Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Conclusion 0000000 00000 0000

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

▲ロト ▲ □ ト ▲ □ ト ▲ □ ト ● ● の Q ()

- Description vs prescription
 - Case of GCC
 - Principles of noniterventional approach

Tools

2 Practice of GCC

- Variants in GCC
- Graphs
- Page layout
- Multi-word units
- Closed paradigms

3 Tools

- Paradigma
- Slovotvorba
- Corpus-based concepts
- 5 Corpus-driven concepts
- 6 Conclusion

 Description
 Practice of GCC
 Tools
 Corpus-based concepts
 Corpus-driven concepts
 Conclusion

 Description
 vs
 prescription
 vs
 prescription
 vs
 vs<

▲□▶ ▲□▶ ▲□▶ ▲□▶ = 三 のへで

Prescriptive situation

- monopoly codification
- discrepancy between the codification and the usage

 Description
 Practice of GCC
 Tools
 Corpus-based concepts
 Corpus-driven concepts
 Conclusion

 Observery concept
 Occord
 O

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)

Practice of GCC 0000000

C Tools 0000 Corpus-based concept

oncepts Corpus

pus-driven concepts

▲ロト ▲ □ ト ▲ □ ト ▲ □ ト ● ● の Q ()

Conclusion

Case of Grammar of Contemporary Czech

VÁCLAV CVRČEK A KOLEKTIV AUTORŮ

MLUVNICE SOUČASNÉ ČEŠTINY



JAK SE PÍŠE A JAK SE MLUVÍ



Description vs prescription Concept of Minimal Intervention (CMI)

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

▲ロト ▲ □ ト ▲ □ ト ▲ □ ト ● ● の Q ()

Concept of Minimal Intervention (CMI)

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

Premises:

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.

Practice of GCC 0000000 Tools

Corpus-based concep

concepts Corp

orpus-driven concepts

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Conclusion

Concept of Minimal Intervention

Premises (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.

Tools

Corpus-based concep

oncepts Corpus

us-driven concepts

Conclusion

Concept of Minimal Intervention

Premises (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.
- 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?

Tools

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.

ション ふぼう メリン メリン しょうめん

What does the CMI approach to language represent?

Tools

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.

ション ふぼう メリン メリン しょうめん

What does the CMI approach to language represent?

Tools

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:

Tools

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:

Tools

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

Description vs prescription ○○○○○○●	Practice of GCC	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

•••

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.

- 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.
- 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.)



• published in 2010 (first part only)



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



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



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



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

estimate (expected 2011)



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

- estimate (expected 2011)
- corpus-based



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

- estimate (expected 2011)
- corpus-based
- descriptive (CMI-style)

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

▲ロト ▲ □ ト ▲ □ ト ▲ □ ト ● ● の Q ()

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Conclusion

Grammar of Contemporary Czech – variants

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

Written Czech: Spoken Czech:			(inst. pl. 'female') (inst. pl. 'female')
Written Czech:	/	2 % mladej	(nom. sg. masc. 'young')
Spoken Czech:		91 % mladej	(nom. sg. masc. 'young')

▲ロト ▲ □ ト ▲ □ ト ▲ □ ト ● ● の Q ()

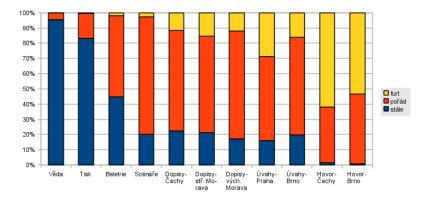
▲□▶ ▲□▶ ▲□▶ ▲□▶ = 三 のへで

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

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Conclusion of the second s

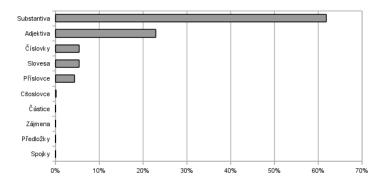
Graph is better than paragraph

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



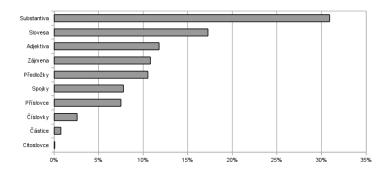
◆□▶ ◆□▶ ◆臣▶ ◆臣▶ ○臣 - のへで

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



▲□▶▲圖▶★≧▶★≧▶ 差 のへで

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



◆□▶ ◆□▶ ◆臣▶ ◆臣▶ ─臣 ─のへで

Page layout –	morphol	ogy			
Description vs prescription	Practice of GCC ○○○●○○○	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

