Reference Data for Czech Collocation Extraction

MWE 2008 Shared Task Resource

Pavel Pecina

pecina@ufal.mff.cuni.cz

Institute of Formal and Applied Linguistics
Charles University, Prague

June 1, 2008

Introduction

Motivation

- to create the reference data set for empirical evaluation of methods for extraction of Czech collocations

Evaluation data sets

1. dependency (syntactical) bigrams from *Prague Dependency Treebank* (*PDT-Dep*)
2. surface (adjacent) bigrams from *Prague Dependency Treebank* (*PDT-Surf*)
3. instances of *PDT-Surf* in *Czech National Corpus* (*CNC-Surf*)

Main features

- annotated as collocational and non-collocational and also assigned to finer-grained categories
- associated with corpus frequency information for easy computation of AM scores
- publicly available from the MWE wiki page `http://multiword.wiki.sourceforge.net/`.

Outline

Prague Dependecy Treebank 2.0

- developed by the Institute of Formal and Applied Linguistics and the Center for Computational Linguistics, Charles University, Prague
- 1 504 847 tokens in 87 980 sentences and 5 338 documents
- complex and interlinked annotation on morphological, analytical (surface syntax), and tectogrammatical (deep syntax) layer
- the annotation is based on the long-standing Praguian linguistic tradition, adapted for the current Computational Linguistics research needs
- available from LDC (catalog number LDC2006T01)
- also available for MWE Shared Task purposes from CU directly

Czech National Corpus

- a project with the aim to build up a large corpus, containing mainly written Czech developed at Institute of CNC, Charles University, Prague
- SYN 2000 and 2005 synchronous corpora containing 242 million tokens
- no manual annotation (no morphology, no syntax)
- automatically assigned part-of-speech taggs (96% accuracy)

genre	*SYN2000*	*SYN2005*
fiction	15 %	40 %
technical literature	25 %	27 %
newspaper, journals	60 %	33 %

PDT Morphological layer

- each word form (token) is assigned a lemma and a morphological tag

Lemma (two parts)

1. *lemma proper* - a unique identifier of the lexical item possibly followed by a number distinguishing different lemmas with the same base forms
2. *technical suffix* - containing additional information about the lemma (semantic or derivational information) – optional.

Morphological tag

- is a string of 15 characters where every position encodes one morphological category using one character

```
<f> ničení <l> ničení_(*3it) <t> NNNS2-----A----
```

