Global Variants in the Czech Language

Jaroslava Hlaváčová, Lukáš Kyjánek, and Magda Ševčíková

September 23-27, 2022
Zuberec, Slovakia

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

Introduction

- Some Czech words can be written in several slightly different ways

Introduction

- Some words can be written in several slightly different ways
  - **inflectional variants = an *n*-tuple of wordforms belonging to the same inflectional paradigm, but different spelling;**
    - ***obchodu~obchodě*** (locative case)

Introduction

- Some words can be written in several slightly different ways
  - **inflectional variants = an *n*-tuple of wordforms belonging to the same inflectional paradigm, but different spelling;**
    - ***obchodu~obchodě*** (locative case)
  - **global variants = an *n*-tuple of lemmas whose difference in spellings propagate to all wordforms and most of their derivationally related words;**
    - ***obchod~vobchod → obchodní~vobchodní***

Introduction

- Some words can be written in several slightly different ways
  - **inflectional variants = an *n*-tuple of wordforms belonging to the same inflectional paradigm, but different spelling;**
    - ***obchodu~obchodě*** (locative case)
  - **global variants = an *n*-tuple of lemmas whose difference in spellings propagate to all wordforms and most of their derivationally related words;**
    - ***obchod~vobchod → obchodní~vobchodní***
- Content
  - How have we identified global variants?
  - What types of global variants have we seen?

Available resources that contain global variants

- MorfFlex 2.0
  - lemma: ***voprášit***_,h_^(^GC****oprášit***) ← Common/Non-standard Czech
  - lemma: ***lavor***_,s_^(^DD****lavór***) ← Standard Czech
  - lemma: ***Dominigue***_,i_^(^DS****Dominique***) ← Distortion/Typo

Available resources that contain global variants

- MorfFlex 2.0
  - lemma: ***voprášit***_,h_^(^GC****oprášit**) ← Common/Non-standard Czech
  - lemma: ***lavor***_,s_^(^DD****lavór**) ← Standard Czech
  - lemma: ***Dominigue***_,i_^(^DS****Dominique**) ← Distortion/Typo
- VALLEX 3.0
  - ***chytit*/*chytnout***^{pf}, ***dozvídat se*/*dovídat se***^{impf}, ***dozvědět se*/*dovědět se***^{pf}

Available resources that contain global variants

- MorfFlex 2.0
  - lemma: ***voprášit***_,h_^(^GC***oprášit***) ← Common/Non-standard Czech
  - lemma: ***lavor***_,s_^(^DD***lavór***) ← Standard Czech
  - lemma: ***Dominigue***_,i_^(^DS***Dominique***) ← Distortion/Typo
- VALLEX 3.0
  - ***chytit*/*chytnout***^{pf}, ***dozvídat se*/*dovídat se***^{impf}, ***dozvědět se*/*dovědět se***^{pf}
- Slovník spisovného jazyka českého (SSJČ)
  - v. = viz (see) in "***obepsati*** v. ***opsati***"
  - comma in "***mysliti***, ***mysleti***"
  - řidč. = řidčeji (rarely) in "***zpěvánka***, řidč. ***zpěvanka***, ***zpívánka***"

Available resources that contain global variants

- MorfFlex 2.0
  - lemma: ***voprášit***_,h_^(^GC*******oprášit***) ← Common/Non-standard Czech
  - lemma: ***lavor***_,s_^(^DD*******lavór***) ← Standard Czech
  - lemma: ***Dominigue***_,i_^(^DS*******Dominique***) ← Distortion/Typo
- VALLEX 3.0
  - ***chytit*/*chytnout***^{pf}, ***dozvídat se*/*dovídat se***^{impf}, ***dozvědět se*/*dovědět se***^{pf}
- Slovník spisovného jazyka českého (SSJČ)
  - v. = viz (see) in "***obepsati*** v. ***opsati***"
  - comma in "***mysliti***, ***mysleti***"
  - řidč. = řidčeji (rarely) in "***zpěvánka***, řidč. ***zpěvanka***, ***zpívánka***"
- long tradition of linguistic studies on the topic of spelling variants

Searching for new global variants

1. extracting variants from the existing resources

Searching for new global variants

1. extracting variants from the existing resources

2. formalising regular patterns
    - automatic *vs.* manual
    - more than one hundred patterns in a form of regular expressions
      e.g., ^o.* ↔ ^vo.* in ***obchodovat~vobchodovat***
    - patterns took into account also morpho-syntactic categories
      e.g., masc.anim ***car*** (*tsar*) *vs.* masc.inanim ***cár*** (*shred*)

