# Corpus grammar – a chance for a descriptive approach

Václav Cvrček

Institute of the Czech National Corpus
Faculty of Arts, Charles University in Prague

4th February 2011

1. Description vs prescription
   - Case of GCC
   - Principles of noniterventional approach
2. Practice of GCC
   - Variants in GCC
   - Graphs
   - Page layout
   - Multi-word units
   - Closed paradigms
3. Tools
   - Paradigma
   - Slovotvorba
4. Corpus-based concepts
5. Corpus-driven concepts
6. Conclusion

# Description vs prescription

## Prescriptive situation

- monopoly codification
- discrepancy between the codification and the usage

# Description vs prescription

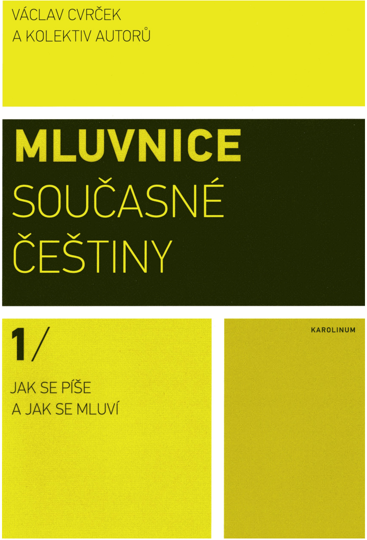
## Prescriptive situation

- monopoly codification
- discrepancy between the codification and the usage

## Descriptive situation

- plurality of codifications (with differences in attitudes and language norms)
- based on objective language data (without rejection of certain variants or varietites)

# Case of Grammar of Contemporary Czech



VÁCLAV CVRČEK
A KOLEKTIV AUTORŮ

**MLUVNICE**
SOUČASNÉ
ČEŠTINY

**1/**

JAK SE PÍŠE
A JAK SE MLUVÍ

KAROLINUM

# Concept of Minimal Intervention (CMI)

**Concept of Minimal Intervention** (Cvrček 2009) – theoretical background for GCC in the question of interventionalism.

# Concept of Minimal Intervention (CMI)

**Concept of Minimal Intervention** (Cvrček 2009) – theoretical background for GCC in the question of interventionalism.

## Premises:

1. There is no reason for linguists to infringe the language development by their interventions, and to disqualify thus speakers for their (natural) linguistic behavior, or purvey arguments for their disqualification.

# Concept of Minimal Intervention

## Premises (2):

2. The language has been evolving (by means of variations and oscillation between variants) into a sensible instrument of communication spontaneously and independently, needing no assistance from linguists.

# Concept of Minimal Intervention

## Premises (2):

2. The language has been evolving (by means of variations and oscillation between variants) into a sensible instrument of communication spontaneously and independently, needing no assistance from linguists.
3. The arbitrary nature of language means draws on their usage, and involves the ways of using the constituents (including their style characteristics and variety affiliation); it is thus beneficial for neither language development, nor its speakers when linguistics with its (institutionalized) interventions violates the very fact of this choice taken by majority.

# What does the CMI approach to language represent?

## Principles (1)

CMI is delimited by the endeavor to minimize linguists' interventional pressure on language and its speakers; the CMI's goal is to bring language situation as close as possible to the condition which is marked by the existence of spontaneously constituted order of lingual and communication norms speakers have appropriated when acquiring their mother tongue, and which is "only" passively recorded by linguists.

# What does the CMI approach to language represent?

## Principles (2)

Since the zero intervention is irreconcilable with the existence of linguistics as the science investigating language and presenting to the public the fruit of research, it is necessary to deliberately weaken potential linguistic interventions by the pluralism of descriptions (descriptive codifications) which should expressly declare the goals they pursue, what (communication) functions they favor; linguistic community should strive to create favorable conditions in order to achieve this goal.

# What does the CMI approach to language represent?

## Principles (3)

CMI as a construction of relation between linguistics, speakers and language does not address concrete properties of language, but the linguistic activity itself. CMI's measure of success is thus not the target condition of language. Sound application of minimal intervention is thus expressed by the stable competition of individual, functionally distinct codifications, which suggest dissimilar means, which are published at various time periods, have various recipients, and continuously track language development.

