CzeSL – an error tagged corpus of Czech as a second language

Barbora Štindlová[1] Svatava Škodová[1]
Jirka Hana[2] Alexandr Rosen[2]

[1]Technical University, Liberec, Czech Republic
[2]Charles University, Prague, Czech Republic

PALC 2011
Practical Applications in Language and Computers
Łódż, 13–15 April 2011

Outline of the talk

Outline of the talk

Learner Corpus (LC)

- A computerized textual database of language as produced by second/foreign language learners (Leech 1998)
- Differs from national corpora:
  - not a representative repository of contemporary language
  - but a repository of **interlanguage**, which is dynamic, varied

Research value of LC

- Language data for the research of *interlanguage*:
  - regularities
  - factors
  - development

CzeSL – a learner corpus of Czech

- First learner corpus of Czech
- For other Slavic languages – Slovene: PiKUST, ... ?
- Part of an acquisition corpus project – *AKCES*
- Other parts: native speakers' classroom language: oral (SCHOLA), written (SKRIPT)

Planned extent in 2012

- 2 million words
- 4 subcorpora according to the learners' L1:
  - Related Slavic language: Russian, Polish
  - Non-Slavic Indo-European language: German, English, French
  - Non-related language: Vietnamese, Arabic
  - L1/2: Romani

Features of CzeSL

- Written and spoken texts
- Original texts – handwritten
- All proficiency levels according to CEFRL
- Various genres and topics
- Metadata on the learner and the task (18 items)

Error annotation

- About 46% of existing LC are annotated
- Partial error annotation:
  - Pronunciation (LeaP)
  - Orthography (TLEC)
  - Syntax (AleSKO)
- Complex error annotation:
  ICLE, FRIDA. FALCO, NICT JLE, CzeSL

Error annotation in CzeSL

- Issues in Czech: rich inflection, derivation, complex agreement rules and information-structure-driven constituent order
- The answer: multi-level annotation scheme
  - Combination of manual and automatic annotation

Automatic annotation

- Automatic assignment of error tags wherever possible, based on comparing faulty and corrected forms
- Standard morphosyntactic tagging and lemmatization

Annotation scheme

- Multi-level design
  – two-stage annotation, three levels,
  allows for:
  - Successive emendation
  - Annotating errors in both single forms and discontinuous strings