Searching for new global variants

1. extracting variants from the existing resources

2. formalising regular patterns
   - automatic *vs.* manual
   - more than one hundred patterns in a form of regular expressions
     e.g., ^o.* ↔ ^vo.* in ***obchodovat~vobchodovat***
   - patterns took into account also morpho-syntactic categories
     e.g., masc.anim ***car*** (*tsar*) *vs.* masc.inanim ***cár*** (*shred*)

3. applying patterns to MorfFlex
   - manual annotation was done because of
     e.g., ***fiala*** (*wallflower*) ⋆ ***fiála*** (*pinnacle*) but ***neandrtalec~neandrtálec*** (*Neanderthal*)

Searching for new global variants

1. extracting variants from the existing resources

2. formalising regular patterns
   - automatic *vs.* manual
   - more than one hundred patterns in a form of regular expressions
     e.g., ^o.* ↔ ^vo.* in ***obchodovat~vobchodovat***
   - patterns took into account also morpho-syntactic categories
     e.g., masc.anim ***car*** (*tsar*) *vs.* masc.inanim ***cár*** (*shred*)

3. applying patterns to MorfFlex
   - manual annotation was done because of
     e.g., ***fiala*** (*wallflower*) ⋆ ***fiála*** (*pinnacle*) but ***neandrtalec~neandrtálec*** (*Neanderthal*)

4. uploading variants to MorfFlex and DeriNet

Global variants into MorfFlex CZ

- 33,477 *n*-tuples
  were annotated in the MorfFlex 2.0
- **18,167 n-tuples**
  were newly added to future ver.
- the main increment is recorded
  for smaller *n*

n	MorfFlex 2.0	after
2	31,919	49,079
3	1,227	2,089
4	121	264
5	16	18
6	187	187
8	4	4
9	1	1
11	1	1
12	1	1

Global variants into DeriNet

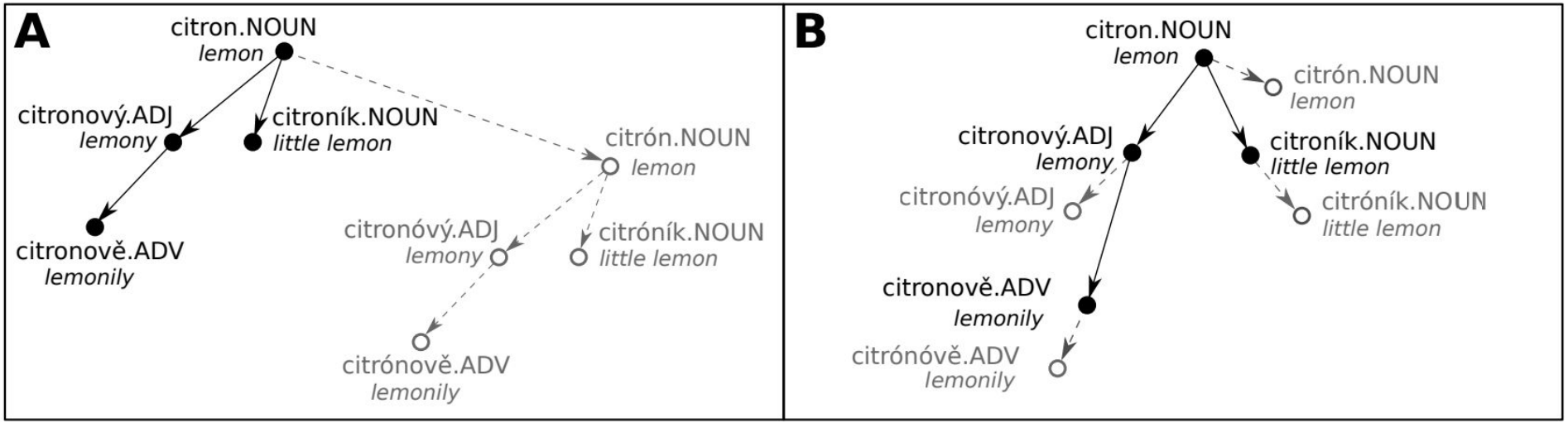

Figure 1: Two possible ways of representing global variants in the rooted trees; (A) making parallel branches, (B) connecting variants to the basic variant (the latter option implemented in DeriNet 2.1).

