A machine translation system into a minority language

Petr Homola & Vladislav Kuboň
RANLP MT workshop 2005, Borovets
September 24, 2005

Overview

- Lower Sorbian — basic facts
- similarity of languages
- MT — a shallow approach
- system architecture
- final remarks & future work

Lower Sorbian

- spoken in Lower Lusatia
- West Slavonic
- centre: Chóśebuz/Cottbus
- ~10,000 speaker

Lower Sorbian (2)

- rich inflection
- free word order (unmarked: SOV)
- archaic features
  - dual, aorist, imperfect, supine
- influenced by German

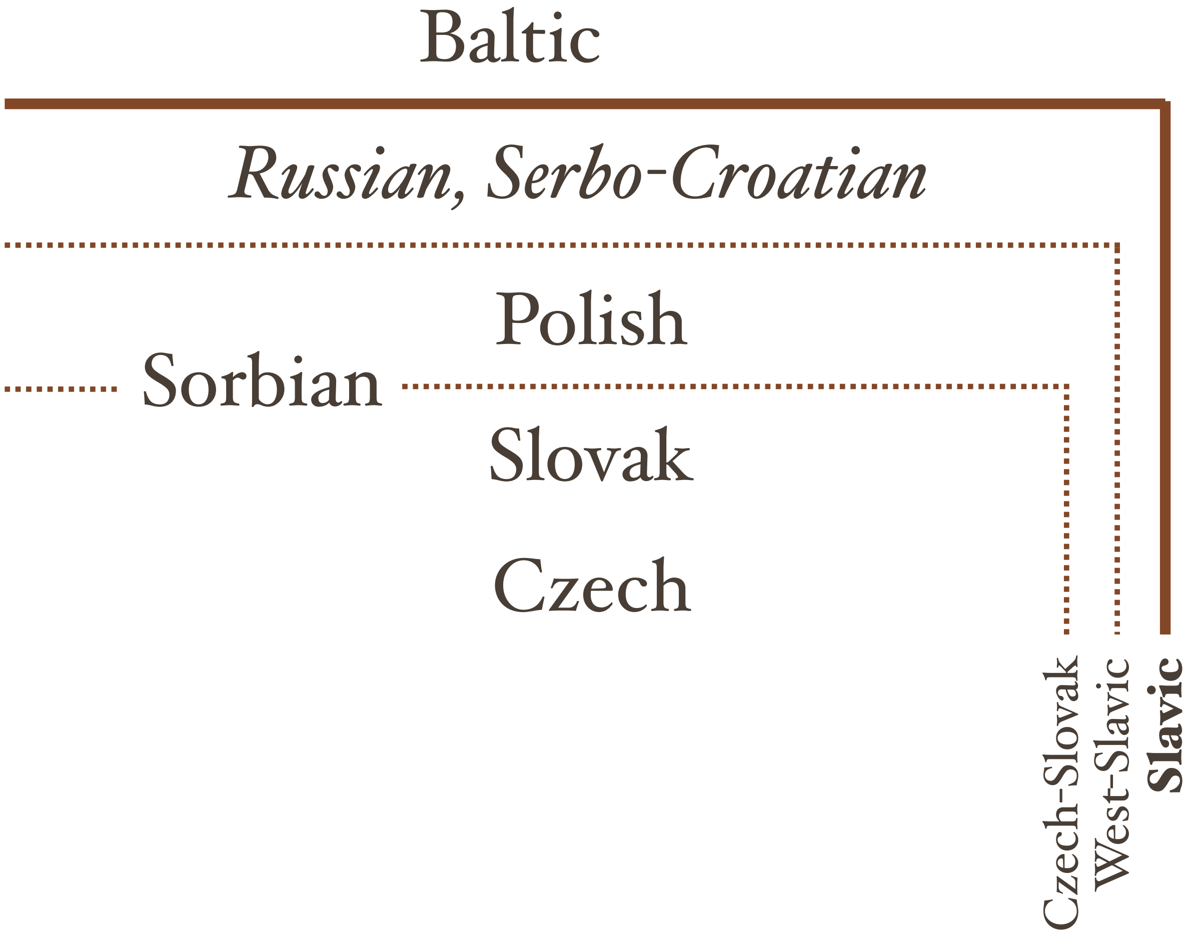
Česílko – motivation

- assumption: MT among closely related languages doesn’t require full syntactic analysis and transfer
- advantages: shallow MT is more robust and simpler to implement
- basic question: how ‘deep’ do we have to analyze sentences?