**Linguistic outputs, results of empirical and synchronic research (esp. those intended for general public) should be based solely on pure description, objective criteria, and representative quantum of relevant linguistic data, that is:**

1. Assessment language phenomena by objectively traceable measurable criteria, i.e. especially: frequency, spoken/written form, regionally-tinted (or nationwide).

Linguistic outputs, results of empirical and synchronic research (esp. those intended for general public) should be based solely on pure description, objective criteria, and representative quantum of relevant linguistic data, that is:

1. Assessment language phenomena by objectively traceable measurable criteria, i.e. especially: frequency, spoken/written form, regionally-tinted (or nationwide).
2. On the other hand, unacceptable are those assessments which are not positively deducible from language data or assume a priori knowledge. Like, for example, attitudes of speakers that often diverge from their actual speech behavior (those attitudes were acquired at school, complying thus with the predominant interventional practice), literariness vs. non-literariness of the language means, or formality vs. informality of the situation the constituent enters, etc.

...

3. Linguistic research should not limit itself to the prescribed language: codification then ceases to be descriptive and becomes contrastive (which is the inherent feature of contemporary prescriptivism). Any linguistic concept or report valid only within the limits of the codified language should be dismissed as incomplete.

...

3. Linguistic research should not limit itself to the prescribed language: codification then ceases to be descriptive and becomes contrastive (which is the inherent feature of contemporary prescriptivism). Any linguistic concept or report valid only within the limits of the codified language should be dismissed as incomplete.
4. Solely extensive and representative corpuses provide researchers with reliable linguistic data to satisfactory measure. Research based on insufficient collection of data should not be regarded as relevant. (It is important to find out clearly in what respect linguists can be their own informants, and in what respect they can not.)

# Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)

# Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)
- designed for students (not academic description)

# Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)
- designed for students (not academic description)
- two parts:

# Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)
- designed for students (not academic description)
- two parts:
  1. introduction to study of language, introduction to study of Czech (incl. history), **phonology**, lexicology, **word formation**, **morphology**, basic syntax, stylistic, orthography (writting system)

# Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)
- designed for students (not academic description)
- two parts:
  1. introduction to study of language, introduction to study of Czech (incl. history), **phonology**, lexicology, **word formation**, **morphology**, basic syntax, stylistic, orthography (writting system)
  2. syntax (expected 2011)

# Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)
- designed for students (not academic description)
- two parts:
  1. introduction to study of language, introduction to study of Czech (incl. history), **phonology**, lexicology, **word formation**, **morphology**, basic syntax, stylistic, orthography (writting system)
  2. syntax (expected 2011)
- corpus-based

## Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)
- designed for students (not academic description)
- two parts:
  1. introduction to study of language, introduction to study of Czech (incl. history), **phonology**, lexicology, **word formation**, **morphology**, basic syntax, stylistic, orthography (writting system)
  2. syntax (expected 2011)
- corpus-based
- descriptive (CMI-style)

# Grammar of Contemporary Czech – variants

All statistically significant variants in written and spoken language (corpora SYN2005 and Oral2006)

# Grammar of Contemporary Czech – variants

All statistically significant variants in written and spoken language (corpora SYN2005 and Oral2006)

| | | | |
|---|---|---|---|
| Written Czech: | 97 % ženami | 3 % ženama | (inst. pl. 'female') |
| Spoken Czech: | 5 % ženami | 95 % ženama | (inst. pl. 'female') |

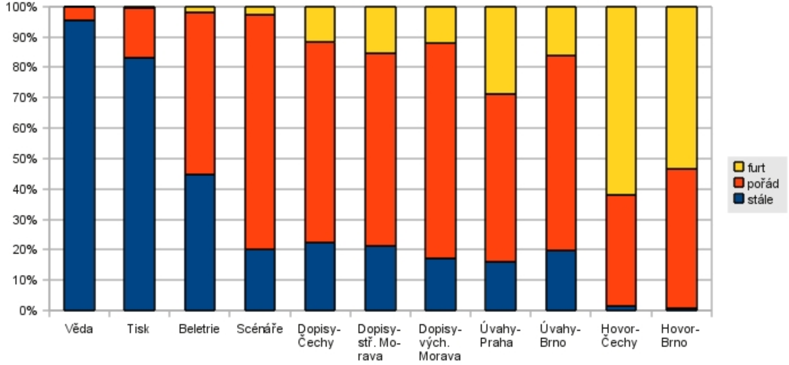
| | | | |
|---|---|---|---|
| Written Czech: | 98 % mladý | 2 % mladej | (nom. sg. masc. 'young') |
| Spoken Czech: | 9 % mladý | 91 % mladej | (nom. sg. masc. 'young') |