- in A, a missing word/variant, e.g., ***citrón***, would disconnect the branch
- in B, there is a need for a representative **basic variant** for each *n*-tuple

Example of global variants from DeriNet 2.1

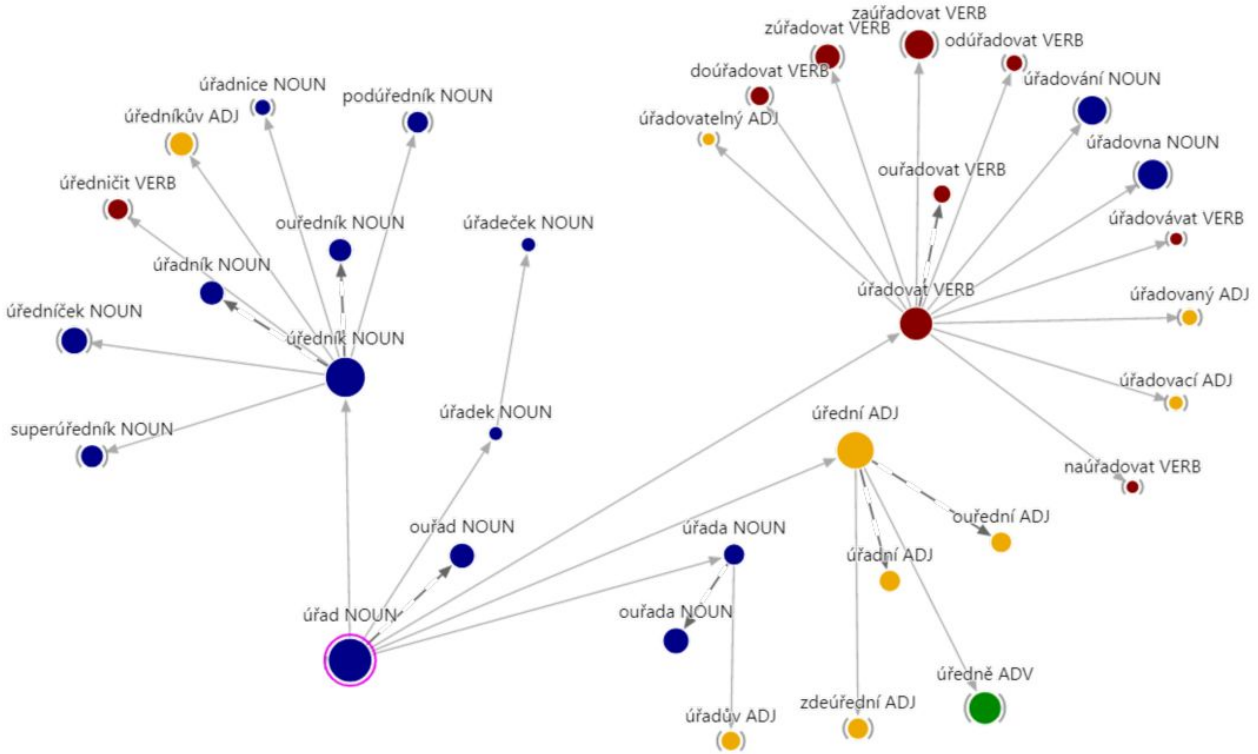

Figure 2: Simplified record of the global variants of the noun *úřad* ~ *ouřad* (*office*) and its derivatives from DeriNet 2.1. Variant relations are represented by dark grey dashed arrows that are shorter than the light grey solid arrows, which represent derivational relations. Size of the nodes corresponds to the token frequency of the lemmas in the corpus SYNv4 [8]. Brackets around nodes indicate that the node's derivatives were hidden for spatial reasons.