Levels of language similarity

- closely related languages (e.g., Czech/Slovak, Upper/Lower Sorbian)
- related languages
  - one family (e.g., Slovak/Polish/Russian)
  - across families (e.g., Polish/Lithuanian)
- other cases (e.g., English/Hungarian)

Češílko: languages



Baltic

Russian, Serbo-Croatian

Polish

Sorbian

Slovak

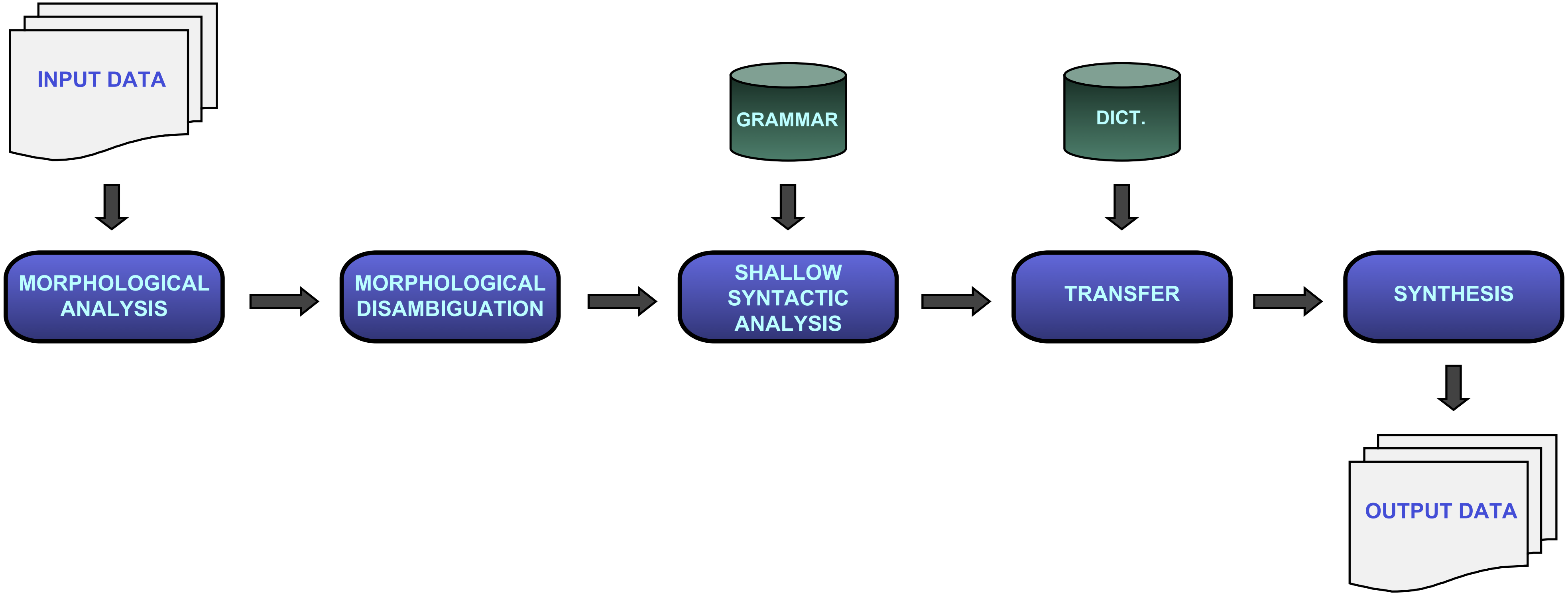
Czech

Czech-Slovak

West-Slavic

Slavic

System architecture



Morphological processing

- analysis
  - over 800.000 lemmas (20 mio inflected word forms)
  - 15 positional tags
- disambiguation
  - stochastical, trained on the Prague Dependency Treebank
  - accuracy 95%

Dictionaries

- domain-related
  - individual words, multiple-word terms
  - organized hierarchically (most specific first)
- general
- translating: lemmas, tagsets