# Graph is better than paragraph

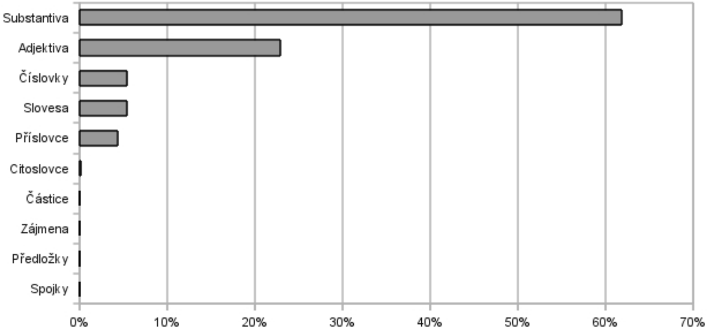
3 variants of ’allways’ – stále, pořád, furt

# Graph is better than paragraph

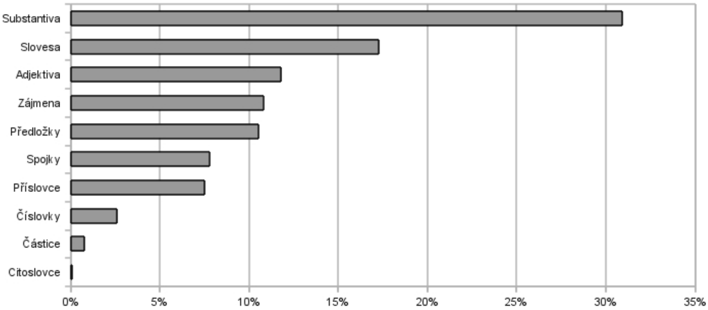
3 variants of 'allways' – stále, pořád, furt



Frequency of parts of speech in lexicon and in texts (token-type distinction).



Frequency of parts of speech in lexicon and in texts (token-type distinction).



# Page layout – morphology

## Main paragraphs of each section (i.e. paradigm):

1. Formal description of the paradigm

# Page layout – morphology

Main paragraphs of each section (i.e. paradigm):

1. Formal description of the paradigm
2. Size of the paradigm (list of the most frequent members)

# Page layout – morphology

## Main paragraphs of each section (i.e. paradigm):

1. Formal description of the paradigm
2. Size of the paradigm (list of the most frequent members)
3. Table with word-forms and variants

# Page layout – morphology

## Main paragraphs of each section (i.e. paragdigm):

1. Formal description of the paradigm
2. Size of the paradigm (list of the most frequent members)
3. Table with word-forms and variants
4. Proportion of frequencies of variants for the whole paradigm + notes

# Page layout – morphology

## Main paragraphs of each section (i.e. paradigm):

1. Formal description of the paradigm
2. Size of the paradigm (list of the most frequent members)
3. Table with word-forms and variants
4. Proportion of frequencies of variants for the whole paradigm + notes
5. Proportion of frequencies of variants for individual lexemes (which differ from overall tendency) + notes

# Page layout – morphology

## Main paragraphs of each section (i.e. paradigm):

1. Formal description of the paradigm
2. Size of the paradigm (list of the most frequent members)
3. Table with word-forms and variants
4. Proportion of frequencies of variants for the whole paradigm + notes
5. Proportion of frequencies of variants for individual lexemes (which differ from overall tendency) + notes
6. Running foot with important information (abreviations, terms etc.)