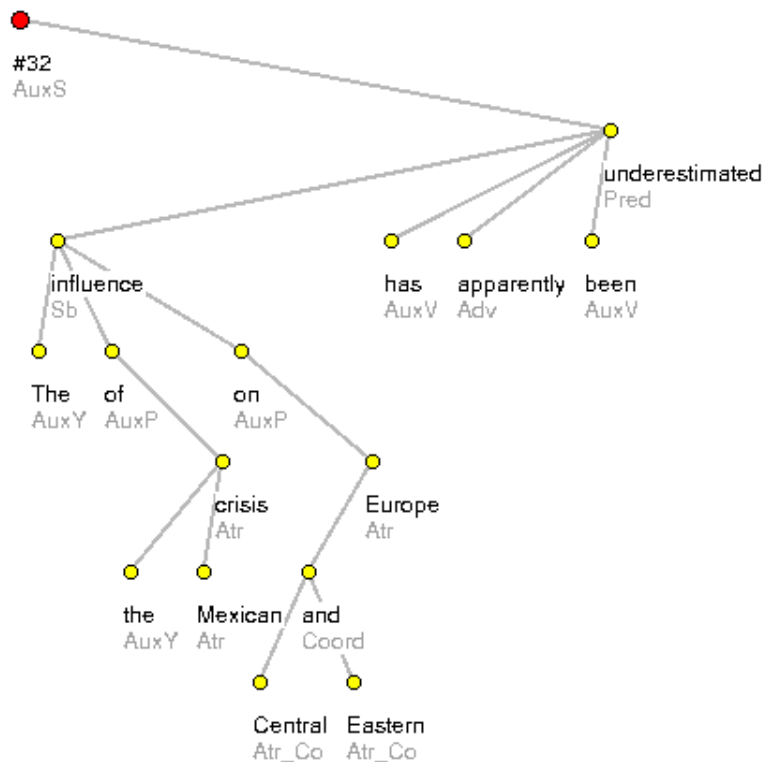
PDT Morphological categories

Pos	*Name*	*Description*	# *Values*
1	**POS**	**Part of speech**	**12**
2	SubPOS	Detailed part of speech	60
3	**Gender**	**Gender**	**9**
4	Number	Number	5
5	Case	Case	8
6	PossGender	Possessor's gender	4
7	PossNumber	Possessor's number	3
8	Person	Person	4
9	Tense	Tense	5
10	**Grade**	**Degree of comparison**	3
11	**Negation**	**Negation**	2
12	Voice	Voice	2
13	Reserve 1	Reserve	-
14	Reserve 2	Reserve	-
15	Var	Variant, style	10

(tagset size: $\sim$ 5000)

PDT Analytical layer

- encoding sentence *dependency structures*
- each word is linked to its *head word* and assigned its *analytical function* (dependency type)
- dependency structure is a tree – a directed acyclic graph having one root



PDT Analytical functions

Afun	*Description*
`Pred`	Predicate, a node not depending on another node
`Sb`	Subject
`Obj`	Object
`Adv`	Adverbial
`Atr`	Attribute
`AtrAtr`	An attribute of any of several preceding (syntactic) nouns
`AtrAdv`	Structural ambiguity between adverbial and adnominal dependency
`AdvAtr`	Dtto with reverse preference
`AtrObj`	Structural ambiguity between object and adnominal dependency
`ObjAtr`	Dtto with reverse preference
`Atv`	Complement (determining), hung on a non-verb. element
`AtvV`	Complement (determining), hung on a verb, no 2nd gov. node
`Pnom`	Nominal predicate, or nom. part of predicate with copula *be*
`Coord`	Coordinated node
`Apos`	Apposition (main node)
`ExD`	Main element of a sentence without predicate, or deleted item
`AuxV`	Auxiliary vb. *be*
`AuxT`	Reflex. tantum
`AuxR`	Ref., neither Obj
`AuxP`	Primary prepos., parts of a secondary p.
`AuxC`	Conjunction (subord.)
`AuxO`	Redundant or emotional item, 'coreferential' pronoun
`AuxZ`	Emphasizing word
`AuxX`	Comma (not serving as a coordinating conj.)
`AuxG`	Other graphic symbols, not terminal
`AuxY`	Adverbs, particles not classed elsewhere
`AuxK`	Terminal punctuation of a sentence

Morphological normalization

- Goal: to canonize morphological variants of words so each collocation can be identified regardless its actual morphological form.
- pure *lemmatization* (using lemmas instead words) not adequate
  (cf. secure area – insecure area, big mountain – (the) highest mountain)
- our apporach: transforming words into combination of:
  1. *lemma proper* – technical suffixes of lemma ignored
  2. *reduced tag* – comprising: *part-of-speech*, *gender*, *grade*, and *negation*.

Morphological normalization: example

Surface form

Id	*Form*	*Lemma*	*Full Tag*	*Parent Id*	*Afun*
1	Zbraně	zbraň	NNFP1-----A----	0	ExD
2	hromadného	hromadný	AANS2----1A----	3	Atr
3	ničení	ničení_^(*3it)	NNNS2-----A----	1	Atr

Normalized form

Id	*Lemma Proper*	*Reduced Tag*	*Parent Id*	*Afun*
1	zbraň	NF-A	0	Head
2	hromadný	AN1A	3	Atr
3	ničení	NN-A	1	Atr

Part-of-speech filtering

Justeson and Katz (1995): focus on precision

- the collocation candidates are passed through a filter which only lets through the patterns that are likely to be 'phrases' (potential collocations)
- patterns suggested A:N (adjective–noun) and N:N (noun–noun)
- a simple heuristics that improves results of collocation extraction methods

