Czechizator – Čechizátor
Czechizator

- lexicon-less “translation” from English to Czech
- lexicon-less “translation” from English to Czech
- usual approach: use a bilingual lexicon
Czechizator

- lexicon-less “translation” from English to Czech
- usual approach: use a bilingual lexicon

- Czechizator approach: use a set of rules instead

rules:
- -ise → -iza
- -tion → -ce

...
Example: Czechizating ITAT titles

- Statistical modelling in climate science
Example: Czechizing ITAT titles

- Statistical modelling in climate science
- Statistické modelování v klimat scienci
Example: Czechizing ITAT titles

- Statistical modelling in climate science
  Statistické modelování v klimat scienci

- 12 years of Unsupervised Dependency Parsing
Example: Czechizing ITAT titles

- Statistical modelling in climate science
  Statistické modelování v klimat scienci

- 12 years of Unsupervised Dependency Parsing
  12 jírů nesupervizované parsování dependence
Example: Czechizing ITAT titles

- Statistical modelling in climate science
  Statistické modelování v klimat scienci

- 12 years of Unsupervised Dependency Parsing
  12 jírů nesupervizované parsování dependence

- Multivariable Approximation by Convolutional Kernel Networks
Example: Czechizating ITAT titles

- Statistical modelling in climate science
  Statistické modelování v klimat scienci

- 12 years of Unsupervised Dependency Parsing
  12 jírů nesupervizované parsování dependence

- Multivariable Approximation by Convolutional Kernel Networks
  Multivariabilní aproximace Konvolucional Kernel networksu
Implementation

- lexical translation: a set of Czechization rules
  - 43 ending-based transformation rules (see later)
  - 33 transliteration rules: th → t, ti → ci, ck → k, ph → f, sh → š, igh → aj, dg → dž, w → v, c → k…
  - 36 hard-coded translations of semi-auxiliaries: be, have, do, and, or, all, this, many, only, main…

- grammar and function words: TectoMT
  - English-Czech machine translation system
  - Czechizator implemented as a TectoMT lexical translation model
I preferred the presentation of David.
I preferred the presentation of David.

TectoMT analysis
Implementation

I preferred the presentation of David.

TectoMT analysis

prefer verb, 1st person, past

presentation noun, definite, object

David noun+of, named ent.
Implementation

I preferred the presentation of David.

TectoMT analysis

prefer verb, 1\textsuperscript{st} person, past

presentation noun, definite, object

David noun+of, named ent.

transfer
Rudolf Rosa: Czechizator - Čechizátor

Implementation

I preferred the presentation of David.

TectoMT analysis

prefer
verb, 1st person, past

presentation
noun, definite, object

David
noun+of, named ent.

Czechization of lemmas
Implementation

I preferred the presentation of David.

TectoMT analysis

prefer verb, 1st person, past

presentation noun, definite, object

David noun+of, named ent.

Czechization of lemmas

preferovat

prezentace

David
I preferred the presentation of David.

TectoMT analysis

prefer
verb, 1st person, past

presentation
noun, definite, object

David
noun+of, named ent.

Czechization of lemmas

preferovat

prezentace

TectoMT transfer of attributes

Rudolf Rosa: Czechizator - Čechizátor
I preferred the presentation of David.

TectoMT analysis

prefer
verb, 1st person, past

presentation
noun, definite, object

David
noun+of, named ent.

Czechization of lemmas

preferovat
verb, 1st person, past

prezentace
noun, accusative

David
noun, genitive, n.e.

TectoMT transfer of attributes
Implementation

I preferred the presentation of David.

TectoMT analysis

prefer verb, 1st person, past

presentation noun, definite, object

David noun+of, named ent.

Czechization of lemmas

preferovat verb, 1st person, past

prezentace noun, accusative

David noun, genitive, n.e.

TectoMT transfer of attributes

TectoMT synthesis
Implementation

I preferred the presentation of David.

TectoMT analysis

prefer
verb, 1st person, past

presentation
noun, definite, object

David
noun+of, named ent.