### 7.1.7.1.4 Vzor ***soudce***

Vzor *soudce* se liší od vzoru *muž* tvarem koncovky v Nsg. Apelativa tohoto vzoru mají kmen zakončený vesměs hláskou *c* a v Nsg. mají koncovku *-e* (o propriích s koncovkou *-e/ě* v Nsg. viz 7.1.7.1.5).

Nejfrekventovanější apelativa patřící k tomuto vzoru: *dárce, důchodce, nástupce, obhájce, obránce, odpůrce, ochránce, poradce, prodejce, průvodce, předchůdce, příjemce, původce, soudce, správce, strážce, tvůrce, vůdce, výrobce, zájemce, zastánce, zástupce*.

Výčet dalších apelativ patřících k tomuto vzoru: *autodopravce, divotvorce/divotvůrce, dohodce, dopravce, dovozce, dozorce, chlebodárce, malovýrobce, mírotvorce/mírotvůrce, mravokárce, nájemce, nálezce, návodce, neplátce, normotvůrce, odhadce, odvozce, ohněstrůjce/ohňostrůjce, oprávce, plátce, podnájemce, podpůrce, porotce, pronájemce, protichůdce, prvovýrobce, přepravce, přestupce, převodce, příkazce, přímluvce, rádce, rozhodce, samoplátce, samosoudce, samovládce, spolutvůrce, spoluvládce, spoluvynálezce, strojvůdce, strůjce, svůdce, šéfporadce, škůdce, únosce, úpadce, úpravce, ústavodárce, velezrádce, velkoprodejce, velkovýrobce, vládce, vlastizrádce, vojevůdce, výherce, vynálezce, výstavce, vývozce, zachránce, zákonodárce, zástavce, zhoubce, zpravodajce, zrádce, žalobce*.

| **Pád** | **Singulár** | **Plurál** |
|---|---|---|
| **Nom** | *soudc-e* | *soudc-i* / *soudc-ové* |
| **Gen** | *soudc-e* | *soudc-ů* |
| **Dat** | *soudc-i* / *soudc-ovi* | *soudc-ům* |
| **Ak** | *soudc-e* | *soudc-e* |
| **Vok** | *soudc-e* / *soudč-e* | *soudc-i* / *soudc-ové* |
| **Lok** | *o soudc-i* / *soudc-ovi* | *o soudc-ích* |
| **Instr** | *soudc-em* | *soudc-i* / *soudc-ema* |

Poznámky k jednotlivým tvarům vzoru *soudce*:

| | |
|---|---|
| *(soudc)-i* / *(soudc)-ovi* | Psaná čeština: skoro vždycky *(soudc)i*<br>Mluvená čeština: údaje nejsou k dispozici, převažuje *(soudc)i* |
| *(soudc)-i* / *(soudc)-ové* | Psaná čeština: výrazně převažuje tvar *(soudc)i*<br>Mluvená čeština: údaje nejsou k dispozici, převažuje *(soudc)i* |
| *(soudc)-i* / *(soudc)-ema* | Psaná čeština: skoro vždycky tvar *(soudc)i*<br>Mluvená čeština: data nejsou k dispozici, převažuje *(soudc)ema* |

Poznámky k jednotlivým substantivům vzoru *soudce*:
V Npl. má většina podstatných jmen (např. *zastánce, výrobce, zájemce, prodejce, poradce, důchodce*) vždy nebo skoro vždycky koncovku *-i*. Koncovka *-ové* se někdy užívá s podstatnými jmény *vládce, vůdce, rádce, svůdce, soudce* a zřídka i se substantivy *správce, strážce, nástupce, tvůrce, průvodce* aj.

V psaných textech je ve Vsg. často až zpravidla zakončení *-ce*, někdy *-če*.

# Multi-word units

- collocations, phrasemes, multi-word (scientific) terms etc.

# Multi-word units

- collocations, phrasemes, multi-word (scientific) terms etc.
- part of the lexicon ⇒ part of grammar

# Multi-word units

- collocations, phrasemes, multi-word (scientific) terms etc.
- part of the lexicon ⇒ part of grammar
- multi-word equivalents for every word class

# Multi-word units

- collocations, phrasemes, multi-word (scientific) terms etc.
- part of the lexicon ⇒ part of grammar
- multi-word equivalents for every word class
- morphology and syntax of multi-word units

# Closed paradigms

## Closed sets of units which are small enough to be listed.

- some nominal paradigms (kuře ’chicken’)

# Closed paradigms

Closed sets of units which are small enough to be listed.