Partial syntactic analysis

- rule based
- analyzing simple constituents (e.g., NP, PP)
- partial (e.g., no embedded sentences in NPs)
- implementation:
  - chunk parser & feature structures
  - similar to LFG

Partial SA (2)

- context-free rules
  - result: c-structure (phrase structure tree)
  - e.g., NP → A N
- constraints (equations for unification)
  - result: f-structure (feature structure)

Transfer: morphology

- different morphological features
- example: *jazyk* “language”
    - Czech: gender=*masc*
    - Sorbian: lemma=*rěc*, gender=*fem*

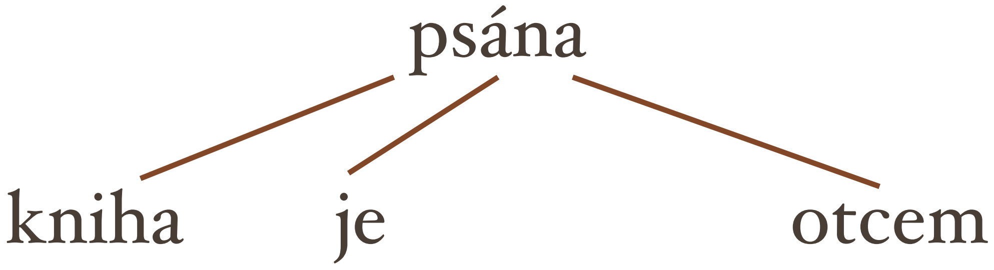
Transfer: problems

- for example:
  - agreement (e.g., *srbský*$_{masc}$ *jazyk* → *serbska*$_{fem}$ *rěc* “Sorbian language”)
  - structural difference (e.g., *kniha*$_{sg}$ *je*$_{aux,3sg}$ *psána*$_{pass.part}$ *otcem*$_{ins}$ → *knigły*$_{pl}$ *se*$_{refl}$ *pišu*$_{3pl}$ *wót nana*$_{gen}$ “a/the book is being written by the father”)

Shallow SA: example

kniha je psána otcem
“a/the book is being written by the father”

syntactic tree



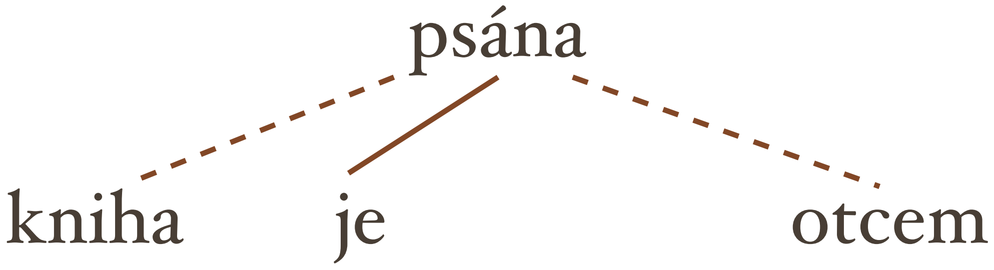
psána

kniha je otcem

Shallow SA: example (2)