Prototypical Cases of Global Variants

Prototypical Cases

- Long and Short vowels
  - ***svíčkař ~ svíčkář*** *(who makes candles)*
  - ***kvikat ~ kvíkat*** *(to oink/squeak)*

Prototypical Cases

- Long and Short vowels
  - ***svíčkař ~ svíčkář*** *(who makes candles)*
  - ***kvikat ~ kvíkat*** *(to oink/squeak)*
- Alveolar vs. Postalveolar/Palatal Consonants
  - ***vlaštovka ~ vlašťovka*** *(a swallow)*
  - ***student ~ študent*** *(student)*
  - ***mrazený ~ mražený*** *(frozen)*

Prototypical Cases

- Long and Short vowels
  - ***svíčkař ~ svíčkář*** *(who makes candles)*
  - ***kvikat ~ kvíkat*** *(to oink/squeak)*
- Alveolar vs. Postalveolar/Palatal Consonants
  - ***vlaštovka ~ vlašťovka*** *(a swallow)*
  - ***student ~ študent*** *(student)*
  - ***mrazený ~ mražený*** *(frozen)*
- Soft and Hard Adjectives
  - ***námezdný ~ námezdní*** *(hired)*
  - ***přívodný ~ přívodní*** *(feed, inflow - e.g. pipe)*

Prototypical Cases

- Long and Short vowels
  - ***svíčkař ~ svíčkář*** *(who makes candles)*
  - ***kvikat ~ kvíkat*** *(to oink/squeak)*
- Alveolar vs. Postalveolar/Palatal Consonants
  - ***vlaštovka ~ vlašťovka*** *(a swallow)*
  - ***student ~ študent*** *(student)*
  - ***mrazený ~ mražený*** *(frozen)*
- Soft and Hard Adjectives
  - ***námezdný ~ námezdní*** *(hired)*
  - ***přívodný ~ přívodní*** *(feed, inflow - e.g. pipe)*
- Prothetic *v-*
  - ***okno ~ vokno*** *(window)*
  - ***zotvírat ~ zvotvírat*** *(to open step by step)*

Prototypical Cases

- Long and Short vowels
  - ***svíčkař ~ svíčkář*** *(who makes candles)*
  - ***kvikat ~ kvíkat*** *(to oink/squeak)*
- Alveolar vs. Postalveolar/Palatal Consonants
  - ***vlaštovka ~ vlašťovka*** *(a swallow)*
  - ***student ~ študent*** *(student)*
  - ***mrazený ~ mražený*** *(frozen)*
- Soft and Hard Adjectives
  - ***námezdný ~ námezdní*** *(hired)*
  - ***přívodný ~ přívodní*** *(feed, inflow - e.g. pipe)*
- Prothetic *v-*
  - ***okno ~ vokno*** *(window)*
  - ***zotvírat ~ zvotvírat*** *(to open step by step)*
- Stylistics (*ú ~ ou, ý ~ ej, th ~ t, s ~ z*)
  - ***mechanismus ~ mechanizmus*** *(mechanism),*
  - ***vytékat ~ vytejkat*** *(flow/leak out)*
  - ***úzký ~ ouzký*** *(narrow),*
  - ***ortopedie ~ orthopedie*** *(orthopedics)*

Prototypical Cases

- Long and Short vowels
  - ***svíčkař ~ svíčkář*** *(who makes candles)*
  - ***kvikat ~ kvíkat*** *(to oink/squeak)*
- Alveolar vs. Postalveolar/Palatal Consonants
  - ***vlaštovka ~ vlašťovka*** *(a swallow)*
  - ***student ~ študent*** *(student)*
  - ***mrazený ~ mražený*** *(frozen)*
- Soft and Hard Adjectives
  - ***námezdný ~ námezdní*** *(hired)*
  - ***přívodný ~ přívodní*** *(feed, inflow - e.g. pipe)*
- Prothetic *v-*
  - ***okno ~ vokno*** *(window)*
  - ***zotvírat ~ zvotvírat*** *(to open step by step)*
- Stylistics (*ú ~ ou, ý ~ ej, th ~ t, s ~ z*)
  - ***mechanismus ~ mechanizmus*** *(mechanism),*
  - ***vytékat ~ vytejkat*** *(flow/leak out)*
  - ***úzký ~ ouzký*** *(narrow),*
  - ***ortopedie ~ orthopedie*** *(orthopedics)*
- Vocalized and Non-vocalised Prefixes
  - ***vpisovat ~ vepisovat*** *(inscribe)*
  - ***střást ~ setřást*** *(shake off)*
  - ***objet ~ obejet*** *(go around)*
  - ***předběhnout ~ předeběhnout*** *(overtake)*
  - ***předepisovat ~ předpisovat*** *(prescribe)*

Foreign Names

- Czech translations of geographic names - usually not variants
  - ***Paris → Paříž, Moscow → Moskva***
    - originals uninflected