- some nominal paradigms (kuře ’chicken’)
- underived adjectives

# Closed paradigms

## Closed sets of units which are small enough to be listed.

- some nominal paradigms (kuře ’chicken’)
- underived adjectives
- pronouns

# Closed paradigms

Closed sets of units which are small enough to be listed.

- some nominal paradigms (kuře ’chicken’)
- underived adjectives
- pronouns
- some types of numerals

# Closed paradigms

## Closed sets of units which are small enough to be listed.

- some nominal paradigms (kuře ’chicken’)
- underived adjectives
- pronouns
- some types of numerals
- prepositions

# Closed paradigms

Closed sets of units which are small enough to be listed.

- some nominal paradigms (kuře ’chicken’)
- underived adjectives
- pronouns
- some types of numerals
- prepositions
- conjunctions

# Basic corpus tools

## Corpora used:

| | |
|---|---|
| SYN2005 | 100M corpus of written Czech, ballanced, lemmatised, morphologically tagged |
| Oral2006 | 1M corpus of spoken Czech (from Bohemia only), informal unprepared dialogues |
| other | PMK (Prague spoken corpus), BMK (Brno spoken corpus), KSK (Private Correspondence Corpus), SYN2006PUB (300M – newspapers) |

Manatee server – Bonito client – Word-sketch engine (© P. Rychlý)

# Paradigma

## Program Paradigma can:

1. Identify lemmas of the same paradigm

# Paradigma

## Program Paradigma can:

1. Identify lemmas of the same paradigm
   - Number of lemmas in the paradigm and their frequency

# Paradigma

## Program Paradigma can:

1. Identify lemmas of the same paradigm
   - Number of lemmas in the paradigm and their frequency
   - Improve delimitation of paradigms

# Paradigma

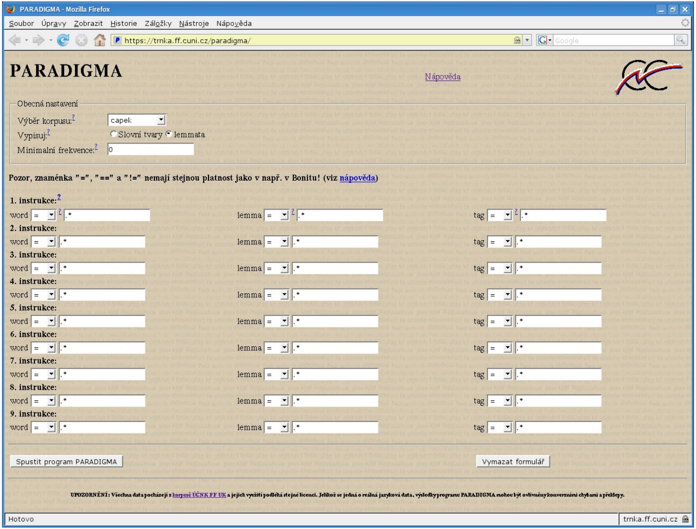
## Program Paradigma can:

1. Identify lemmas of the same paradigm
   - Number of lemmas in the paradigm and their frequency
   - Improve delimitation of paradigms
2. Find out all homonymous word-forms (e.g. nouns – verbs)

# Paradigma

## Program Paradigma can:

1. Identify lemmas of the same paradigm
   - Number of lemmas in the paradigm and their frequency
   - Improve delimitation of paradigms
2. Find out all homonymous word-forms (e.g. nouns – verbs)
3. Improve automatic morphological analysis