▲□▶▲□▶▲□▶▲□▶ ▲□ ● のへで

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

Formal description of the paradigm

Description vs prescription OCODOCOCO Practice of GCC OCODOCOCO Provide of GCC OCOPUs-based concepts Corpus-driven concepts Corpus-driven concepts Conclusion Conclusion

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

- Formal description of the paradigm
- Size of the paradigm (list of the most frequent members)

Description vs prescription OOOOOOOO
Practice of GCC OOOOOOO
Corpus-based concepts OOOO
Corpus-based concepts Corpus-driven concepts Conclusion
Conclusion

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

- Formal description of the paradigm
- Size of the paradigm (list of the most frequent members)

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Table with word-forms and variants

Description vs prescription OCODOCOCO Practice of GCC OCODOCOCO Prage layout – morphology Corpus-based concepts Corpus-based concepts Corpus-driven concepts Corpus-driven concepts Conclusion

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

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

Description vs prescription OCODE OCO Practice of GCC OCODE OCO OCODE OCODE

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

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

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

 Proportion of frequencies of variants for individual lexemes (which differ from overall tendency) + notes Description vs prescription OCODE OCO Practice of GCC OCODE OCO OCODE OCODE

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

- Formal description of the paradigm
- Size of the paradigm (list of the most frequent members)
- Table with word-forms and variants
- Proportion of frequencies of variants for the whole paradigm + notes
- Proportion of frequencies of variants for individual lexemes (which differ from overall tendency) + notes
- 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/č v Nsg. viz 7.1.7.1.5).

Nejfrekventovanější apelativa patřící k tomuto vzoru: dárce, důchodce, nástupce, obhájce, obránce, odpůrce, ochránce, porudce, prodejce, privodce, předchůdce, příjemce, původce, soudce, sprivce, strážee, tvárce, vádce, výrobce, zájemce, zastánce, zástupce.

Výčet dálšíh apelaiv patříckí k tomato vzon: anadoporece, divorarezdivoráne, divolace, doproce, dovoze, dozor, elibodatre, androvánek, ministračnetiniváne, mroválov, minister, androka, ediparnitovác, nepláte, nomováre, odhude, odvoze, ohnistnije, člohbastnije, opráve, pláte, poducjime, popodpisce, pontec, pomojane, printikale, provývake, přímyroce, přintege, poducke, pláte, poducjime, ninove, nieke, nrihoder, samojalice, samosade, samovláde, spoluritver, spolavláde, spisovale, příste, printipe, volike, volike, žeforudne, klade, ninose, špade-teimrev, stávnoste, veleznáke, velikopradyce, volike, vidiační doke, rámor, špade-teimrev, stávnoste, veleznáke, velikopradyce, velkovývboc, hláče, vidiační koče, vidier, veleznéke, velikov, velikove, vyšance, výstorec, svistorec, stávnodáre, stávne, bodnec, parvodke, prinde, vidier, kader, spisovate, vistorec, výstorec, výstorec, svistorec, výstorec, vistorec, svistorec, výstorec, v

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-ich
Instr	soudc-em	soudc-i / soudc-ema

Poznámky k jednotlivým tvarům vzoru soudce:

(soudc)-i / (soudc)-ovi	Psaná čeština: skoro vždycky (soudc)i
	Mluvená čeština: údaje nejsou k dispozici, převažuje (soudc)i
(soudc)-i / (soudc)-ové	Psaná čeština: výrazně převažuje tvar (soudc)i
	Mluvená čeština: údaje nejsou k dispozici, převažuje (soudc)i
(soudc)-i / (soudc)-ema	Psaná čeština: skoro vždycky tvar (soudc)i
	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, dáchodce) vždy nebo skoro vždycky koncovku -i. Koncovka -ové se někdy užívá s podstatnými jmény vládce, vládce, vládce, vládce, svádce, soudce a zřídka i se substantivy správce, stráčce, nástapce, tvírace, privodce aj.

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

Rod: mužský = maskulinum (mask.), ženský = fermininum (fem.), střední = neutrum (neut.); Číslo: jednotné = singulár (sg.), množné = plurál (pl.); Pád: 1 nominativ (N, nom.), 2 genitiv (G, gen.), 3 dativ (D, dat.), 4 akuzativ (A, ak.), 5 vokativ (V, vok.), 6 lokil (L, lok.), 7 instrumentál (L, instr.)