Foreign Names

- Czech translations of geographic names - usually not variants
  - ***Paris ≠ Paříž, Moscow ≠ Moskva***
    - originals uninflected
- Person names (in Latin script) - usually global variants
    - not translated, but inflected (***Shakespeare***)
    - including typical typos, non-standard orthography (nespisovné)
  - ***Abdulah ~ Abdullah ~ Abduláh***

Foreign Names

- Czech translations of geographic names - usually not variants
  - ***Paris ⌖ Paříž, Moscow ⌖ Moskva***
    - originals uninflected
- Person names (in Latin script) - usually global variants
    - not translated, but inflected (***Shakespeare***)
    - including typical typos, non-standard orthography (nespisovné)
  - ***Abdulah ~ Abdullah ~ Abduláh***
- Exceptions:
  - Slavic names with ending *-ski* or *-skij* - inflectional variants
    - ***Čajkovský ~ Čajkovskij ~ Čajkovski*** … only for singular nominative and vocative
  - Ancient Greek names with endings *-es* or *-és*
    - ***Empedokles ~ Empedoklés***

Nouns with more paradigms

- Usually not global variants
- one wordform for lemma: ***kužel*** *(cone)*, ***chmel*** *(hop)*, ***korbel*** *(mug)*
  - combined inflectional paradigm (hrad + stroj) merged, wordforms = inflectional var.

Nouns with more paradigms

- Usually not global variants
- one wordform for lemma: ***kužel*** *(cone)*, ***chmel*** *(hop)*, ***korbel*** *(mug)*
  - combined inflectional paradigm (hrad + stroj) merged, wordforms = inflectional var.
- different wordforms of lemma: ***kapuca/kapuce*** *(hood);* ***brambor/brambora*** *(potato)*
  - should be merged, too, but due to different lemmas not yet
    - the same gender (*kapuca*) - no problem, lemmas might be inflectional variants
    - different genders (*brambor*) - **Principle of morphological differentiation**: the paradigm must have a single gender … PROBLEM

Nouns with more paradigms

- Usually not global variants
- one wordform for lemma: ***kužel*** *(cone)*, ***chmel*** *(hop)*, ***korbel*** *(mug)*
  - combined inflectional paradigm (hrad + stroj) merged, wordforms = inflectional var.
- different wordforms of lemma: ***kapuca/kapuce*** *(hood);* ***brambor/brambora*** *(potato)*
  - should be merged, too, but due to different lemmas not yet
    - the same gender (*kapuca*) - no problem, lemmas might be inflectional variants
    - different genders (*brambor*) - **Principle of morphological differentiation**: the paradigm must have a single gender … PROBLEM
- One wordform, different genders:
  - ***kredenc*** (masc. inan. / fem.)
  - ***tenor, hajzl*** (masc. inan. / masc. anim.)

Conclusion

- Inventory of Czech global variants
- Precise distinction between global and inflectional variants (variant lemma is not enough for global variant)
- New version of DeriNet
- MorfFlex - new links in future official edition (concept is already implemented in version 2.0 from 2020)
- Fuzzy border between variants and non-variants (*pécéčko*+*písíčko*, *zvýhodněný*+*zvýhodnělý*) … to be reconsidered

Questions?

Comments?

Suggestions?

Prototypical Cases … too detailed !!!

- **Vocalized and Non-vocalised Prefixes**

v-	*s-*	*vz-*	*roz-*	*od-*	*pod-*	*nad-*	*ob-*	*před-*
ve-	*se-*	*vze-*	*roze-*	*ode-*	*pode-*	*nade-*	*obe-*	*přede-*

  - *vpisovat ~ vepisovat (inscribe), vemlouvat ~ vmlouvat (ingratiate)*
  - *střást ~ setřást (shake off)*
  - *vzplát ~ vzeplát (flare up), vzedmout ~ vzdmout (surge)*
  - *rozsmutnit ~ rozesmutnit (make sad), rozebírat ~ rozbírat (disassemble)*
  - *odjet ~ odejet (leave), odečítat ~ odčítat (subtract)*
  - *podebírat ~ podbírat (scoop up), podjet ~ podejet (go under)*
  - *objet ~ obejet (go around), obestavět ~ obstavět (build around)*
  - *předběhnout ~ předeběhnout (overtake), předepisovat ~ předpisovat (prescribe)*

Acknowledgement

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, and the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), and LINDAT/CLARIAH-CZ project of the Ministry of Education (LM2015071, LM2018101).