PARADIGMA - Mozilla Firefox

https://trnka.ff.cuni.cz/paradigma/

# PARADIGMA

Nápověda

Obecná nastavení

Výběr korpusu:? capek

Vypisuj:? ○ Slovní tvary ● lemmata

Minimalní frekvence:? 0

**Pozor, znaménka "=", "==" a "!=" nemají stejnou platnost jako např. v Bonitu! (viz nápověda)**

**1. instrukce:**?

word = ? .* lemma = ? .* tag = ? .*

**2. instrukce:**

word = .* lemma = .* tag = .*

**3. instrukce:**

word = .* lemma = .* tag = .*

**4. instrukce:**

word = .* lemma = .* tag = .*

**5. instrukce:**

word = .* lemma = .* tag = .*

**6. instrukce:**

word = .* lemma = .* tag = .*

**7. instrukce:**

word = .* lemma = .* tag = .*

**8. instrukce:**

word = .* lemma = .* tag = .*

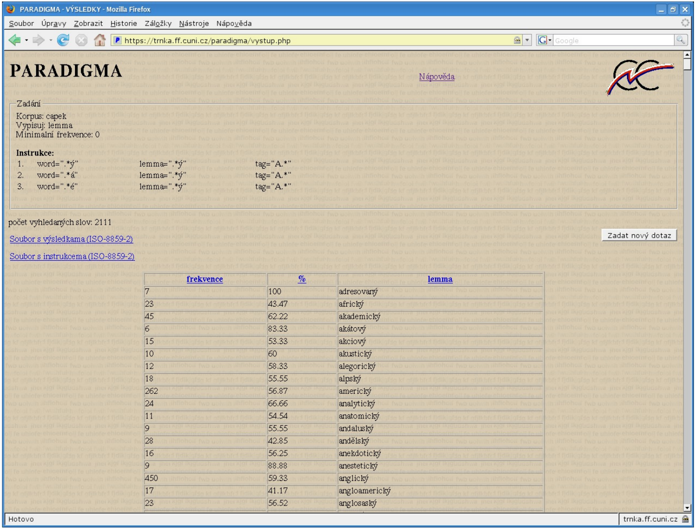
**9. instrukce:**

word = .* lemma = .* tag = .*

Spustit program PARADIGMA

Vymazat formulář

**UPOZORNĚNÍ: Všechna data pocházejí z korpusů ÚČNK FF UK a jejich využití podléhá stejné licenci. Jelikož se jedná o reálná jazyková data, výsledky programu PARADIGMA mohou být ovlivněny konverzními chybami a překlepy.**


PARADIGMA - VÝSLEDKY - Mozilla Firefox
Soubor Úpravy Zobrazit Historie Záložky Nástroje Nápověda
https://trnka.ff.cuni.cz/paradigma/vystup.php
Google
Nápověda
Zadat nový dotaz
Hotovo
trnka.ff.cuni.cz

# PARADIGMA

Zadání

Korpus: capek
Vypisuj: lemma
Minimalní frekvence: 0

**Instrukce:**

1. word=".*ý" lemma=".*ý" tag="A.*"
2. word=".*á" lemma=".*ý" tag="A.*"
3. word=".*é" lemma=".*ý" tag="A.*"

počet vyhledaných slov: 2111

Soubor s výsledkama (ISO-8859-2)

Soubor s instrukcema (ISO-8859-2)

| frekvence | % | lemma |
|---|---|---|
| 7 | 100 | adresovaný |
| 23 | 43.47 | africký |
| 45 | 62.22 | akademický |
| 6 | 83.33 | akátový |
| 15 | 53.33 | akciový |
| 10 | 60 | akustický |
| 12 | 58.33 | alegorický |
| 18 | 55.55 | alpský |
| 262 | 56.87 | americký |
| 24 | 66.66 | analytický |
| 11 | 54.54 | anatomický |
| 9 | 55.55 | andaluský |
| 28 | 42.85 | andělský |
| 16 | 56.25 | anekdotický |
| 9 | 88.88 | anestetický |
| 450 | 59.33 | anglický |
| 17 | 41.17 | angloamerický |
| 23 | 56.52 | anglosaský |

