 150	213	1,464
f ≥ 200	179	1,324
f ≥ 250	165	1,247
f ≥ 300	147	1,170
f ≥ 350	135	1,076

Table 2. The set of current complete verbs and their frequency in BNC50.

	tokens	BNC50 coverage
all	495,553	10.61%
f < 250	32,206	0.69%
f < 300	37,148	0.80%
f < 350	41,056	0.88%
f ≥ 250	463,347	9.92%
f ≥ 300	458,405	9.81%
f ≥ 350	454,497	9.73%

Table 3. How the current set of complete verbs covers BNC50 corpus.

* **Conclusion:** In the current PDEV there are (only) 100-200 complete verb entries that are applicable for our experiments designed for PDEV validation.

*** 3) Inconsistencies in the current PDEV data

* A) Inconsistencies in the current pattern database

- Several types of inconsistency have been detected
  - data written in fields designed for different kind of data
  - inconsistent coding - separators, etc. (..., "|", ",")
  - chaoticly written data, for which there were no systematic fields

- Some mistakes are "systematic", and those can be corrected easily.

- Some mistakes were done "intentionally", because the PDEV form did not provide options to encode the needed data systematically.

- Conclusion: Thorough manual revision of all patterns is necessary for serious experiments. The revision will go hand in hand with copying the entries into the PDEV2 format (see below).

* B) Inconsistencies in manually tagged reference corpus data

- significant disagreement in tagging between Patrick and "historical Patrick" on a sample of complete verbs (in the beginning of 2010)

- In our opinion the main (natural) sources of inconsistency in tagged data are
  - the historical development (changes) of the CPA method
  - occasional (minor) shifts in the interpretion of PDEV patterns
  - (mainly:) missing written rules for tagging

- Conclusion: Thorough revision of the existing reference sample data is necessary. The revision should be based on
  - the currently already existing "guidelines for annotators"
  - revised patterns in the PDEV2 form (see A))

* C) Inconsistencies in using sematic types

- have not been explored/mapped yet

*** 4) Steps towards further systematic development

*** A) Documentation of both PDEV components and the related procedures**

- is necessary for consistent work (especially in a team)
- should consist of

  * "Guidelines for PDEV Lexicographers" - to improve the consistency of patterns - two parts:
    - procedural part = how lexicographers should work when they create a PDEV entry
    - technical part = how lexicographers should use the PDEV form to write PDEV patterns properly, vcetne definic lingvistickych kategorii a prikladu

  * Documentation/definitions of Semantic Types

  * "Guidelines for PDEV Annotators" - to improve the consistency of both pattern interpretation and the manually tagged data

  * Technical report on PDEV validation = the description and the results of performed experiments, especially
    - the degree of inter-annotator agreement
    - analysis of both frequency and sources of disagreement

  * Technical specification of PDEV forms (describes even the implementation of the pattern database, including dtd schema)

*** B) Validation and correction**

- Each PDEV entry in the test sample should be validated using the IAA test.

- In case of significant amount of disagreement (if better pattern definitions do not help):
  -> Analyse the types/sources of disagreement and modify the method. Then repeat the test.
  * The method can be modified by
    a) a change in the pattern structure (PDEV patterns form), or
    b) a change of the metody of pattern writing (Guidelines for Lexicographers), or
    c) a change in the interpretation of existing patterns (Guidelines for Annotators)

- Currently we are training two anotators. Our experience shows that the training is demanding and time consuming, but without that the "good" IAA seems to be impossible.

- **Conclusion:** Documentation and validation of the PDEV data is our current goal. First "pilot validation test" is planned to be done in January.

  Without a serious empirical test, the NLP community cannot recognize and will not believe that PDEV is a valuable source for NLP. To perform such a test we need a "reasonable" sample of consistent PDEV data, which, however, is not available yet (in the existing PDEV database stored in Brno).

*** 5) The design of PDEV2 form

- **the current specification**
  - the layout
  - the XML specification: includes the technical part of Guidelines for Lexicographer

- **the current implementation**

- **examples of some differences between the "original PDEV" and PDEV2**

*** 6) What has been done since last year

* We have written **Guidelines for Annotators**. Silvie and Patrick agreed on the final version that has already been published on the "official" CPA web pages.

* We have designed and implemented a **new PDEV web form** that provides lexicographers with all they need to consistently describe PDEV patterns. As the number of changes/improvements is quite big, we call it "PDEV2". Currently we are testing the implementation.

* We have hired and are training **two qualified annotators**. In January they should be ready to perform IAA test on a sample of test verbs.

* We have designed and implemented infrastructure tools for **generating and storing random samples** of corpus verb occurrences. Those tools are necessary to make serious experiments and to have possibility to analyse the causes of disagreement.

* We have developed a **tool for analysing verb arguments** in manually tagged sentences (where the verb was assigned a pattern). Its output is a sketch of nouns that are likely to form a semantic type.

* We have developed a **simple pattern recognizer** - just to have a baseline for further experiments.

*** 7) Future work

* **A) The nearest future: First validation attempt:**
  - in January 2011
  - 10-20 "representative" sample verbs
  - PDEV data with revised consistency
    - revised patterns in the PDEV2 form
    - revised random reference samples
  - 2 annotators, 50 random occurrences per verb

* **B) Directions of further research in 2011**
  - integration of PDEV data with existing resources at UFAL
  - evaluation in the machine translation framework