Description vs prescription	Practice of GCC ○○○○○●○	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion
Multi-word ι	units				

◆□▶ ◆□▶ ◆ □▶ ◆ □▶ ● □ ● ● ● ●

Description vs prescription	Practice of GCC ○○○○○●○	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion
Multi-word u	inits				

◆□▶ ◆□▶ ◆ □▶ ◆ □▶ ● □ ● ● ● ●

• part of the lexicon \Rightarrow part of grammar

0000000	0000000	0000		
Multi-word ı	inits			

▲□▶ ▲□▶ ▲ □▶ ▲ □▶ ▲ □ ● ● ● ●

- part of the lexicon \Rightarrow part of grammar
- multi-word equivalents for every word class

Multi-word ı	inits				
Description vs prescription	Practice of GCC ○○○○●○	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

ション ふぼう メリン メリン しょうめん

- part of the lexicon \Rightarrow part of grammar
- multi-word equivalents for every word class
- morphology and syntax of multi-word units

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

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

• some nominal paradigms (kuře 'chicken')

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

closed paradigins

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

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

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Conclusion

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Closed paradigms

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

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

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Conc

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

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

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Co 0000000 0000 0000

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

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Conclusion

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

Practice of GCC 0000000 Tools

Corpus-based conce

d concepts Co

rpus-driven concepts

ション ふぼう メリン メリン しょうめん

Conclusion

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)

▲□▶ ▲□▶ ▲□▶ ▲□▶ = 三 のへで

Paradigma

Program Paradigma can:

Identify lemmas of the same paradigm

Description vs prescription Practice of GCC Tools Corpus-ba

Corpus-based concept

Corpus-driven

イロト 不得 とうほう イヨン

3

Conclusion

Paradigma

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

Practice of GCC 0000000

Tools ●000 Corpus-based concep

ts Corpus-driver

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Conclusion

Paradigma

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

Tools

Corpus-based concep

epts Corpus-driv

6 Conclusion

◆□▶ ◆□▶ ◆三▶ ◆三▶ ・三 つく⊙

Paradigma

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

Practice of GCC 0000000

Tools ●000 Corpus-based concep

pts Corpus-drive

cepts Conclusion

◆□▶ ◆□▶ ◆三▶ ◆三▶ ・三 つく⊙

Paradigma

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

Practice of GCC

Corpus-based co

Tools ○●○○

d concepts Corpus-c

us-driven concepts

Conclusion

oubor Úpr <u>avy Z</u> obrazit <u>H</u> istorie Zá				
🗉 - 🗼 - 🧭 📀 🚹 🗷 https://	📾 💌 🔀 - Google	📾 💌 🔀 - Google		
PARADIGMA		Népověda	RC	
Obecná nastavení				
Výběr korpusu: ² capek	🔁 stillelikiplig trappered fjelikiplig trappeted fjelikiplig trappeted fjelikiplig trappeted fjelit.			
Vypisuj. ² CSlovni tva	ry 🕫 lemmata			
Minimalní frekvence:	Now for all finds the childle to start and the two blacks and show the child			
instrukce: ²	temma = _] ² .•	tag = _] ² [•	sykted fisik/also kt of a processor algorithms alofe chicefrit oli fau un altiohus two u	
instrukce:				
ord = .	lemma = 💌 .*	tag = 💌 .*		
instrukce:				
ord = 💌 .*	lemma =*	tag =*		
instrukce:	vote ethicetint oil te uniote ethicetint oil te uniote ethicetint oil te uniote ethicet Non shriothoir two ucivon sigliothoir	nfoj fe uhinte etitoethfoj te uhinte-etitoethfoj fe uhinte etitoethfoj te u		
ord = 💽 .*	lemma =*	tag = 🔄 .*		
ord =	lemma = 📩 .*	tag =*		
instrukce:	tourned =			
ord = 💌 .*	lemma = 💌 .*	tag = 💌 .*		
instrukce:				
ord = 💌 .*	lemma =*	teg = 🔄 .*		
instrukce:	preskjol Rigbostova, preskjol Rigbostova, preskjol Rigbostova, preskjol Ri	iducativa, jinerkidi Ridocativa, jinerkidi Ridocativa, jinerkidi Ridocativ Moli je ubiote etilopintoli je vojo		
ord = 🔄 .*	lemma =*	tag = 💌 .*		
ord = .	temma = 🔄 .*	teg =*		
	and the children is a state of the second of the second state of t	in for the universe choose in the state of the second state in the second state is the second state in the second state is the		
Spustit program PARADIGMA		Vymazat formulář		
	ázeji z kzepcně ÚČNN PP UK a jejich vychtří podlehá ctojná licenci. Jelkná ce jedná o redná ja			

