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 noninterventional 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

<table>
<thead>
<tr>
<th>Prescriptive situation</th>
</tr>
</thead>
<tbody>
<tr>
<td>monopoly codification</td>
</tr>
<tr>
<td>discrepancy between the codification and the usage</td>
</tr>
</tbody>
</table>
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

KARLINUM
Concept of Minimal Intervention (CMI) (Cvrček 2009) – theoretical background for GCC in the question of interventionalism.
Concept of Minimal Intervention (CMI) (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.
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.
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.)
published in 2010 (first part only)
Grammar of Contemporary Czech – basic information

- published in 2010 (first part only)
- designed for students (not academic description)
<table>
<thead>
<tr>
<th>Description vs prescription</th>
<th>Practice of GCC</th>
<th>Tools</th>
<th>Corpus-based concepts</th>
<th>Corpus-driven concepts</th>
<th>Conclusion</th>
</tr>
</thead>
</table>

**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:
  - introduction to study of language, introduction to study of Czech (incl. history), phonology, lexicology, word formation, morphology, basic syntax, stylistic, orthography (writing system)
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 (writing 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 (writing 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 (writing system)
  2. syntax (expected 2011)
- corpus-based
- descriptive (CMI-style)
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).
Main paragraphs of each section (i.e. paradigm):

1. Formal description of the paradigm
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)
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
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
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
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 (abbreviations, terms etc.)
Vzor soudce se liší od vzoru muž tvarem koncovky v Nsg. Apelativa tohoto vzoru mají kmen zakončený vesměs hláskou -e a v Nsg. mají koncovku -e (o propриích s koncovkou -elé v Nsg. viz 7.1.7.1.5).

Nejřídkovějším apelativa patřící k tomuto vzoru: dáře, důchodce, nástupce, obhájce, obránce, odpáře, ochránce, poradce, prodeje, přípravec, převede, příjemce, příznivce, správce, strážce, tvůrce, vědec, výrobců, zájemce, zástupce.

Výčet dalších apelativ patřících k tomuto vzoru: autodopravce, divotvorce, dobrodružce, dobrodružka, rozhodce, rozhodčí, tvůrce, vůdce, výrobce, zákonodárce, zástupce, zhoubce, zpravodajce, zrádce, žalobce.

**Poznámky k jednotlivým substantivům vzoru soudce:**

<table>
<thead>
<tr>
<th>Pád</th>
<th>Nom</th>
<th>Gen</th>
<th>Dat</th>
<th>Ak</th>
<th>Vok</th>
<th>Lok</th>
<th>Instr</th>
</tr>
</thead>
<tbody>
<tr>
<td>soudc-e</td>
<td>soudc-e</td>
<td>soudc-i</td>
<td>soudc-ů</td>
<td>soudc-e</td>
<td>soudc-ů</td>
<td>o soudc-i</td>
<td>soudc-em</td>
</tr>
<tr>
<td>soudc-ové</td>
<td>soudc-ovi</td>
<td>soudc-ové</td>
<td>soudc-ové</td>
<td>soudc-ům</td>
<td>soudc-ové</td>
<td>o soudc-ích</td>
<td>soudc-ema</td>
</tr>
</tbody>
</table>

Poznámky k jednotlivým tvarům vzoru soudce:

- **Nom:**
  - (soudc-i) / (soudc-ovi) Pán se: skoro vždy nevšední (soudc-i)
  - (soudc-ové) Pán se: skoro vždy nevšední (soudc-ové)

- **Dat:**
  - (soudc-ové) Pán se: výrazně převažuje tvar (soudc-i)
  - (soudc-i) Pán se: výrazně převažuje tvar (soudc-ové)

- **Vok:**
  - (soudc-ů) Pán se: je nejžádoucí (soudc-i)
  - (soudc-ů) Pán se: je nejžádoucí (soudc-ové)

- **Instr:**
  - (soudc-em) Pán se: je nejžádoucí (soudc-ové)
  - (soudc-ema) Pán se: je nejžádoucí (soudc-i)

Poznámky k jednotlivým substantivům vzoru soudce:

V Npl. má většina podstatných jmen (např. zastánce, výrobce, zájemce) vždy nebo skoro vždy v koncovku -i. Koncovka -ové se někdy užívá s podstatnými jmény vládce, vědec, rádce, svědec, soudce a zřídka i se substantivy správce, strážce, nástupce, tvůrce, přípravec aj.