Our approach: focus on recall

- filter out candidates having POS patterns that *never* form a collocation (to keep the cases with POS patterns that can *possibly* form a collocation)

Part-of-speech filtering

Pattern	*Example*	*Translation*
A:N	trestný čin	*criminal act*
N:N	doba splatnosti	*term of expiration*
V:N	kroutit hlavou	*shake head*
R:N	bez problémů	*no problem*
C:N	první republika	*First Republic*
N:V	zranění podlehnout	*succumb*
N:C	Charta 77	*Charta 77*
D:A	volně směnitelný	*free convertible*
N:A	metr čtvereční	*squared meter*
D:V	těžce zranit	*badly hurt*
N:T	play off	*play-off*
N:D	MF Dnes	*MF Dnes*
D:D	jak jinak	*how else*

A – adjectives, N – nouns, C – numerals, V – verbs,
D – adverbs, R – prepositions, T – particles

Frequency filtering

- limit on bigrams occurring more than five times.
- motivation: not to bias the evaluation
- the less frequent candidates do not meet the requirement of sufficient evidence of observations needed by some methods

Candidate Data Sets

PDT-Dep

- 12 232 dependency bigrams from PDT consisting of a normalized head word and its modifier, plus their dependency type

PDT-Surf

- 10 021 surface bigrams (pairs of adjacent words) from PDT consisting of normalized components
- 974 of these bigrams do not appear in *PDT-Dep* test sets (if we ignore the syntactical information)

CNC-Surf

- 9 868 surface bigrams from PDT occuring in SYN2000 and SYN2005
- 153 do not occur in SYN2000 and SYN2005 corpora more than five times

Manual annotation

Definition:

"A collocation expression is a syntactic and semantic unit whose exact and unambiguous meaning or connotation cannot be derived directly from the meaning or connotation of its components." Choueka (1988)

- The annotation was performed independently by three experts without knowledge of context
- The annotators were instructed to judge any bigram which could *eventually* appear in a context where it has a character of collocation, as a *collocation*.
- During the annotation the annotators also attempted to classify each collocation into one of the following categories.

Annotation categories

1. stock phrases, frequent unpredictable usages
   zásadní problém (major problem), konec roku (end of a year)
2. names of persons, organizations, geographical locations, and other entities
   Pražský hrad (Prague Castle), Červený kříž (Red Cross)
3. support verb constructions
   mít pravdu (to be right), činit rozhodnutí (make decision)
4. technical terms
   předseda vlády (prime minister), očitý svědek (eye witness)
5. idiomatic expressions
   studená válka (cold war), visí otazník (hanging question mark ∼ open question)

- not intended as a result of the process but rather as a way how to clarify and simplify the annotation
- any bigram assigned to any of the categories by all annotattors we considered a collocation