Tools ○●○○

• 🐵 • 🞯 🙆 🏠	https://tmka.ff.cuni.cz/paradigma	/vystup.php		📾 💌 💽 • Google
ARADIGM		owning in the united school inform two working sinflicture and show the color of the school should be address of the school in Reglacement of the school school and the united school schoo	<u>Népověda</u>	RC
Zadání Korpus: capek /ypisuj: lemma Minimalní frekvence: 0				
instrukce:				
1. word=".*ý"	lemma=".*ý*	tag="A.*"		
2. word=".*á"	lemma=".*ý*	tag="A.*"		
8. word=".*é"	lemma=".*ý*	tag="A.*"		
et vyhledaných slov: 2111				
ubor s výsledkama (ISO-8				Zadat nový dota
	of te uniote chiceInfoll fe uniote chi			
ubor s instrukcema (ISO-8	on re uniore-criscenti on re unicce eni			
Construction of the Difference	on re uniore-criscenti on re unicce eni			
Contra Concerna Contra Universit	1859-2) frekvence	<u> </u>	lemma	
Contra Concerna Contra Universit	1859-2) <u>frekvence</u> 7	100	adresovaný	
Contra Concerna Contra Universit	2859-2) frekvence 7 23	100 43.47		
Contra Concerna Contra Universit	1859-2) <u>frekvence</u> 7	100	adresovaný	
Contra Concerna Contra Universit	2859-2) frekvence 7 23	100 43.47	adresovaný africký	
construction of the printing of	1859-2) frekvence 7 23 45	100 43.47 62.22	adresovaný africký akademický akátový	
construction of the printing of	18899-20 7 23 45 6	100 43.47 62.22 83.33	adresovaný africký akademický akátový akotový	
construction of the printing of	frekvence 7 23 45 6 15 10	100 43.47 62.22 83.33 53.33 60	adresovaný africhý akademichý akatový akciový akatichý	
construction of the printing of	8859-21 7 23 45 6 15 10 12	100 43.47 62.22 83.33 53.33 60 58.33	adresovaný africhý akademichý akátový akatoloj akatoloj alextický	
construction of the printing of	frekvener 7 23 45 6 15 10 12 18	100 43.47 62.22 83.33 53.33 60 58.33 55.55	adresovaný adrichý akademický akadový akadrový akadroký akadroký akadroký akadroký	
construction of the printing of	5539-20 7 23 45 6 15 15 10 12 12 18 262	100 43.47 62.22 83.33 53.33 60 58.33 55.55 56.87	adresovang adrichý akademický akatový akatový akatoty alagorický alagorický alagotický	
construction of the printing of	2829-20 7 frskvenee 22 45 6 15 10 12 18 202 24	100 43.47 62.22 83.33 53.33 60 58.33 55.55 56.87 66.66	indersovergi adricity adademicity adademicity adatory adatory adaptory adaptory adaptory adaptory adaptory adaptory adaptory	
construction of the printing of	8399-20 7 5 45 45 6 15 10 12 18 20 20 24 11	100 43.47 62.22 83.33 53.33 60 58.33 55.55 56.87 66.66 54.54	adresovanej adartický adadomický adatoroj adatoroj adatoroj adaptický adaptický anartický anartický	
Contra Concerna Contra Universit	5859-20 7 23 45 6 15 10 12 18 202 24 4 11 9	100 43.47 62.22 83.33 53.33 60 58.33 55.55 56.87 66.66 54.54 55.55	indersovergi an	
Contra Concerna Contra Universit	1992) 7 23 45 6 15 10 12 14 18 24 4 11 9 2 2	100 43.47 62.22 83.33 60 59.33 55.55 56.87 66.66 54.54 55.55 42.85	indersovangi antro bij kladorni kladovi aktoroty	
Contra Concerna Contra Universit	Extense 7 23 45 6 15 10 16 12 17 24 26 24 17 24 26 24 27 24 28 24 16 12 17 25 28 24 16 12	100 43.47 62.22 83.33 53.33 60 55.35 56.87 66.66 54.54 55.55 42.85 56.25	datasovardy datasoly	
Contra Concerna Contra Universit	1992) 7 23 45 6 15 10 12 10 12 24 11 9 9 24 11 9 9 24 11 9 9 2 16 9 9	100 43,47 62,22 83,33 60 53,33 55,55 56,87 66,66 54,55 55,55 42,85 66,25 88,88	intersorverig Andready Andread	
Contra Concerna Contra Universit	Extense 7 23 45 6 15 10 16 12 17 24 18 24 19 24 20 24 21 24 22 24 24 24 25 24 26 24 27 28 28 29 20 20 26 29 200 200	100 43,47 62,22 83,33 53,33 60 55,55 56,87 66,66 54,54 55,55 54,54 55,55 54,54 55,55 54,54 55,55 56,25 88,88 59,33	datasovardy datasoly	
Construction of the Difference	1992) 7 23 45 6 15 10 12 10 12 24 11 9 9 24 11 9 9 24 11 9 9 2 16 9 9	100 43,47 62,22 83,33 60 53,33 55,55 56,87 66,66 54,55 55,55 42,85 66,25 88,88	intersorverig Andready Andread	