Czechization of lemmas

preferovat
verb, 1st person, past

prezentace
noun, accusative

David
noun, genitive, n.e.

TectoMT transfer of attributes
Transformation rules for adjectives

- **partial**  →  **parciální**
- **stable**  →  **stabilní**
- **tolerant**  →  **tolerantní**
- **tolerated**  →  **tolerovaný**
- **turkic**  →  **turkický**
- **practical**  →  **praktický**
- **native**  →  **nativní**
- **regular**  →  **regulární**
- **fatal**  →  **fatální**
- **nervous**  →  **nervózní**
- **parsed**  →  **parsovaný**
- **parsing**  →  **parsující**
- **park**  →  **parkový**
What is it good for?

- translations sometimes “reasonable”
  - scientific titles and abstracts, marketing texts
What is it good for?

- translations sometimes “reasonable”
  - scientific titles and abstracts, marketing texts:
  - Accenture Operations combines technology that digitizes and automates business processes, unlocks actionable insights, and delivers everything-as-a-service with our team's deep industry, functional and technical expertise.
  - Operacions acenturu kombinuje technologii, která digitizuje a automuje procesy businosti, unlokuje akcionabilní insajty a deliveruje everything-as-a-service s funkcionální a technickou expertizou dípové industrie našeho tímu.
What is it good for?

- translations sometimes “reasonable”
  - scientific titles and abstracts, marketing texts
- still, only a proof of concept & a fun application
  - not really useful as a standalone tool
  - *maybe* as a starting point for later post-editing
What is it good for?

- translations sometimes “reasonable”
  - scientific titles and abstracts, marketing texts
- still, only a proof of concept & a fun application
  - not really useful as a standalone tool
  - *maybe* as a starting point for later post-editing
- potential: combine with TectoMT lexical models
  - frequent words: translation model trained from data
  - infrequent words: insufficient training data, Czechize!
Complementing TectoMT

- rare/unseen words not well handled by TectoMT
  - unreliable translation for rare words, none for unseen
- e.g. scientific terms
  - large number and growing, rare in data
  - often rather regular translations → can be Czechized
- anaphora → anafora
- hypotactical → hypotaktický
- circumfixal → cirkumfixální
Complementing TectoMT

- rare/unseen words not well handled by TectoMT
  - unreliable translation for rare words, none for unseen
- e.g. scientific terms
  - large number and growing, rare in data
  - often rather regular translations → can be Czechized
    - anaphora → anafora
    - hypotactical → hypotaktický
    - circumfixal → cirkumfixální
- current issues: named entities get Czechized
  - usually should be avoided, but detection insufficient
Conclusion

- lexicon-less lexical “translation” module
  - transformation (endings) and transliteration rules
- grammar and aux words handled by TectoMT
  - Czechization of lemmas on t-layer
- Czechization of scientific titles sometimes “good”
  - but still not really useful
- work in progress: integrate into TectoMT
  - complement existing lexical models
  - Czechize rare and unseen words, e.g. science terms
Thank you for your attention

Rudolf Rosa
rosa@ufal.mff.cuni.cz

Czechizator – Čechizátor

http://ufal.mff.cuni.cz/czechizator

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

http://ufal.mff.cuni.cz/rudolf-rosa/
Examples: parsing papers

- The theory of parsing, translation, and compiling
- Accurate unlexicalized parsing
- An efficient context-free parsing algorithm
- Seven principles of surface structure parsing in natural language
- Head-driven statistical models for natural language parsing
- Parsing by chunks
- Shallow parsing with conditional random fields
Examples: parsing papers

- Teorie parsování, translace a kompiluje
- Akuratová unlexikalizovaná parsování
- Eficientová kontext-fríová parsování algoritmus
- 7 principů struktur surface parsování v naturální langvaži
- Híd-drivenové statistické modely pro naturální langvaž parsování
- Parsují Chunky
- Šalovujte, parsujete s kondicionálními randomovými Fieldy