# Slovotvorba

### Program Slovotvorba can:

1. Identify related (derived) words

# Slovotvorba

## Program Slovotvorba can:

1. Identify related (derived) words
2. According to specifications find all words with the same formal relationship

# Slovotvorba

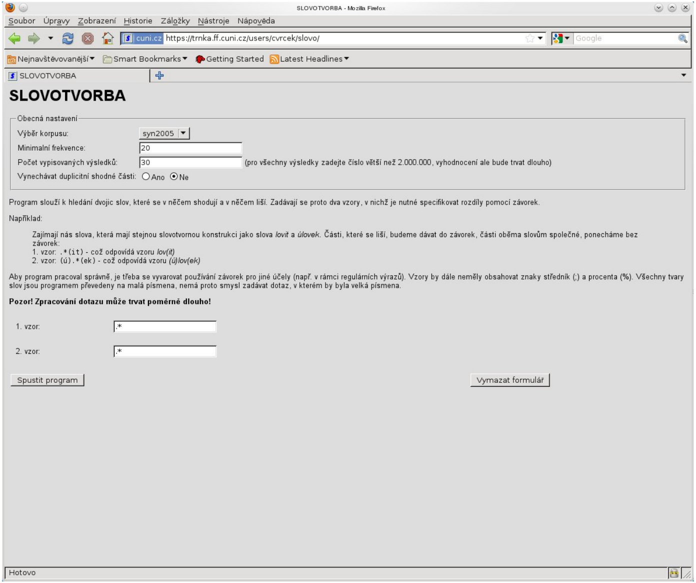
## Program Slovotvorba can:

1. Identify related (derived) words
2. According to specifications find all words with the same formal relationship
3. Identifying what's identical and what's different

# Slovotvorba

## Program Slovotvorba can:

1. Identify related (derived) words
2. According to specifications find all words with the same formal relationship
3. Identifying what's identical and what's different
4. Reveal frequency correspondence


SLOVOTVORBA - Mozilla Firefox
Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda
cuni.cz https://trnka.ff.cuni.cz/users/cvrcek/slovo/
Google
Nejnavštěvovanější Smart Bookmarks Getting Started Latest Headlines
SLOVOTVORBA
Hotovo

## SLOVOTVORBA

Obecná nastavení

Výběr korpusu: syn2005

Minimalní frekvence: 20

Počet vypisovaných výsledků: 30 (pro všechny výsledky zadejte číslo větší než 2.000.000, vyhodnocení ale bude trvat dlouho)

Vynechávat duplicitní shodné části: ○ Ano ● Ne

Program slouží k hledání dvojic slov, které se v něčem shodují a v něčem liší. Zadávají se proto dva vzory, v nichž je nutné specifikovat rozdíly pomocí závorek.

Například:

> Zajímají nás slova, která mají stejnou slovotvornou konstrukci jako slova *lovit* a *úlovek*. Části, které se liší, budeme dávat do závorek, části oběma slovům společné, ponecháme bez závorek:
> 1. vzor: `.*(it)` - což odpovídá vzoru *lov(it)*
> 2. vzor: `(ú).*(ek)` - což odpovídá vzoru *(ú)lov(ek)*

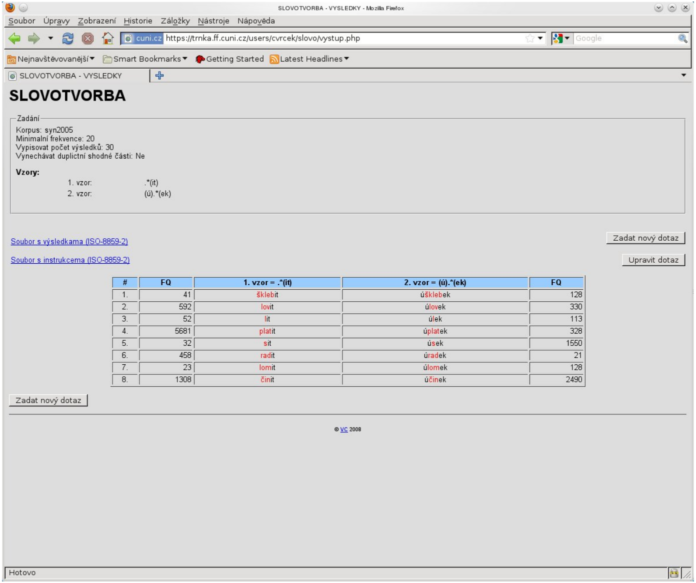
Aby program pracoval správně, je třeba se vyvarovat používání závorek pro jiné účely (např. v rámci regulárních výrazů). Vzory by dále neměly obsahovat znaky středník (;) a procenta (%). Všechny tvary slov jsou programem převedeny na malá písmena, nemá proto smysl zadávat dotaz, v kterém by byla velká písmena.