V psaných textech je ve Vsg. často až zpravidla zakončené -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 $\Rightarrow$ part of grammar
Multi-word units

- collocations, phrasemes, multi-word (scientific) terms etc.
- part of the lexicon $\Rightarrow$ part of grammar
- multi-word equivalents for every word class
Multi-word units

- collocations, phrasemes, multi-word (scientific) terms etc.
- part of the lexicon $\Rightarrow$ 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)
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
Program Paradigma can:

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

Výběr korpusu:

Vypínavý slovníky:

Minimalní frekvence:

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

1. instrukce

2. instrukce

3. instrukce

4. instrukce

5. instrukce

6. instrukce

7. instrukce

8. instrukce

9. instrukce

Spuštění programu PARADIGMA

Vymazat formulář

Upozornění: Vizualizace dat je k dispozici v korpusu ČOÚFF Mohou však omezovat sovětské regulace. Zvlášť je však vhodná jazyková data, výsledky programu PARADIGMA nebudou však integrace se všemi otevřenými skupinami a pravidly.
PARADIGMA

Zadání
Korpus ooprek
Vypisují lemma.
Minimalní frekvence: 0

Instrukce:
1. word="*5" lemma="*5" tag="A.*"
2. word="*6" lemma="*5" tag="A.*"
3. word="*5" lemma="*5" tag="A.*"

počet vyhledaných slov: 2111

Soubor s výsledným (ISO-8859-2)
Soubor s instrukcí (ISO-8859-2)

<table>
<thead>
<tr>
<th>frekvence</th>
<th>%</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>100</td>
<td>adresovaný</td>
</tr>
<tr>
<td>43.47</td>
<td>34</td>
<td>afrikační</td>
</tr>
<tr>
<td>62.22</td>
<td>25</td>
<td>akademický</td>
</tr>
<tr>
<td>83.33</td>
<td>16</td>
<td>abstraktivní</td>
</tr>
<tr>
<td>53.33</td>
<td>10</td>
<td>abstraktivní</td>
</tr>
<tr>
<td>60</td>
<td>5</td>
<td>akceptor</td>
</tr>
<tr>
<td>58.58</td>
<td>5</td>
<td>akceptor</td>
</tr>
<tr>
<td>55.55</td>
<td>4</td>
<td>akceptor</td>
</tr>
<tr>
<td>56.57</td>
<td>3</td>
<td>akceptor</td>
</tr>
<tr>
<td>66.66</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>54.54</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>55.55</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>42.85</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>56.25</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>88.88</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>59.33</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>41.17</td>
<td>1</td>
<td>akceptor</td>
</tr>
<tr>
<td>56.52</td>
<td>1</td>
<td>akceptor</td>
</tr>
</tbody>
</table>
Program Slovotvorba can:

1. Identify related (derived) words
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
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

- Obecná nastavení
  Část: syn2005  
  Minimalní frekvence: 20  
  Počet vypisovaných výsledků: 30  
  Vynochovat duplicitní shodné části: Ano  

Program slouží k hledání dvojic slov, které se v něčem shodují a v něčem líší. Zadávají se proto dva vzory, v nichž je nutné specifikovat rozdělení pomocí závek.

Například:

Zajímají nás slova, která mají stejné slovotvorné konstrukce jako slovo život a slovec. Části, které se líší, budeme dávat do závek, části oběma slovům společné, ponecháme bez závek:
1. vzor: "(1st) - co odpovídá vzhůru život(a)"
2. vzor: "(2st) - co odpovídá vzhůru život(a)"

Aby program pracoval správně, je třeba se vyvarovat používání závek pro jiné účely (např. v některých regulacích výrazů). Vzory by dala namísto obsahovat znaky střední (.) a procenta (%). Všechny tvary slov jsou programem přesouveny na malá písmena, nemá proto slyšet žadat dát rozsah, v kterém by byla velká písmena.

Poznámka! Zpracování dotazu může trvat poměrně delší čas!

1. vzor: *  
2. vzor: *  

Vynést formulář  

spustit program
SLOVOTVORBA

Zadání
Korpus sym2005
Minimální frekvence: 20
Vypísnat počet výsledků: 30
Výsledky v úplnosti uložit do souboru.

Vzory:
1. vzor: "(*4)
2. vzor: "(*6)

<table>
<thead>
<tr>
<th>#</th>
<th>F0</th>
<th>1. vzor = &quot;(*4)</th>
<th>2. vzor = &quot;(*6)</th>
<th>F0</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>41</td>
<td>blebit</td>
<td>užklebek</td>
<td>128</td>
</tr>
<tr>
<td>2</td>
<td>562</td>
<td>leat</td>
<td>olek</td>
<td>330</td>
</tr>
<tr>
<td>3</td>
<td>32</td>
<td>lit</td>
<td>ušek</td>
<td>113</td>
</tr>
<tr>
<td>4</td>
<td>568</td>
<td>plant</td>
<td>ulanek</td>
<td>368</td>
</tr>
<tr>
<td>5</td>
<td>32</td>
<td>min</td>
<td>ušek</td>
<td>1650</td>
</tr>
<tr>
<td>6</td>
<td>458</td>
<td>zvid</td>
<td>ušedek</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>23</td>
<td>tom</td>
<td>ulon</td>
<td>128</td>
</tr>
<tr>
<td>8</td>
<td>1008</td>
<td>čnit</td>
<td>ušinek</td>
<td>2480</td>
</tr>
</tbody>
</table>

© VJC 2010
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)
Why choose the descriptive approach to grammar over the prescriptive?

1. because that’s what users will appreciate
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
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
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.

...instead of conclusion
Thank you for your attention!