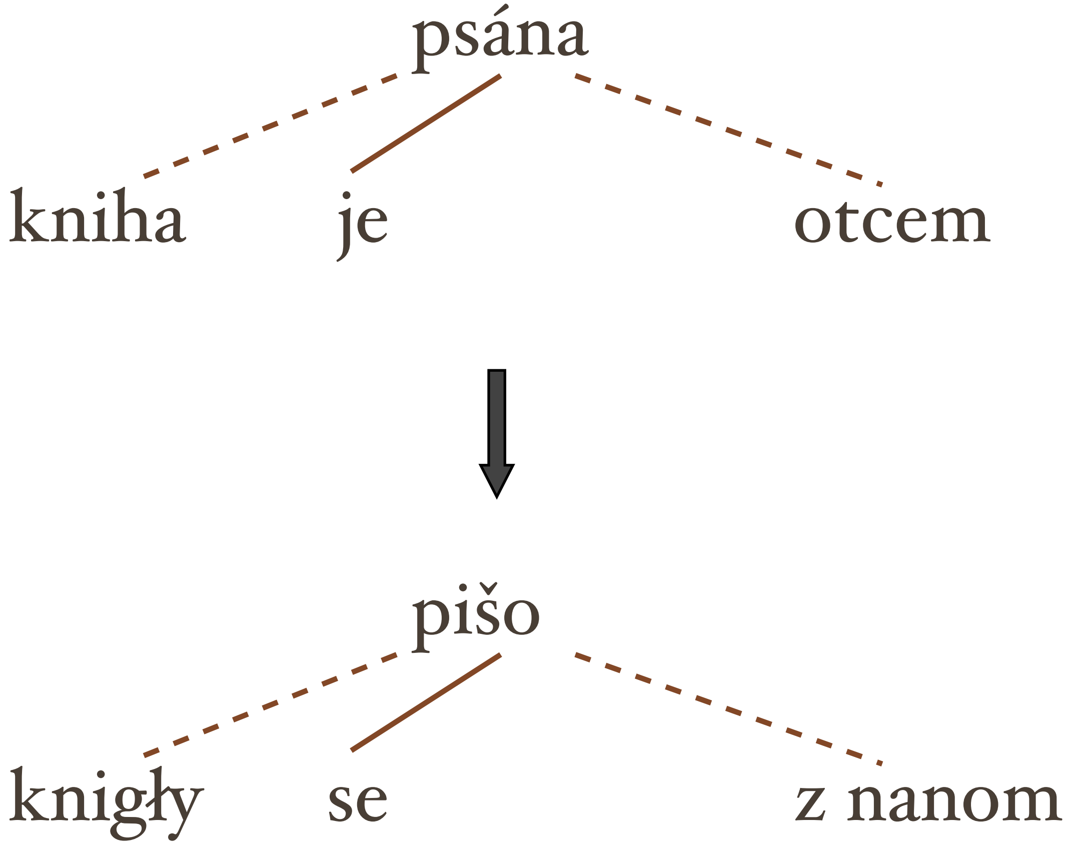
kniha je psána otcem
"a/the book is being written by the father"