**Pozor! Zpracování dotazu může trvat poměrně dlouho!**

1. vzor: .*

2. vzor: .*

Spustit program  Vymazat formulář


SLOVOTVORBA - VYSLEDKY - Mozilla Firefox
Soubor Úpravy Zobrazení Historie Záložky Nástroje Nápověda
cuni.cz https://trnka.ff.cuni.cz/users/cvrcek/slovo/vystup.php
Google
Nejnavštěvovanější Smart Bookmarks Getting Started Latest Headlines
SLOVOTVORBA - VYSLEDKY
Zadat nový dotaz
Upravit dotaz
Zadat nový dotaz
Hotovo

# SLOVOTVORBA

Zadání

Korpus: syn2005
Minimalní frekvence: 20
Vypisovat počet výsledků: 30
Vynechávat duplictní shodné části: Ne

**Vzory:**

1. vzor: .*(it)
2. vzor: (ú).*(ek)

Soubor s výsledkama (ISO-8859-2)

Soubor s instrukcema (ISO-8859-2)

| # | FQ | 1. vzor = .*(it) | 2. vzor = (ú).*(ek) | FQ |
|---|---|---|---|---|
| 1. | 41 | šklebit | úšklebek | 128 |
| 2. | 592 | lovit | úlovek | 330 |
| 3. | 52 | lit | úlek | 113 |
| 4. | 5681 | platit | úplatek | 328 |
| 5. | 32 | sit | úsek | 1550 |
| 6. | 458 | radit | úradek | 21 |
| 7. | 23 | lomit | úlomek | 128 |
| 8. | 1308 | činit | účinek | 2490 |

© VC 2008

## Corpus-based concepts and advantages of corpus approach:

- Relatively complete and precise description

### Corpus-based concepts and advantages of corpus approach:

- Relatively complete and precise description
- Based on real language data (important for descriptive nature)

## Corpus-based concepts and advantages of corpus approach:

- Relatively complete and precise description
- Based on real language data (important for descriptive nature)
- Differences of language forms (written vs spoken)

### Corpus-based concepts and advantages of corpus approach:

- Relatively complete and precise description
- Based on real language data (important for descriptive nature)
- Differences of language forms (written vs spoken)
- Closed classes

### Corpus-based concepts and advantages of corpus approach:

- Relatively complete and precise description
- Based on real language data (important for descriptive nature)
- Differences of language forms (written vs spoken)
- Closed classes
- Lots of examples

## Corpus-driven concepts and desiderata for future work:

- collocations and multi-word units

### Corpus-driven concepts and desiderata for future work:

- collocations and multi-word units
- colligations on the level of two positions (some words co-occur with certain grammatical categories)

### Corpus-driven concepts and desiderata for future work:

- collocations and multi-word units
- colligations on the level of two positions (some words co-occur with certain grammatical categories)
- colligations on the level of one position (some words are unusually often in certain grammatical categories)

# ...instead of conclusion

**Why choose the descriptive approach to grammar over the prescriptive?**

1. because that's what users will appreciate

# ...instead of conclusion

## Why choose the descriptive approach to grammar over the prescriptive?

1. because that’s what users will appreciate
2. (even if they won’t) because we do not have the right to intervene to the language development

# …instead of conclusion

## Why choose the descriptive approach to grammar over the prescriptive?

1. because that’s what users will appreciate
2. (even if they won’t) because we do not have the right to intervene to the language development
3. (even if we have) because we do not know how to regulate the language

# ...instead of conclusion

**Why choose the descriptive approach to grammar over the prescriptive?**

1. because that's what users will appreciate
2. (even if they won't) because we do not have the right to intervene to the language development
3. (even if we have) because we do not know how to regulate the language
4. (but mostly) because that's our job and that's what we have data for.

**Thank you for your attention!**