Levels of annotation

LEVEL 0

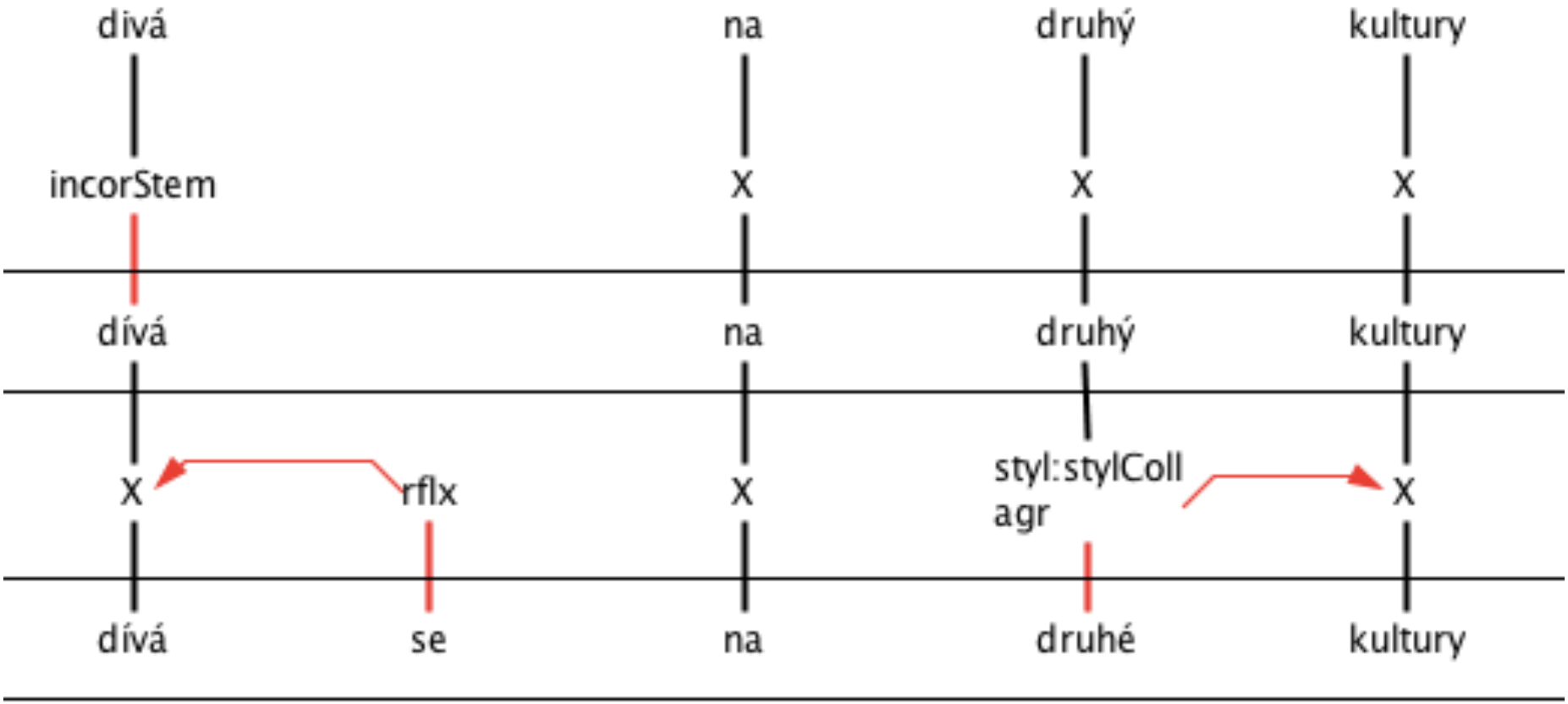
- Transcribed input

LEVEL 1

- Orthographical and morphological emendation of isolated forms
- Result:
  - String of existing Czech forms
  - Sentence as a whole can still be incorrect

LEVEL 2

- All other types of errors
- Syntactic, lexical, word order, usage, style, reference, negation, overuse/underuse of syntactic items
- Result: grammatically correct sentence


dívá
incorStem
na
X
druhý
X
kultury
X
dívá
na
druhý
kultury
X
rflx
X
styl:stylColl
agr
X
dívá
se
na
druhé
kultury

Taxonomy of errors

- 2 stages of error emendation
- Minimal intervention in the original
- 22 manually added tags + 10 automatic error tags

Outline of the talk

Sample

- 67 texts, about 150 words each
- 9373 tokens
- 7995 words (excluding punctuation)
- CEFRL level A2–B1
- Various L1s
- 14 annotators, each text by two

A measure of IAA: Kappa

- A naive measure: identical choices / number of choices
- Kappa penalizes cases with fewer choices (agreement by chance is higher)
- Kappa = 1 – perfect agreement
- Kappa = 0 – random agreement
- Kappa > 0.4 – reasonable

IAA results

on 9848 tokens

Tag	A1 only	A2 only	Both A1 and A2	Kappa
incor	168	130	894	0.84
incorStem	167	165	559	0.75
incorInfl	173	130	250	0.61
wbd	14	21	45	0.72
fw	25	17	18	0.46
agr	82	99	110	0.54
dep	99	118	87	0.43
neg	11	9	9	0.47
styl	19	14	10	0.38
lex	107	131	74	0.37
use	60	74	19	0.21
sec	45	18	4	0.11

Examples of high IAA

Agreement error

kappa = 0.54

(1) Viděl malého Petra

(2) Viděl *malou Petra

Why not still higher?

Different emendations

L0: *Věci budou *težki*
A1 – L1: *těžký*, L2: *těžké* + AGR
A2 – L1: *těžké*, L2: *těžké*

Wrong choice of a tag

due to misunderstanding of a grammar concept by the annotator: agreement vs. valency

(3) *kvůli jeho *životním/životnímu stylu*
'for his lifestyle'

(4) *každý *muset/musí řešit ten problém*
'everyone has to solve the problem'

Examples of low IAA

Lexical error

kappa = 0.37

Due to semantic proximity of lexemes annotators disagree about the need for correction:

(5) *když se dívám na *?druhý/jiný kultury*
'when I look at other cultures'

On the other hand, some lexemes are distant enough and annotators agree about the need for for correction:

(6) **housenky/housky kupuju v pekařství*
'I buy caterpillars in the baker's shop'

Some reasons for low IAA

- Errors of type **lex** involve a high degree of subjective judgement, thus cannot aim at high IAA.
- Errors of type **sec** – highly formal specific, due to primary errors.