partial syntactic trees



Transfer: example

shallow



psána

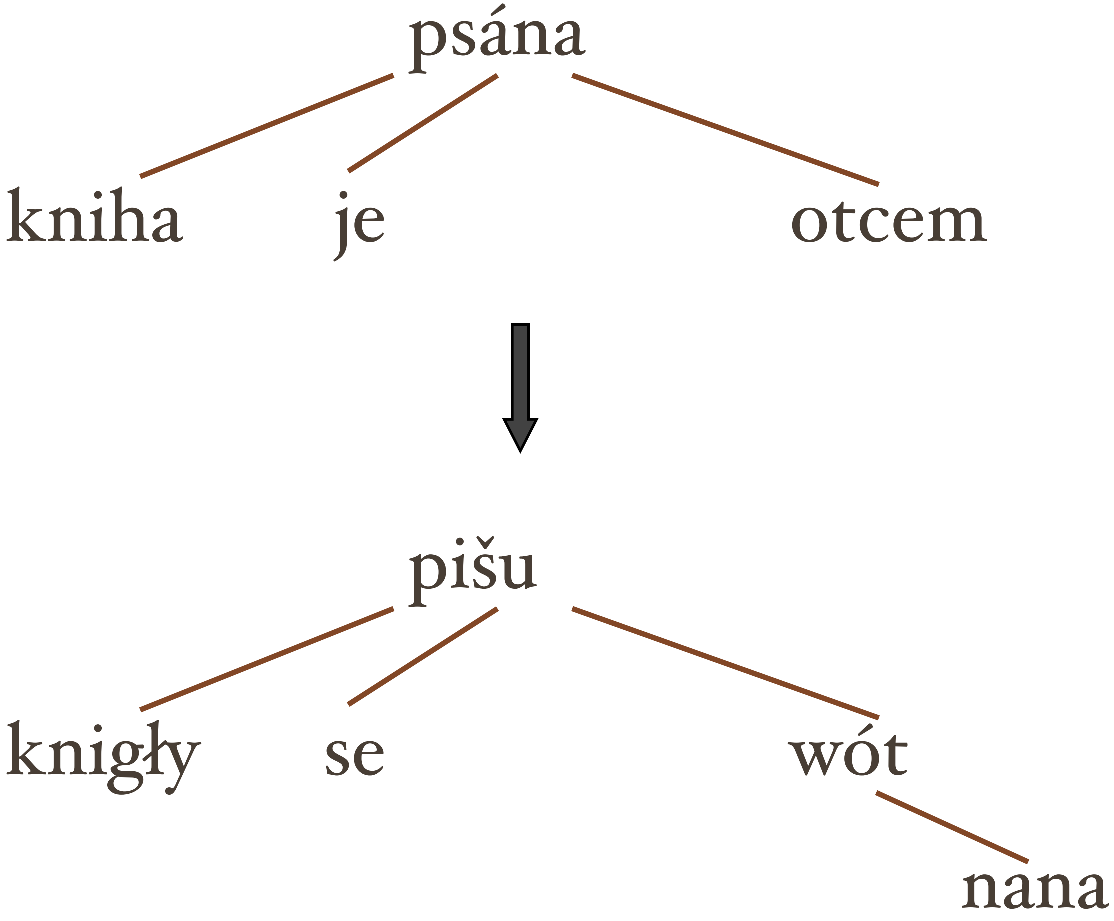
kniha je otcem

pišo

knigły se z nanom

Transfer: example (2)

deep



psána

kniha

je

otcem

pišu

knigły

se

wót

nana

Example: input FS

- #input[POS: ‘subst’, CASE: #nom[], ORDER: ‘1’, CAPITAL: ‘1’, FORM: ‘Praha’, NEG: ‘pos’, GENDER: #fem[], LEMMA: ‘Praha’, ORIGTAG: ‘NNFS1-----A----’, NUMBER: #sg[]]

Example: rule

- {SPAN: 3,
  COND: ‘fs1#subtype(”word”) & fs2#type() = “filler” &
  fs3#subtype(”word”)
  & fs1#featureValue(”POS”) = “adj” &
  fs3#featureValue(”POS”) = “subst”
  & setFsVar(”GENDER”, fs1#unifyAtt(fs3, “GENDER”)) !
  null & setFsVar(”CASE”, fs1#unifyAtt(fs3, “CASE”)) ! null &
  setFsVar(”NUMBER”, fs1#unifyAtt(fs3, “NUMBER”)) ! null’,
  NEW: ‘clone(fs3)#setFsAtt(”ADJ”,
  fs1)#setTextAtt(”PHRASE”, fs1#featureValue(”PHRASE”) +
  “ “ +
  fs3#featureValue(”PHRASE”))#replaceFsAtt(”GENDER”,
  getFsVar(”GENDER”))#replaceFsAtt(”CASE”,
  getFsVar(”CASE”))#replaceFsAtt(”NUMBER”,
  getFsVar(”NUMBER”))’,
  LOG: “‘R1’”}

Evaluation

- tool: Trados Translator's Workbench
- translated text corrected manually to ensure grammaticality
- average accuracy ~ weighted average of accuracy over all sentences
  - weight: the length of the sentence (number of words)

Evaluation (2)

from Czech into Lower Sorbian

	tagger	*manual*
no parser	92%	93%
shallow	93%	95%

Evaluation (3)

source language: Czech

target language	*w eighted avg.*	*synt. analysis*
Slovak	90%	none
Polish	71.4%	none
Lithuanian	87.6%	shallow
Lower Sorbian	93%	shallow

Limits

- only local dependencies
- no non-projective structures
- valence (verbs, adjectives…)
- information structure (topic/focus articulation)

In progress: deep analysis

- taking verbal valence into account
- goal: recognize all projective dependencies
- comparison with shallow approach
  - shallow: ~40% sentences translated correctly
  - deep: lower variance of ill-formed sentences

Final remarks

- shallow MT is sufficient to produce raw translation between related languages
- saves work of human translators
- comparatively easy to implement
  - only 10 syntactic rules (~40 for deep analysis so far)

Future work

- many errors caused by the tagger
  - try to use non-disambigusted output
- other problem: semantic ambiguity
  - e.g., Sorbian: *dajo*
    - 1. “to give”
    - 2. “there is…”

Thank you