Description vs prescription Practice of GCC **Tools** Corpus-based concepts C

<□> <@> < E> < E> = E = のへで

Slovotvorba

Program Slovotvorba can:

Identify related (derived) words

Tools

0000

Corpus-based con

l concepts Cor

pus-driven concepts

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Conclusion

Slovotvorba

Program Slovotvorba can:

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

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Slovotvorba

Program Slovotvorba can:

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

Tools

0000

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Slovotvorba

Program Slovotvorba can:

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

Tools

0000

Reveal frequency correspondence

Tools ○○○●

۵ 🕑		SLOVOTVORBA - Mozilla Fielfox	ی ک
	storie Zál <u>o</u> žky <u>N</u> ástroji		
🔶 🌳 🔹 🐯 😫 🔛 🔟	cuni.cz https://tmka.ff.ci	uni.cz/users/cvrcek/slovo/	☆ ▼ 🚺 🗙 Google
🛅 Nejnavštévovanéjší 🔻 🛅 Smar	t Bookmarks 🔻 🌘 Gettir	ng Started 🔝Latest Headines 🖲	
SLOVOTVORBA	4		
SLOVOTVORBA			
- Obecná nastavení			
Výběr korpusu:	syn2005 🔻		
Minimalni frekvence:	20		
Počet vypisovaných výsledků:	30	(pro všechny výsledky zadejte číslo větší než 2.000.000, vyhodnoc	eni ale bude trvat dlouho)
Vynechávat duplicitní shodné části	i: O Ano ⊙ Ne		
Program slouží k hladání dvojic slov	ktorá so v něřom shodují o :	v něčem liši. Zadávají se proto dva vzory, v nichž je nutné specifikovat ro	zdíly nomoci závorak
Napfiklad:		- needen nee zaaang ee prete and rzery, i menz je name apoenneeder re	and bounder foreign.
Zajimaji nás slova, která maji r závorek: 1. vzor: .*(it) - což odpovidá 2. vzor: (ú) .*(ek) - což odpo	i vzoru <i>lov(i</i> t)	kci jako slova fovit a úlovefi. Části, které se liší, budeme dávat do závor	ek, části oběrna slovům společné, ponechárne bez
Aby program pracoval správně, je třeb slov jsou programem převedeny na m	sa se vyvarovat používání zár alá pismena, nemá proto sr	vorek pro jiné účely (např. v rámci regulárnich výrazů). Vzory by dále ner nysl zadávat dotaz, v kterém by byla velká pismena.	měly obsahovat znaky středník (;) a procenta (%). Všechny tva
Pozor! Zpracování dotazu může tr	vat pomérné dlouho!		
1. vzor:			
2. vzor:			
Spustit program			Vymazat formulár
Hotovo			
Piotovo			Je