Outline of the talk

Questions

- How far can we get without manual annotation?
- Does it make sense to use morphosyntactic taggers, parsers, spell-checkers on both emended and ill-formed input?
- So far, we tried two taggers and a spell-checker.

Taggers

Taggers use different default strategies to handle faulty forms.

- *Morče*: includes morphological analyzer, lexically-driven
- *TnT*: more sensitive to syntactic context
- Both include a method to handle unknown words.

Do they have something interesting to say about incorrect forms?

(7) Tady je vecne dobra programa navstevy.
here is always? good programme of the visit
'This place is always worth visiting.'

— emendable as:

(8) Tady je vždy dobrý program návštěvy.

What the taggers say about **programa**:

- *Morče*: **genitive** masculine singular, lemma *programus*
  – morphology-based interpretation
- *TnT*: **nominative** neuter singular
  – syntax-based interpretation

– unfortunately, not enough nice results like this in our data

Comparison of taggers (Morče vs. TnT)

The sample:

- no. of texts: 93
- no. of tokens: 12681
- no. of words (excluding punctuation): 10727

Comparison of tagger results on ill-formed words:

- ill-formed tokens (= unidentified and guessed by Morče): 1323 (8.9%)
- ill-formed tokens where taggers agree: 405 (28.8%)
- ill-formed tokens where taggers disagree: 918 (71.2%)

Evaluation of tagger results on L0 vs. L1:
(next slide)

Tags on L0 and L1 – percentages of agreement

	L0m x L0t	L0m x L1	L0t x L1	L0m x L1	L0t x L1
# tokens	918	2589	2589	314	314
Tag	0	84.1	79.0	19.1	26.1
POS	39.2	89.6	88.7	43.9	52.5
SubPOS	37.1	89.2	87.9	42.0	49.7
Gender	23.9	88.8	88.2	36.0	46.5
Number	36.9	91.1	91.2	49.0	63.1
Case	31.2	89.0	86.5	43.0	51.3
PossGen	98.6	99.8	99.9	98.4	99.7
PossNr	99.5	99.8	99.7	99.0	99.7
Person	68.1	96.3	94.2	81.8	76.1
Tense	70.6	96.7	95.3	83.1	77.4
Grade	78.3	96.4	96.9	75.2	81.5
Negation	74.4	95.3	93.8	73.9	74.2
Voice	70.6	96.7	95.5	83.1	78.7

Numbers of tags assigned to ill-formed words

POS	Morče	Tnt
adjective	158	94
adverb	118	21
gradable adverb	31	11
noun	499	441
preposition	10	-
particle	8	-
finite verb	32	129
infinitive	7	41
l-participle	10	119
passive participle	1	29

Morče vs. TnT

- Morphological / syntactic interpretation of faulty forms: unconfirmed, more experiments needed
- *TnT* loses ground in a context with many errors
- *Morče* strongly disprefers verbs
- *TnT* better on faulty forms, *Morče* better in general

Spell-checker I

- michalisekSpell, by Michal Richter (2010)
- combines morphology with context
- Modes: spell-checker, proof-reader, diacritics assigner
- The sample
  - no. of texts: 67
  - no. of tokens: 9373
  - no. of words (excluding punctuation): 7995
- Evaluated on:
  - identical emendations on L1: 9069 tokens (96.8%)
  - identical emendations on L2: 8549 tokens (91.2%)
- Ill-formed tokens:
  - ill-formed total (= unidentified and guessed by morce): 918
  - ill-formed with identical emendations on L1: 786
- Results for ill-formed tokens with identical emendations on L1:

Spell-checker II

- where diacritics assigner agrees with L1: 552 (70.2%)
- where proof-reader agrees with L1: 639 (80.5%)
- where diacritics assigner followed by proof reader agrees with L1: 644 (81.9%)

Outline of the talk

Conclusion

- Morphosyntactic errors are easy to formalize and lead to a high Kappa – incor, agr, dep
- Semantic errors depend on subjective judgement, should standard measures be applied?
- Projecting morphosyntactic annotation of L1 and L2 onto L0 straightforward and useful
- Extracting useful information from multiple taggers applied to L0 not proved viable so far
- Proof-reader has a relatively high degree of success, could be a part of a fully automatic chain, with a tagger as the next step

Acknowledgments

Thanks to

other members of the project team, namely Karel Šebesta, Milena Hnátková, Tomáš Jelínek, Vladimír Petkevič, and Hana Skoumalová