Interannotator agreement

Agreement scores

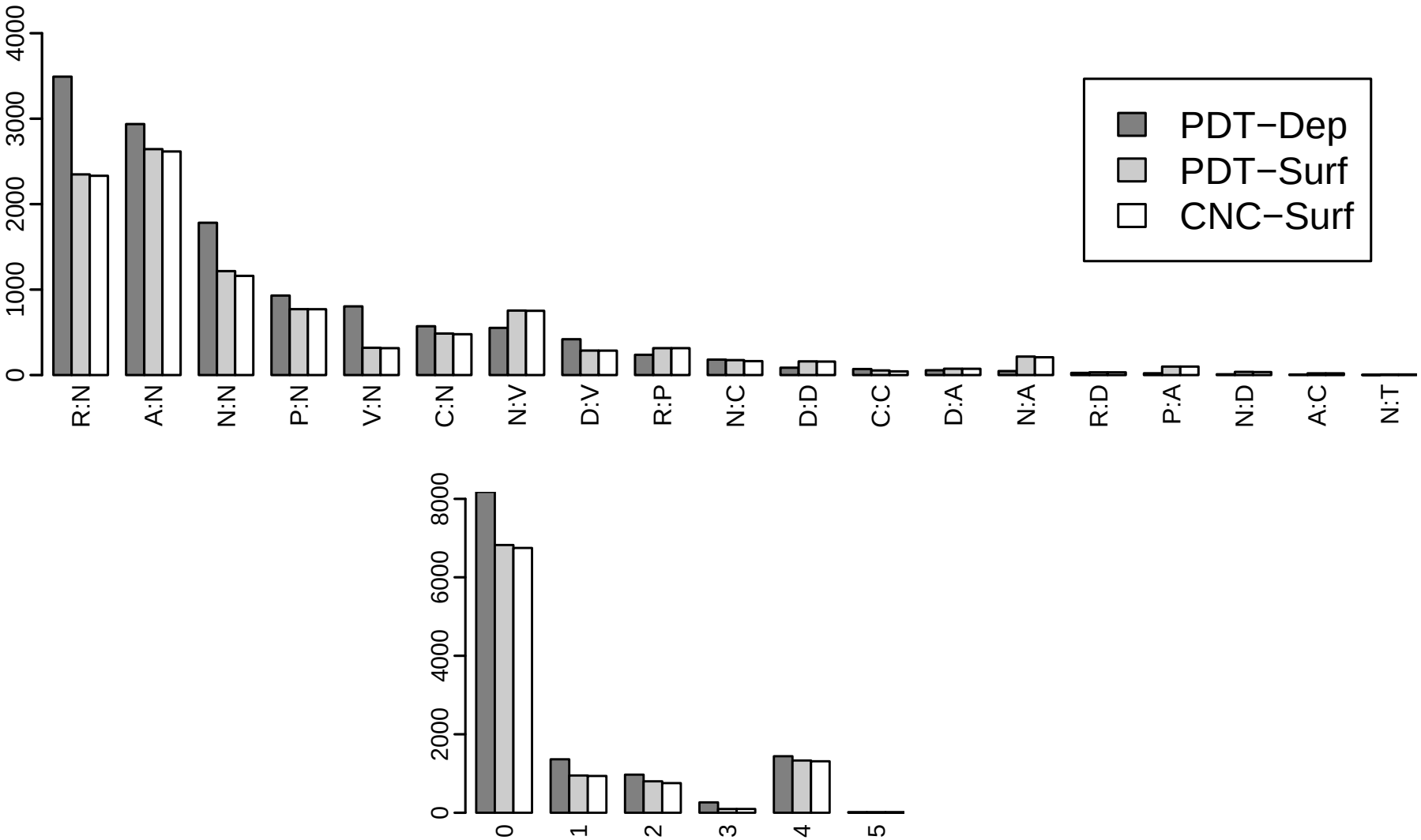
annotations	*fine grained*		*merged*	
	accuracy	*Fleiss' κ*	*accuracy*	*Fleiss' κ*
A1–A2	72.1	0.49	79.5	0.55
A2–A3	71.1	0.47	78.6	0.53
A1–A3	75.4	0.53	82.2	0.60
A1–A2–A3	61.7	0.49	70.1	0.56

Confusion matrices (fine grained and merged categories)

	0	*1*	*2*	*3*	*4*	*5*
0	7 066	644	135	78	208	3
1	590	265	125	0	96	0
2	13	8	621	0	46	1
3	74	0	1	185	0	0
4	409	442	87	0	1075	7
5	25	3	2	2	15	6

	0	*1*
0	7 066	1 068
1	1 111	2 987

Annotation: POS pattern and category distribution



Summary statistics

Reference Data Set	*PDT-Dep*	*PDT-Surf*	*CNC-Surf*
sentences	87 980		15 934 590
tokens	1 504 847		242 272 798
words (no punctuation)	1 282 536		200 498 152
bigram types	635 952	638 030	30 608 916
after frequency filtering	26 450	29 035	2 941 414
after part-of-speech filtering	12 232	10 021	1 503 072
collocation candidates	12 232	10 021	9 868
sample size (%)	100	100	0.66
true collocations	2 557	2 293	2 263
baseline precision (%)	21.02	22.88	22.66

Thank you!