ractice of GCC

Tools Corpus-base

is-based concepts

Corpus-driven concepts

Conclusion

🕨 🗣 🔹 🕄 🔕	🏠 🚺 🖬	uni.cz https://trnka.f	f.cuni.cz/users/cvrcek/slovo/vystup.j	php	्रे 🔻 🚺 🕈 Google	
Nejnavštěvovanější 🔻	Smart E	Bookmarks 🔻 🌘 Ge	tting Started 🔝Latest Headlines 🖲	•		
SLOVOTVORBA - VYSL	EDKY	4				
LOVOTVOR	BΔ					
Zadání Korpus: syn2005 Minimalní frékvence: 20 Vypisovat počet výsledků Vynechávat duplictní shor	: 30 Iné části: N	8				
Vzory:						
1. vzor:		.*(it)				
2. vzor:		(ú).*(ek)				
						Zadat nový dot
oubor s výsledkama (ISO-	8859-2)					2.auar novy do
oubor s instrukcema (ISO	-8859-2)					Upravit dot
	#	FQ	1. vzor – .*(it)	2. vzor - (ú)."(ek)	FQ	
	1.	41	šklebit	úšklebek	128	
			lovit	úlovek	330	
	2.	592		úlek		
	2.	52	lit		113	
	2. 3. 4.	52 5681	platit	úplatek	328	
	2. 3. 4. 5.	52 5681 32	platit sit	úplatek úsek	328	
	2. 3. 4. 5. 6.	52 5681 32 458	platit sit radit	úplatek úsek úradek	328 1550 21	
	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radk lomit	úplatek úsek úradek úlomek	308 1650 21 128	
	2. 3. 4. 5. 6.	52 5681 32 458	platit sit radit	úplatek úsek úradek	328 1550 21	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radk lomit	úplatek úsek úradek úlomek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radk lomit	úplatek úsek úradek úlomek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radk lomit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zadat nový dotez	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zedat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	
Zadat nový dotaz	2. 3. 4. 5. 6. 7.	52 5681 32 458 23	platit sit radit Iomit činit	óplatak disek űradek űörmek űörmek	308 1650 21 128	

◆□▶ ◆□▶ ◆臣▶ ◆臣▶ ─臣 ─のへで

Description vs prescription	Practice of GCC 0000000	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

◆□▶ ◆□▶ ◆三▶ ◆三▶ ・三 つく⊙

Corpus-based concepts and advantages of corpus approach:

• Relatively complete and precise description

Description vs prescription	Practice of GCC 0000000	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

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

◆□▶ ◆□▶ ◆三▶ ◆三▶ ・三 つく⊙

Description vs prescription	Practice of GCC 0000000	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

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

・ロト ・ 母 ト ・ ヨ ト ・ ヨ ト ・ ヨ

• Differences of language forms (written vs spoken)

Description vs prescription	Practice of GCC	Tools	Corpus-based concepts	Corpus-driven concepts	Conclusion

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

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

- Differences of language forms (written vs spoken)
- Closed classes

Description vs prescription	Practice of GCC 0000000	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

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

(日) (日) (日) (日) (日) (日) (日)

- Differences of language forms (written vs spoken)
- Closed classes
- Lots of examples

Description vs prescription	Practice of GCC 0000000	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

▲□▶ ▲□▶ ▲ □▶ ▲ □▶ ▲ □ ● ● ● ●

Corpus-driven concepts and desiderata for future work:

• collocations and multi-word units

Description vs prescription	Practice of GCC 0000000	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

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)

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Description vs prescription	Practice of GCC 0000000	Tools 0000	Corpus-based concepts	Corpus-driven concepts	Conclusion

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)

◆□▶ ◆□▶ ◆三▶ ◆三▶ ・三 つく⊙

• colligations on the level of one position (some words are unusually often in certain grammatical categories)

▲□▶ ▲□▶ ▲□▶ ▲□▶ = 三 のへで

Why choose the descriptive approach to grammar

Why choose the descriptive approach to grammar over the prescriptive?

Decause that's what users will appreciate

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts Conclusion

...instead of conclusion

Why choose the descriptive approach to grammar over the prescriptive?

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

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus-driven concepts 00000000 0000 0000 Conclusion

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

...instead of conclusion

Why choose the descriptive approach to grammar over the prescriptive?

- because that's what users will appreciate
- (even if they won't) because we do not have the right to intervene to the language development
- (even if we have) because we do not know how to regulate the language

Description vs prescription Practice of GCC Tools Corpus-based concepts Corpus 00000000 00000 0000

◆□▶ ◆□▶ ◆□▶ ◆□▶ ● □ ● ○○○

...instead of conclusion

Why choose the descriptive approach to grammar over the prescriptive?

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

Description vs prescription	Practice of GCC	Tools	Corpus-based concepts	Corpus-driven concepts	Conclusion

Thank you for your attention!

◆□▶ ◆□▶ ◆ □▶ ◆ □▶ ● □ ● ● ● ●